THE UNIVERSITY
of EDINBURGH

Entity-based Coherence in Statistical Machine Translation
A Modelling and Evaluation Perspective

Dominikus Wetzel

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2017

Abstract

Natural language documents exhibit coherence and cohesion by means of interrelated structures both within and across sentences. Sentences do not stand in isolation from each other and only a coherent structure makes them understandable and sound natural to humans. In Statistical Machine Translation (SMT) only little research exists on translating a document from a source language into a *coherent* document in the target language. The dominant paradigm is still one that considers sentences independently from each other. There is both a need for a deeper understanding of how to handle specific discourse phenomena, and for automatic evaluation of how well these phenomena are handled in SMT.

In this thesis we explore an approach how to treat sentences as dependent on each other by focussing on the problem of pronoun translation as an instance of a discourse-related non-local phenomenon. We direct our attention to pronoun translation in the form of cross-lingual pronoun prediction (CLPP) and develop a model to tackle this problem. We obtain state-of-the-art results exhibiting the benefit of having access to the antecedent of a pronoun for predicting the right translation of that pronoun. Experiments also showed that features from the target side are more informative than features from the source side, confirming linguistic knowledge that referential pronouns need to agree in gender and number with their target-side antecedent. We show our approach to be applicable across the two language pairs English-French and English-German.

The experimental setting for CLPP is artificially restricted, both to enable automatic evaluation and to provide a controlled environment. This is a limitation which does not yet allow us to test the full potential of CLPP systems within a more realistic setting that is closer to a full SMT scenario. We provide an annotation scheme, a tool and a corpus that enable evaluation of pronoun prediction in a more realistic setting. The annotated corpus consists of parallel documents translated by a state-of-the-art neural machine translation (NMT) system, where the appropriate target-side pronouns have been chosen by annotators. With this corpus, we exhibit a weakness of our current CLPP systems in that they are outperformed by a state-of-the-art NMT system in this more realistic context. This corpus provides a basis for future CLPP shared tasks and allows the research community to further understand and test their methods.

The lack of appropriate evaluation metrics that explicitly capture non-local phenomena is one of the main reasons why handling non-local phenomena has not yet been widely adopted in SMT. To overcome this obstacle and evaluate the coherence of translated documents, we define a bilingual model of entity-based coherence, inspired

by work on monolingual coherence modelling, and frame it as a learning-to-rank problem. We first evaluate this model on a corpus where we artificially introduce coherence errors based on typical errors CLPP systems make. This allows us to assess the quality of the model in a controlled environment with automatically provided gold coherence rankings. Results show that this model can distinguish with high accuracy between a human-authored translation and one with coherence errors, that it can also distinguish between document pairs from two corpora with different degrees of coherence errors, and that the learnt model can be successfully applied when the test set distribution of errors comes from a different one than the one from the training data, showing its generalization potentials.

To test our bilingual model of coherence as a discourse-aware SMT evaluation metric, we apply it to more realistic data. We use it to evaluate a state-of-the-art NMT system against post-editing systems with pronouns corrected by our CLPP systems. For verifying our metric, we reuse our annotated parallel corpus and consider the pronoun annotations as proxy for human document-level coherence judgements. Experiments show far lower accuracy in ranking translations according to their entity-based coherence than on the artificial corpus, suggesting that the metric has difficulties generalizing to a more realistic setting. Analysis reveals that the system translations in our test corpus do not differ in their pronoun translations in almost half of the document pairs. To circumvent this data sparsity issue, and to remove the need for parameter learning, we define a score-based SMT evaluation metric which directly uses features from our bilingual coherence model.

Lay Summary

Natural language text in documents such as newspaper articles is generally meaningful to a human; it is said to be coherent. This coherence is achieved by establishing relations between parts of the text both within and across sentences. Sentences therefore do not stand in isolation from each other. Only a coherent structure makes them understandable and sound natural to humans. Furthermore, coherence is part of the more general linguistic concept of discourse. In Statistical Machine Translation (SMT) only little research exists on translating a document from a source language into a *coherent* document in the target language. The most prevalent methods consider sentences independently. Therefore, there is both a need for a deeper understanding of how to handle specific discourse phenomena, and for automatic evaluation of how well these phenomena are handled in SMT.

In this thesis we explore an approach how to treat sentences as dependent on each other by focussing on the problem of pronoun translation as an instance of a discourse-related non-local phenomenon. We develop a computational model for cross-lingual pronoun prediction (CLPP) to predict the translation of a pronoun in the context of a bilingual document. We obtain state-of-the-art results exhibiting the benefit of giving the model access to the entity a pronoun refers to. Experiments also showed that features of the model extracted from the target side of the document are more informative than features from the source side, confirming linguistic knowledge that pronouns referring to an entity need to agree in gender and number with this entity. We show our approach to be applicable across the two language pairs English-French and English-German.

The experimental setting for CLPP was artificially restricted by focussing on a small number of pronouns and using human-authored translations. This enabled automatic and fast evaluation in a controlled environment. However, this limitation does not yet allow us to test the full potential of CLPP systems within a more realistic setting that is closer to a full SMT scenario. We provide an annotation scheme and tool, and a bilingual corpus that enable evaluation of pronoun prediction in a more realistic setting. The annotated corpus consists of a collection of documents automatically translated by a state-of-the-art neural machine translation (NMT) system, where the appropriate target-side pronouns have been chosen by annotators. With this corpus, we exhibit a weakness of our current CLPP systems in that they are outperformed by a state-of-the-art NMT system in this more realistic context. This corpus provides a basis

for future research on CLPP and allows the research community to further understand and test their methods.

The lack of appropriate evaluation metrics that explicitly capture non-local phenomena is one of the main reasons why handling non-local phenomena has not yet been widely adopted in SMT. To overcome this obstacle and evaluate the coherence of translated documents, we define a bilingual model of coherence specifically focussing on entities. Our model is inspired by work on monolingual coherence modelling, and it is designed to learn how to rank bilingual documents taking the coherence of both source and target side into account. We first evaluate this model on a bilingual corpus where we artificially introduce coherence errors based on typical errors CLPP systems make. This allows us to assess the quality of the model in a controlled environment and to automatically obtain true coherence rankings without requiring humans to provide such a judgement. Results show that this model can distinguish with high accuracy between a human-authored translation and one with coherence errors, that it can also distinguish between two documents with different degrees of coherence errors, and that the model generalizes well across different variations of coherence error introduction.

To test our bilingual model of coherence as a discourse-aware SMT evaluation metric, we apply it to more realistic data. We use it to evaluate translations of a state-of-the-art NMT system against translations where we automatically corrected pronouns with our CLPP model. A good SMT evaluation metric judges automatic translations according to how humans would judge them. We reuse our annotated bilingual corpus and consider the pronoun annotations as an approximation of how humans would judge the coherence of each document. Experiments show that the model has far lower performance in ranking translations according to their entity-based coherence than on the artificial corpus, suggesting that it has difficulties generalizing to a more realistic setting. Analysis reveals that the automatic translations we use for learning the model parameters and testing it as SMT evaluation metric are too similar to each other. To circumvent this lack of data and diversity, which is a problem for data-driven models, we define a simpler score-based SMT evaluation metric which directly uses features from our bilingual coherence model without the need to learn parameters.

Acknowledgements

I would like to thank my first supervisor Bonnie Webber for accepting me into the PhD program, for the critical discussions in our frequent meetings, and the pointers she gave me during my PhD. I would also like to thank my second supervisor Adam Lopez for his valuable feedback on various paper drafts and this thesis.

I am grateful for the funding provided by the European Union that enabled my research (MosesCore 288487 and HimL 644402).

Thanks also to both my examiners, Jon Oberlander and Jörg Tiedemann, for their interest and constructive discussion.

Many thanks also go to all the members of the StatMT group for the discussions during weekly meetings and for the welcoming and social atmosphere. Furthermore, I would like to thank those people of the group who spent their time on setting up and keeping running the computing infrastructure.

Thanks also to all my great office mates who made our workspace very productive and enjoyable. Furthermore, thanks to the ILCC level 3 lunch group for the many pleasant lunches and conversations, and to all the friendly people in the Informatics Forum in general.

I am grateful to Rico Sennrich, who provided me with great guidance and the support I needed during the last year of my PhD.

Many thanks also go to the Edinburgh Composers' Orchestra, and the Red Wine Circle including Michael Graham, Thomas Grossi and Riinu Ots for all the new, old and unconventional music we played and the fun and relaxing time together.

I am also very grateful for the continuous moral and financial support throughout my studies from my parents, Rita Endres-Wetzel and Adolf Wetzel, and for my whole family including my sister Felicitas Wetzel and brother Nikolai Endres for listening and encouraging me and for enjoying live together.

Finally, I want to thank Lea Frermann for her tireless support and motivation throughout the entire time of my PhD, for the many research-related discussions and her feedback, which all helped me proceed with my thesis. Thanks and ♡s for being there and for exploring the world with me.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Dominikus Wetzel*)

Contents

Acronyms

AOD average out-degree. 34, 108–110, 116, 118, 121, 126, 155

APT accuracy of pronoun translation. 18, 20, 29

BPE byte-pair encoding. 77

CLPP cross-lingual pronoun prediction. iii–vi, ix, x, 2, 3, 5–8, 10, 11, 14, 16, 17, 19, 35, 43–62, 64, 66, 68–76, 79–81, 85, 91–98, 100, 102, 103, 106, 117, 118, 120, 125, 126, 129, 131–133, 153, 157, 162–164, 166–172, 174

CRF Conditional Random Field. 59, 62, 63, 173

DRT Discourse Representation Theory. 4, 22

EDU elementary discourse unit. 26

EGM entity grid model. x, xi, 32, 33, 40, 106, 107, 109, 110, 118, 124, 126, 129, 130, 133, 136, 137, 139, 141, 143, 145, 147, 149, 151, 153, 155, 164, 169, 173, 174

EGraph entity graph. 33–35, 107–109, 111–113, 115–117, 124, 126, 169, 173

EGrid entity grid. 32, 33, 35, 36, 107, 114, 169, 174

KM Kuhn-Munkres. 38, 39, 112–114, 116

LM Language Model. 18, 46, 47, 49, 54, 57, 60, 63, 64, 66–71, 94, 171

MaxEnt Maximum Entropy. 16, 47, 57, 59, 62, 63, 66, 72

NMT neural machine translation. iii–vi, x, 2, 5, 19, 20, 69, 73–75, 77, 79, 93, 98–100, 102, 103, 131, 134, 148, 164, 168, 170, 171, 174

POS Part-of-Speech. 16, 22, 45, 48, 49, 56, 59, 70, 92–94, 133

RST Rhetorical Structure Theory. 4, 24–26

SMT Statistical Machine Translation. iii–vi, ix–xi, 2, 3, 5–7, 9, 13–15, 17–29, 31, 34–36, 39, 40, 43–50, 52, 54, 56, 58–62, 64, 66, 68–76, 78, 80, 82, 84, 86, 88, 90–96, 98, 100, 102, 105, 106, 118, 123, 125–127, 129–171, 174, 175

SVM Support Vector Machine. 33, 119

Chapter 1

Introduction

Natural text in documents, such as newspaper articles, consists of a sequence of sentences that follow a coherent structure. They are on a specific topic, they describe events and their participants, and how they are related to each other. The following text is an excerpt of a newspaper article:

(1) Eleven years ago, Sufjan Stevens sits on the stage in the Prime Club (now Luxor) in Cologne. Beside him stands a flip chart on which the shy-seeming folk singer has drawn the picturesque US state of Michigan in felt-tip pen. The entire audience, some 40 people, is virtually mesmerised by Stevens' performance. Referring to different places, which each time he marks on the map, he talks about the stories behind his meticulously and subtly contrived songs. Where they originated, and what it looks like, in his home country. – WMT16 newstest

It introduces the main participant (i.e. Sufjan Stevens) at the beginning, provides references to the context (i.e. the location and other participants in the room) and is about a coherent topic (i.e. a concert performance). Taking a sentence out of this document or changing the order of sentences would make the article less intelligible and less coherent:

(2) Referring to different places, which each time he marks on the map, he talks about the stories behind his meticulously and subtly contrived songs. Eleven years ago, Sufjan Stevens sits on the stage in the Prime Club (now Luxor) in Cologne.

The first sentence refers to a participant (i.e. *he*), which has not yet been introduced and it is unclear what his talking has to do with something that happened eleven years ago.

The major body of work on Statistical Machine Translation (SMT) has the underlying assumption that sentences can be translated independently of each other.[1] Only in recent years, research on document-level and discourse-aware SMT started to attract attention, working on lifting the strong independence assumptions in traditional SMT that confine translation to a very local context. There has been work on analysing and handling non-local phenomena within and across sentences (e.g. (Tiedemann, 2010; Hardmeier et al., 2013a; Xiong et al., 2013; Guillou, 2016)). This interest has been fuelled by workshops, such as the Discourse in Machine Translation Workshop (DiscoMT), in being organized to encourage contributions in this field. Further examples are the two shared tasks for pronoun translation and prediction in 2015 (Hardmeier et al., 2015) and the cross-lingual pronoun prediction (CLPP) shared task in 2016 (Guillou et al., 2016). Both the workshop and the shared task are going to be held again in 2017, showing the consistent interest in this field, and also showing that it is a problem far from being solved.

Discourse phenomena are much more complex than local phenomena (such as word-order, grammaticality, or fluency that SMT was mostly concerned with). At the same time elements relating to the discourse and signalling discourse structure can be very sparse. For example, establishing a discourse relation by means of a discourse connective (such as *although, but, and*, etc.) contributes a lot to the coherence and cohesion of a text, however, it is often only affected by one word out of the entire paragraph, sentence or clause. These issues provided the motivation that lead to the CLPP shared tasks, focussing on one very specific problem, i.e. pronoun translation, in greater detail. Furthermore, they also yielded the creation of phenomena-specific corpora, e.g. ParCor (Guillou et al., 2014), and evaluation sets, e.g. PROTEST (Guillou and Hardmeier, 2016).

On the other end, discourse phenomena in SMT require new, more suitable evaluation metrics. The frequently used evaluation metrics such as BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) are only defined to measure overlapping translations *within* a sentence against one or more reference translations. Furthermore they are typically restricted to a small context window of adjacent words

[1] Throughout this thesis we use SMT to refer to automatic machine translation in general, including neural machine translation (NMT) approaches. Whenever required, a more specific term is used.

(e.g. four words for BLEU). However, pronouns can have non-local dependencies that can occur in preceding sentences and that influence the actual form of the pronoun. These existing metrics therefore cannot capture non-local phenomena that researchers started to model in the recent years. A few discourse-related evaluation metrics have been proposed for SMT, however, most of them have a number of drawbacks, e.g. that they are verified against sentence-level human judgements rather than human judgements obtained from the entire document context, or that they stay within the sentence context (cf. Section 2.3).

1.1 Thesis Statement

In this thesis we focus on the two interrelated problems of pronoun translation and discourse-aware SMT evaluation. Our starting point is an abstraction from the full SMT scenario in order to simplify the problem at hand and thus to better understand the requirements. Further along in the thesis, we point out what the shortcomings of these simplifications are, proposing and experimenting with ways to move back towards the full SMT scenario.

Pronoun translation in SMT is a problem that needs to be tackled and better understood, and we believe that CLPP is an approach that makes this possible. CLPP systems have so far been evaluated in a simplified and restricted setting. To further understand and correctly model pronoun translation, we need to relax this artificial setup, in order to close the gap between abstraction and realistic setting. For this part of the thesis, we therefore present our work on CLPP exploiting similarities in grammar and structure across two language pairs to study the problem and define a path towards a more realistic setting.

To further understand the role of CLPP within the context of a full SMT system and to enable more research on discourse-aware SMT systems in general, we first require an adequate SMT evaluation metric that takes non-local context and sparsity into account. Pronoun translation is part of the larger phenomenon of entity-based coherence. We believe that there is a systematic correspondence between source- and target-side entity coherence and that this can be exploited to define such an evaluation metric. In the second part of the thesis, we therefore explore these systematic correspondences and define and experiment with such a discourse-aware SMT evaluation metric.

1.2 Motivation and Overview

Text coherence refers to the set of concepts that make a text semantically meaningful and deals with the connectedness of a document. There are quite a few phenomena that contribute to the overall coherence of a document. There is topical coherence, which is concerned with the semantic relatedness of topics that are observed and inferred from documents (e.g. Barzilay and Lee (2004) using topics and topic transitions as content models, or Misra et al. (2008) inferring topics with Latent Dirichlet Allocation (Blei et al., 2003) to detect semantic coherence of a document). There is relational coherence that is concerned with the discourse structure (e.g. *cause and effect* or *conjunction* from Rhetorical Structure Theory (RST) (Mann and Thompson, 1988)) which relates text spans with each other, using overt or implicit cues called discourse connectives (cf. Prasad et al., 2008). There is also event-based coherence, which focuses on the connectedness and relation between events described in a text (e.g. scripts (Schank and Abelson, 1977), narrative structure and event schemata (Chambers and Jurafsky, 2008, 2009)). Finally, there is entity-based coherence which is concerned with entities occurring or being mentioned in a text, and how they are referred to. The latter is strongly based on concepts such as coreference chains and pronominal anaphora. Even though presented as discrete concepts grouped under coherence, some of them are more directly related to each other (e.g. entities also play a central role in events and Discourse Representation Theory (DRT) (Kamp, 1981), or world-knowledge and expectations, and linguistic constraints together contribute to how pronominal referring expressions are interpreted (Kehler and Rohde, 2013)). Despite the importance of each of the mentioned types of coherence, we will focus on entity-based coherence throughout this thesis. In the context of translation, entity-based coherence consists of basic building blocks such as pronoun translation or prediction, anaphora and coreference resolution.

One of the building blocks that contributes to the coherence of a document are anaphoric pronouns. In the context of translation, pronouns have been identified as posing difficulties that need to be overcome (Le Nagard and Koehn, 2010; Hardmeier, 2014; Guillou, 2016). They vary in their use across languages, they have to fulfil certain grammatical constraints, such as agreement with their antecedents and they can have highly non-local dependencies, i.e. their antecedent might be in sentences preceding or following the current sentence. This problem has been picked up by two shared tasks focussing solely on the cross-lingual prediction and translation of

pronouns (Hardmeier et al., 2015; Guillou et al., 2016).

The general setup of the shared tasks for CLPP is a set of parallel documents with human-authored source and target sides from which target-side pronouns have been removed and replaced with a placeholder. Via word alignments, source- and target-side pronouns are related to each other. The task is then to predict for each source-side pronoun instance, the most likely translation of the source pronoun, choosing from a closed set of classes. A detailed description of the shared task setups is provided in Section 3.1.1. One of the major advantages of these CLPP shared tasks is that the investigation of automatic pronoun translation was done outside of a full SMT system working with human-authored translations. This allows for evaluating performance on just this phenomenon without introducing other variation from the translation process. In full SMT systems, it is often not straightforwardly possible to verify if handling a particular sparse problem (e.g. pronoun translation) caused improvements in translated pronouns in the final SMT output, or whether it just caused unrelated changes in the resulting translation hypotheses. Furthermore, the setup with a human-authored target side allows for automatically obtaining gold labels and hence large amounts of training data.

With coherence in mind, we therefore first focus on pronoun translation and propose a CLPP system that attempts to find a solution to this problem (Chapter 3). The CLPP shared tasks view this problem as a cross-lingual prediction task. This enables us to gain insight into particular problems pronoun translation faces, what specific context is required to deal with pronoun translation and what the similarities are that guide the pronoun translation process across language pairs. At the same time, we realize that the CLPP shared task abstracts away considerably from the realistic final goal (and hence from many difficulties yet to come) of a fully pronoun-aware SMT system in its limited focus and form of data it operates on.

To advance research on that end, we propose to remove the requirement of human-authored target-side data for CLPP systems. In Chapter 4 we replace this target side with an automatic translation generated by a state-of-the-art NMT system. The intertwined requirement that comes with this automatic translation is that gold labels for predicted target-side pronouns have to be provided manually, since automatic extraction of these gold labels is no longer possible from an automatically translated target side. We therefore propose an annotation and data collection scheme to obtain these gold labels. The corpus created in that way provides a basis for other researchers to experiment with and better understand their CLPP systems in a more realistic setup.

Furthermore, we use this corpus as basis for verifying our SMT evaluation metric (cf. Chapter 6) and whether it correlates with these human judgements.

Evaluation of CLPP systems in the setup of the shared tasks can be done automatically, which is a great advantage. However, it has the underlying assumption that the rest of the translation remains unchanged. In particular it assumes that pronoun antecedents do not change in translation, therefore the gold pronoun label will always remain correct. Achieving greater flexibility and a more realistic SMT setting demands taking more context into account in evaluation. It has been shown that pronoun translation evaluation based on matching just the pronouns in the reference translation too harshly penalizes the SMT system (Luong and Popescu-Belis, 2016). If the translated pronoun does not match the reference, it is counted as an error even though in the particular translation, it might in fact be the right pronoun, since it agrees with its antecedent.

Furthermore, prior work on document-level and discourse-aware SMT often showed no improvements with traditional sentence-based evaluation metrics such as BLEU. It has been commented that this is due to the lack of SMT evaluation metrics that go beyond sentence boundaries. There have been only very few attempts to tackle this problem and none that focussed on entity-based coherence. This is why we propose to focus on this aspect of discourse to develop an SMT evaluation metric that is capable of capturing the entity-based coherence of translations. As a first step towards this SMT evaluation metric, we propose a model of bilingual coherence that is inspired by earlier work on monolingual coherence modelling (Chapter 5). This follows our hypothesis that there is a strong correlation of entity-based coherence in the source document and its human translation. In a bad translation this correlation is violated and systematic differences can be identified to rank this bad translation accordingly with respect to better translation hypotheses.

We then use the bilingual model of coherence as a basis for our discourse-aware SMT evaluation metric (Chapter 6). Earlier work on SMT evaluation for discourse or document-level phenomena often used existing SMT system outputs. It was never defined exactly what these systems modelled, hence it is unclear whether any of these systems actually handle discourse phenomena or not. If in earlier work a good correlation with human judgements could be shown in that it ranked certain SMT systems higher, then this was most likely because these SMT systems handled the captured discourse phenomenon by chance. We want to test SMT systems that specifically handle a discourse phenomenon and want to find out in this context whether our discourse-

aware evaluation metric can capture relevant differences. This way we can find out if we actually handled the discourse phenomenon properly in the SMT system. This motivates the use of at least two SMT systems for verifying our SMT evaluation metric: One that acts as a baseline without explicitly modelling discourse phenomena, and one in which such a discourse-aware model has been integrated. We therefore propose to integrate our CLPP system predictions as post-editing step into a baseline translation. In this way, we combine our efforts on the modelling side (using CLPP in context) and our efforts on the evaluation side (based on our bilingual coherence model). And we gain some insights as to what degree the CLPP system is applicable and sufficient to solving pronoun translation.

1.3 Contributions

The following list summarizes the main contributions of this thesis:

- Two CLPP systems for handling pronoun translations. They handle pronoun predictions for English-French and English-German showing that the same approach generalizes over language pairs. They also provide insight into what requirements are necessary and what features are useful for pronoun translation.

- A manually annotated English-German corpus for pronouns in context of automatic translations. This enables us to test CLPP systems on realistic data and could be used by a future CLPP shared task. Furthermore it acts as a proxy for document-level human coherence judgements required for verification of our discourse-aware SMT evaluation metric.

- A bilingual model of entity-based coherence that models the relationship between source- and target-side coherence in parallel documents. We evaluated this model on an automatically created parallel corpus where we artificially manipulated the coherence on the target-side based on typical errors made in pronoun prediction.

- A discourse-aware SMT evaluation metric based on our bilingual coherence model that operates on the document level for enabling more informed research on and evaluation of document-level SMT. This metric is tested for correlation with human judgements obtained at the document level.

1.4 Relation to Published Work

In Wetzel et al. (2015) and Wetzel (2016) we describe our work on CLPP for submission to the CLPP shared tasks 2015 and 2016, respectively, which is covered in Chapter 3 in this thesis.

Chapter 2

Background and Related Work

We first give a brief introduction to pronouns and coreference chains with regards to our thesis (Section 2.1). We then give an overview of existing work on automatic pronoun translation for SMT (Section 2.2). Following this, we present work on evaluation metrics for SMT that go beyond the sentence boundary and take discourse-phenomena into account (Section 2.3). After that we provide an overview of entity-based coherence modelling (Section 2.4). Finally, we discuss coreference resolution evaluation methods and how they relate to our SMT evaluation metric (Section 2.5).

2.1 Pronouns and Coreference Chains

Pronouns and coreference chains play a central role in this thesis. In this section we give an overview of the phenomena relevant to our work.

2.1.1 Pronouns

Pronouns are a frequent group of words that fulfil a variety of different functions. Guillou (2016) provides a detailed overview of the different functions pronouns can take. There are anaphoric pronouns (which have an antecedent in preceding or following discourse), pleonastic pronouns (that act as dummy subjects and do not have a referent in discourse), event pronouns (linking to events in discourse), pronouns with an extra-textual referent, with a speaker/addressee referent and with a generic referent. In this thesis we generally focus on personal pronouns (i.e. anaphoric pronouns) and non-referential pronouns (i.e. pleonastic pronouns).

Personal pronouns are referential pronouns that establish a link in discourse to their

antecedents. These are words or phrases that provide enough context so that it can be understood what or who the pronoun refers to. In the following example *Clinton* is the antecedent of the personal pronoun *her*.

(1) [Clinton]$_1$ maintains a large lead among women and moderates, but those leads have narrowed. [Her]$_1$ support among men has dropped considerably and Sanders only trails [her]$_1$ by 5 points. – WMT16 newstest

Referential pronouns can further be divided into two groups. Those whose antecedent occurs within a sentence (i.e. *intra-sentential* pronouns). And those whose antecedent occurs in a preceding or following sentence (i.e. *inter-sentential* or *cross-sentential* pronouns).

Non-referential pronouns (also referred to as pleonastic pronouns) are used to fill a required syntactic position, but do not have an antecedent in discourse. An example is given as follows.

(2) The Met Office predicts, *it* will be raining in Edinburgh tomorrow.

In this thesis, non-referential pronouns play a role in our work on CLPP in two ways. First, they need to be treated differently to referential pronouns since they do not have antecedents in discourse which they have to agree with. Second, since they cannot be part of a coreference chain, their identification acts as feature to counterbalance erroneous decisions by automatic coreference resolution systems.

2.1.2 Pronouns Across Languages

In our work on CLPP, we focus on the translation of the two English pronouns *it* and *they*, and their typical translations into French and German as given in Tables 2.1 and 2.2, respectively. Descriptions of French and German pronouns are based on definitions from Guillou et al. (2016). With English as the source language, there are ambiguous pronouns which need to be resolved in order to correctly translate them into French and German.

For example *it* may be a pleonastic pronoun translating to *man* or *es* in German as shown in Example (3).

(3) a. [...] I could only imagine what *it* must be like to be trapped in that hell.

ce, c’	used in the construction *c’est* (i.e. *it is*)
elle, elles	3rd person subject position feminine personal pronoun (singular and plural)
il, ils	3rd person subject position masculine personal pronoun (singular and plural)
	il is also used as pleonastic pronoun (*il pleut*, i.e. *it rains*) or as pronoun with generic referent
ça, ç’, cela,	demonstrative pronoun (the second form is used before words starting with a vowel)
on	indefinite pronoun

Table 2.1: French pronouns used in our work on CLPP.

er	3rd person subject position masculine personal pronoun (singular)
sie	3rd person subject position personal pronoun (singular feminine and plural)
	if the case of characters is ignored, this can also be the polite form of a 2nd person subject position personal pronoun (*Sie*, i.e. *you*)
es	3rd person subject position neutral personal pronoun (singular)
	also used as pleonastic pronoun (*es regnet*, i.e. *it rains*)
man	indefinite pronoun

Table 2.2: German pronouns used in our work on CLPP.

b. [...] konnte ich mir nur vorstellen, wie *es* wäre, in dieser Hölle
[...] could I myself only image how it would-be in this hell
gefangen zu sein.
trapped to be
– CLPP16 test set

If *it* is used as a referential pronoun, it has to agree in number and gender with the target-side antecedent. Since German has grammatically gendered pronouns, a decision between *er, sie, es* has to be made. Consider Example (4), where *sie* has to be chosen, since *Sklaverei* (i.e. *slavery*) has feminine gender.

(4) a. [Slavery]$_1$ exists everywhere, nearly, in the world, and yet [*it*]$_1$ is illegal everywhere in the world.

b. [Sklaverei]$_1$ existiert fast überall auf der Welt, obwohl [*sie*]$_1$
slavery exists nearly everywhere in the world although it
überall auf der Welt verboten ist.
everywhere in the world illegal is
– CLPP16 test set

Furthermore, pronouns do not necessarily have a counterpart in the other language. In Example (5), a different formulation is used in German, which does not require a pleonastic pronoun.

(5) a. The air is thick with heat and dust, and *it*'s hard to breathe.

b. Die Luft ist stickig von Hitze und Staub und das Atmen fällt schwer.
the air is thick from heat and dust and the breathing is hard
– CLPP16 test set

2.1.3 Coreference Chains

Pronouns form a part of a bigger structure in discourse. Together with coreferring nouns, they are grouped into equivalence classes, where all elements in each class refer to the same entity. These equivalence classes are referred to as coreference chains. Consider this excerpt of a document:

(6) Last year at TED [I]$_1$ gave an introduction to [the LHC]$_2$. And [I]$_1$ promised to come back and give you an update on how [that machine]$_2$ worked. So this is it. And for those of you that weren't there, [the LHC]$_2$ is the largest scientific experiment ever attempted – 27 kilometers in circumference. [Its]$_2$ job is to recreate the conditions that were present less than a billionth of a second after

> [the universe]$_3$ began, up to 600 million times a second. – TED talk from Brian Cox on the Large Hadron Collider

There are two main entities in this document, one is *Brian Cox* and the other one is the *Large Hadron Collider*.[1] These entities are mentioned several times throughout the document either using nouns to refer to them, or pronouns. All mentions with the same index form a coreference chain. A coreference chain with only a single member is also often referred to as a *singleton*. An example of this is marked with the index 3 above, but for better visibility not all singletons are marked.

Throughout the thesis, but especially in Chapters 5 and 6, we will make extensive use of coreference chains. In these cases, we only focus on coreference chains that contain more than one element, i.e. we do not look at singletons.

2.2 Automatic Pronoun Translation

Le Nagard and Koehn (2010) present one of the early instances of research on pronoun translation for SMT. They focus on English-French translation of the pronouns *it* and *they*. These can be translated in multiple ways. They point out that one of the most important factors for disambiguation of the source pronoun is the target-side antecedent, and that these antecedents often occur in preceding sentences. They first identify the antecedent of a pronoun on the source side and then determine the target-side antecedent via word alignment links. From this target-side antecedent, gender information is extracted. This information is then added to the source-side pronoun, essentially disambiguating its intended target-side meaning. At training time, the target-side translation of the antecedent is given by the parallel corpus. At test time, they resort to using the output of a baseline translation of the full document to determine the target-side antecedent. For identifying the antecedent on the source-side, they employ two different rule-based anaphora resolution algorithms (i.e. Hobbs (1978), and Lappin and Leass (1994)). They both operate on syntactic parse trees, identifying the most likely noun antecedent with a set of heuristics on how to traverse the trees and sentences, and how to distinguish similarly plausible antecedent candidates. Le Nagard and Koehn (2010) automatically filter out pleonastic uses of *it*, so that these algorithms do not try to find an antecedent for these non-referential pronouns. They train a phrase-based SMT system on data where the source-side pronoun is disambiguated by adding the gender

[1] There are other entities, such as *introduction*, *experiment*, etc. which we leave out for clarity.

of the target-side antecedent. They investigate the lexical translation probabilities of each trained system. In the baseline they observe a strong bias towards the masculine pronoun in French. Systems with the anaphora resolution algorithms show a shift towards the correct translation. The BLEU scores, however, remain almost the same between the baseline and their proposed system. Also, a manual inspection of pronoun translation reveals, that baseline and extended systems have around the same count of correctly translated pronouns. On a subset of the test set, the performance of anaphora resolution was verified and it was found that only 56% of the pronouns were correctly resolved. This leads the authors to conclude that the negative result is mostly due to bad performance of the anaphora resolution algorithms. One of the features in our CLPP system identifies the features of the target-side antecedent in a similar way via the source and word alignments. We use a state-of-the-art coreference resolution system instead of manually modified algorithms. Gender features are extracted based on a lexicon, whereas we increase coverage by using a morphological tagger. Furthermore, we handle pleonastic pronouns and null-translations, which are ignored in the above work. In the above presented two pass-decoding process the antecedents get fixed by the baseline translation, however in the second translation phase, there is no mechanism that ensures that the antecedent gets translated in the same way. The previously determined antecedent feature might therefore be no longer valid.

Guillou (2012) presents work on English-Czech pronoun translation. Their work follows closely the work by Le Nagard and Koehn (2010), however, their main focus is to rule out sources of error in the resolution and feature extraction process. Instead of automatically resolving coreference on the source side, they take a manually annotated corpus for pronoun coreference (the BBN Pronoun Coreference and Entity Type Corpus). This corpus also identifies pleonastic pronouns, which are removed in the work presented. The semantic head of the source-side antecedent is identified from syntactic gold parses. The sentence-aligned target-side data is taken from the PCEDT2.0 corpus, which comes with automatic word alignments between the source corpus and the target documents, and provides gender, number and animacy information for each token. The baseline system is a phrase-based SMT system trained on the plain parallel corpus. The extended system is trained on data, where the source-side pronouns have been annotated by the gender, number and animacy features of the target-side antecedent. At training time the target-side is given by the parallel corpus. At test time, the translations of the antecedent (required to identify agreement features) are taken from the translation of the baseline system, following the approach by Le Nagard and Koehn

(2010). The agreement features from these antecedents at test time are obtained from a dictionary. In order to avoid a different translation of the antecedent when translating with the extended system, the same word alignments were used during training of both SMT systems. This does not give a guarantee, but manual inspection showed that the antecedent rarely changed. An automatic evaluation is proposed targeted towards pronoun translation evaluation. Among others, there is the count of how many times the translated pronoun is in a list of valid possible translations (compiled by hand) for the given source pronoun and agrees with the antecedent. A matching against the reference translation is not required. Manual evaluation is also performed. Counts are collected which of the two system translations of pronouns is preferred if the translation differs (overall, and only in those cases where the agreement features were correctly identified). This is done on a small sample of translations. Automatic results show only minor improvements of a few pronouns in the extended system. Manual evaluation results show that the extended system is marginally preferred. In those cases where the annotated system received correct agreement features, a clearer preference is shown for this system. No BLEU scores are given, so it is not possible to say how strong the baseline is. A weak baseline would result in bad antecedent translations. Furthermore, the authors comment that the annotators had difficulties to judge the systems due to errors in syntax in the translation. This suggests that the overall translation quality was not great to start with. While using manually resolved coreference is a great way of reducing the preprocessing errors, we want to focus on a more realistic setting, where we can base our models on automatically resolved coreference.

Hardmeier et al. (2013b) investigate pronoun translation by framing it as a pronoun prediction problem. They focus on English-French pronoun translation of subject-position third person pronouns. The basis for the task is a parallel corpus with human-authored translations and word alignments between each source and target sentence. The translation of a closed set of source pronouns has to be predicted by choosing among a closed set of target-side pronouns. Features are extracted for a baseline classifier and two different neural network architectures. Source-side features are one-hot vector representations of a 3-word window around the source-side pronoun. Similarly to Le Nagard and Koehn (2010) target-side antecedents are obtained via sources-side coreference resolution and word alignments. In this work, instead of relying on the best identified antecedent, a weighted average over all possible antecedent candidates is computed. In the first neural network, the weights correspond to the probabilities from the coreference resolution system BART that follows a mention-pair model. In

the second neural network, these probabilities are jointly learnt with the pronoun prediction task. The baseline is a Maximum Entropy (MaxEnt) classifier. In the neural networks the source-side and target-side features are separately mapped to a lower dimensional space. This layer is then mapped to a hidden layer expected to learn a representation for gender and number features. Finally, a softmax layer produces the pronoun predictions. Experiments are run on two data sets. One is from the IWSLT corpus containing TED talks. The other one is from the NewsCommentary6 corpus. The performance of the baseline is generally better on the TED data, than on the News-Commentary6 data. Also in the baseline, the rare class for the plural feminine pronoun in French (*elles*) has a very low recall. Compared to the baseline, the accuracy of the first neural network model is more or less the same. Recall is slightly better at the cost of precision. The performance on *elles* is slightly better. The results for the second neural network are even better on accuracy, and provide a better balance between precision and recall for *elles*. This work inspired the succeeding CLPP shared tasks (Hardmeier et al., 2015; Guillou et al., 2016) to which we submitted CLPP systems. Contrary to this model, we also use target-side context around the pronoun that is to be predicted. Instead of taking an average over all target-side antecedent candidates, we take the best antecedent and extract explicit gender and number features from it. This is also in contrast to their model, which hopes to learn these two features within the hidden layer of the neural network. These explicit features were shown to be important in our feature ablation study.

Luotolahti et al. (2016) present their approach to CLPP with a deep recurrent neural network. According to the official results of the CLPP16 shared task (Guillou et al., 2016), their system performed best in terms of macro-averaged recall. For each target-side pronoun, the input to the network is defined as follows: left and right sentence context of the aligned source-side words (inclusive), left and right sentence context of the target-side pronoun (exclusive), where the former comes in three different flavours using lemmata, Part-of-Speech (POS) tags or a combination of both. The first layer is a randomly initialized word-embedding layer, followed by two layers of gated recurrent units (GRUs), followed by a rectified linear unit (ReLU), with a final softmax layer that provides a distribution over the possible pronoun classes. For training, they modify the loss function such that each pronoun instance is weighted inversely proportional to the frequency of the pronoun class label. This indirectly optimizes for macro-average recall, the official metric, since it penalizes errors on low frequency class labels more. The primary system submission uses only sentence-internal context (of a maximum

number of 50 context words). Different experimental setups reveal that without the modification to the loss function (i.e. no weighting), a drop of 4-6% (absolute) of macro-averaged recall is observed. If the strictly sentence-internal context is relaxed to allow for crossing sentence boundaries, a drop in performance is also observed. Finally, if the Europarl corpus is removed from the training set, performance drops considerably by about 7-14% (absolute) in macro-average recall. In contrast to our CLPP submission, they do not explicitly identify the antecedent of a pronoun, nor do they use gender features from the pronoun antecedent. In fact, in all of the cases where the noun antecedent is outside of the sentence context, their system has no chance of seeing the antecedent. Their good results are partially explained by optimizing their model towards the official shared task metric. Even more substantially contributing to the results is the use of the Europarl corpus. We do not use this corpus, since we require clear document boundaries for coreference resolution, which Luotolahti et al. (2016) do not need. Extending their model to cross the sentence boundaries lead to worse results, suggesting that just adding more context in form or raw sequences of words is not sufficient, and explicitly identifying the antecedent might also be necessary in their model.

Rios Gonzales and Tuggener (2017) present work on pronoun translation focussing on a pro-drop language on the source side. Subject pronouns can be implicit in these languages (i.e. null-subjects), so they have to be made explicit when translating into a language that is not a pro-drop language. They experiment on the language pair Spanish-English and focus on null-subject and possessive pronouns. English only distinguishes gender with persons, and uses neutral gender otherwise. Furthermore, the gender of persons remains the same across the two languages, so it can be easily determined from the source-side. Possessive pronouns in Spanish are only marked for number determined by the possessed object. In English, however, they need to agree in gender and number with the possessor. For making null-subjects explicit and obtaining grammatical features for possessive pronouns, the source-side is first processed with a coreference resolution system. They adapt the incremental entity-mention model from CorZu (Tuggener, 2016) to Spanish. For this they include all verbs used in conjunction with null-subjects as markables. This allows them to make the null-subjects explicit and to copy over gender and number information from the coreference chain it belongs to. The processed source-side with explicit null-subjects and additional gender and number information is then used for training a phrase-based SMT system (Koehn et al., 2007). The training, development and test data is chosen from the NewsCom-

mentary11 corpus to include frequent cases of null-subjects and possessive pronouns. Results are therefore obtained based on a fairly small training set and on a non-standard test set. In experiments, it was determined that the Language Model (LM) has a high bias towards translating masculine pronouns. The LM was therefore trained on data with a large number of sentences sampled for feminine pronouns, in addition to unmodified large monolingual corpora. Results are measured in BLEU and accuracy of pronoun translation (APT), which we explain in detail in Section 2.3. Both metrics show improvements over a baseline, with the former they are only minor, but with the latter they show strong improvements with respect to pronouns. An oracle experiment with three manually annotated source documents for coreference show that coreference resolution errors are one reason for limiting the performance and that there is still room for improvement. In our experiments, we do not have the problem of missing subjects, since our source language is always English. So this method is not applicable to our systems. Furthermore, in the above work, the grammatical features, such as gender and number can be determined from the source language, whereas in our work, we need to have knowledge from target-side coreference chains, since our target-side languages (French and German) pose grammatical gender and number requirements on pronouns, that do not exist in English.

Luong et al. (2017) present work on Spanish-English. They handle possessive pronouns and determiners, and third person subject-position pronouns. Both of these can be translated in multiple ways into English as described above (cf. Rios Gonzales and Tuggener, 2017). This work employs a translation model from their earlier work on English-French (Luong and Popescu-Belis, 2016), extending it to make it suitable to the new language pair. This is an easier problem, since all the features required can be obtained from the source side. They propose a coreference model, which learns the probability of a pronoun translation, given the source pronoun and source antecedent features (gender, number and humaneness). When estimating the probabilities from a parallel corpus, the source side is first processed with CorZu for Spanish (as described in the previous paper) to obtain antecedent links. In the estimation, the n-best list of antecedent candidates and their respective scores in this list is taken into account. At test time, the best antecedent for each source pronoun is identified. The pronoun is then marked-up with the features from the antecedent. These special tokens are then handled by a secondary translation model, that is backed by the coreference model above. Two phrase-based SMT systems with and without the secondary translation model are trained on a small subset of the NewsCommentary11 corpus and tested on a

non-standard test set from the same corpus. A small increase (2% absolute) in the APT score is observed between the baseline and the extended system. Using oracle annotations on a small subset of sentences on the source-side (both antecedent and features), a large increase of 35% (absolute) in the APT score can be observed. The authors point out, that one of their shortcomings is that their coreference model does not take context into account, other than the extracted antecedent features. Furthermore, they do not consider null-subjects. In their earlier work on English-French, they have to use a two-pass decoding setup, since the antecedent in the target language has to be determined first. This is not necessary in the Spanish-English language pair, since all the required information can be determined in the source. Our CLPP systems take local context into account when making pronoun predictions. Furthermore, the approach in this paper is not applicable to English-German or English-French, since in both of these language pairs, the relevant features have to be determined on the target side. On the other hand, they integrate their coreference model directly into a phrase-based SMT system at decoding time, whereas we only present a post-editing variant.

Miculicich Werlen and Popescu-Belis (2017) present work on using coreference information for translating from Spanish to English. They propose two approaches. The first one is based on reranking translation hypotheses, by using coreference resolution evaluation metrics that compare source- and target-side chains. The second one is a post-editing approach, where the best translation for a source-side chain is obtained based on pair-wise target-side mention scores. Both approaches are evaluated manually and with BLEU on a non-standard test set of 10 documents from the AnCora-ES corpus. The first approach relies on gold standard annotations from the AnCora-ES corpus on the source-side and the Stanford statistical coreference resolution system (Clark and Manning, 2015) on the target-side. The target-side chains are first projected to the source-side via word alignments. This provides the input to the three standard coreference evaluation metrics MUC, B^3 and CEAF-m (cf. Section 2.5), with the source-side annotations as gold standard, and the projected chains as system chains. This approach is motivated by showing that there is a correlation between translation quality according to BLEU and the three individual coreference metrics, when comparing the reference translation, a commercial NMT system and a baseline phrase-based SMT system. This approach is used in a reranking framework, where among the n-best lists for each sentence in a document, the combination of sentences that maximize the average of the above coreference metrics is searched. The search is approximated in a beam search by incrementally selecting the sentence that maxi-

mizes the above score, pruning away lower scoring ones. Sentences leading to the same coreference score are pruned if they have a lower BLEU score. The second approach is a two-stage decoding setup with a final post-editing step. In the first stage, a baseline translation is obtained. In this translation, the source-side mention head words (given by gold coreference annotations) are projected to the target side and inserted there. In a second stage, an n-best list of possible translations for each mention in context is obtained. This is done by using the modified translation from the first step as new source document. A filtered phrase-table is used where all phrases with words other than mention words are removed. The decoding then produces the above mentioned n-best lists. A final translation for each mention in a cluster is then selected by maximizing the pair-wise mention score on the target-side. This pair-wise score is obtained from the pre-trained Stanford statistical coreference resolution system. Experiments are run on a test set with ten documents from the AnCora-ES corpus. The reranking approach and post-editing approach are compared to a phrase-based SMT baseline and an NMT system. Evaluation is done by BLEU to test for overall translation quality. The accuracy of pronoun translation is determined with the same method as described in (Luong and Popescu-Belis, 2016) using the APT metric (and as summarized in Section 2.3). Both reranking and post-editing approaches improve the accuracy of pronoun translation with respect to the baselines. However, in the reranking approach BLEU decreases considerably, while in the post-editing system, BLEU remains the same. The latter observation is an indication that with the post-editing system the general translation quality is not affected, while improving pronoun translation. A manual evaluation of four documents is performed where the annotator has to compare a mention in a translated sentence to the reference translation and judge whether it is wrong, acceptable or equal to the reference. The post-editing system reduces the number of wrong mention translations, whereas the reranking system performs worse than the baseline. Both approaches are based on the hypothesis that a good translation should have similar coreference chains compared to the source-side. While the authors integrate this concept into a translation system, we use this hypothesis to define an SMT evaluation metric that can evaluate document coherence. The post-editing approach shows promising results, however automatic and manual evaluation of pronoun or mention translation, respectively, only verify against the reference translation. The antecedent information which both influences pronoun and mention translation, is not considered there. It therefore remains to be shown that translations of pronouns and mentions are valid translations with respect to how their antecedent was translated. Another draw-

back of the approach is that it relies on gold annotations on the source side whereas we explicitly do not have this requirement. Furthermore, the first approach is based on the assumption that implicit pronouns in Spanish are made explicit, otherwise the coreference resolution evaluation metric would penalize potentially correct translations if they do not have an explicit counterpart on the source side. Finally, experiments are run on a non-standard and small ten-document test set. If the approach generalizes remains to be seen.

2.3 Evaluation Metrics for Document-level SMT

The most commonly used evaluation metric for automatically assessing translation quality of SMT systems is BLEU (Papineni et al., 2002). It is a precision-oriented metric and is designed to capture adequacy and fluency of proposed translations. This is done by measuring the n-gram overlap between a proposed translation and one or more reference translations (conventionally n ranges between 1 and 4). As such it is applied sentence by sentence, and averaged over the entire corpus. A brevity penalty specific to that corpus is computed in order to penalize overly short translations with respect to their reference. This metric therefore makes independent decisions for each sentence and neglects the fact that sentences in a text are interconnected via various aspects such as topic, discourse connectives, coreference chains, etc. which make a text cohesive and coherent. Furthermore, it ignores possibly correct translations of pronouns that do not match the reference, but are correctly agreeing with the translation of their antecedent. Finally, discourse-level phenomena such as pronoun translation often involve only one word out of the entire sentence. BLEU however gives equal importance to every word and such a small change is not adequately captured (cf. Hardmeier, 2014, Section 6.4.4).

Comelles et al. (2010) state that up to this point, work on SMT is only evaluated segment by segment, where a segment is usually a sentence.[2] They propose an SMT evaluation metric based on discourse representations. These representations focus on important parts of a document (i.e. who did what to whom, when, where and why?). Furthermore, these representations are not simply concatenations of individual sentences, however they span across longer parts of the document, e.g. via anaphoric links or discourse connectives. The authors claim that having such a discourse representation is essential in determining the semantic equivalence of an automatic translation

[2]This work is sometimes cited with a different author order as (Gimenez et al., 2010)

and the corresponding reference document. The discourse structures in their work are extracted with Boxer (Bos, 2008) (from the C&C Tools (Curran et al., 2007)), which follow the DRT (Kamp, 1981). They are represented as first-order logic formulae encoding entities and the relationship between them. Boxer can be applied to individual sentences, but also to a document as a whole. This is what Comelles et al. (2010) take advantage of. They explore three different metric formulations. The first one is based on the number of matching subpaths (of length four) between automatic translation and the reference. The second and third one compute the average lexical (or alternatively POS tag-based) overlap between automatic translation and reference. Of these three metrics, they explore one variant based on discourse representations extracted sentence by sentence, and one extracted for the entire document at once. For testing their method, they compute the correlation between their evaluation metrics and human judgements on Arabic-English documents. Document-level human judgements were not available so to obtain them, they simply averaged sentence-level judgements. Results show a higher correlation with human judgements for the sentence-by-sentence variants over the document-level ones. Furthermore, the sentence-based evaluation metric METEOR shows a similarly high correlation with human judgements as the first variant. The authors attribute this negative result to three possible causes. First, there might be parser errors by Boxer on SMT output. Second, the human judgements might be biased towards sentence-level evaluation. Third, the similarity measures employed might not be expressive enough. While this work takes discourse structure into account, they only verify their evaluation metric against human judgements that were elicited sentence by sentence. This is one of the major differences to our work. Evaluation metrics that take cross-sentence context into account, have to be verified by human judgements that also considered the same cross-sentence context.

Hardmeier and Federico (2010) touch on the topic of evaluating discourse-level phenomena in SMT. Their main work focusses on pronoun translation between English and German. They propose a word-dependency model between words aligned to a source-side pronoun and words aligned to its source-side antecedent. This is formulated as a bi-gram language model and integrated as a feature function into the phrase-based SMT decoder Moses. BLEU scores are virtually the same between a baseline and the extended system. They propose a pronoun-specific evaluation method to get a more detailed insight into performance of pronoun translation. Via word alignments between the source document and reference translation, they collect the correct translations of source-side pronouns. Similarly, via word alignments from the decoder

between the source document and automatic translation, they collect proposed translations of source-side pronouns. Their evaluation score then captures the number of correctly translated pronouns. However, since word alignments can be one-to-many (including one-to-none), instead of accuracy they define precision and recall. The numerator of both is defined as the count of a particular word occurring in the automatic translation limited by the number of times this word occurs in the reference translation. This count is then summed over all aligned words of the automatic translation. The denominators for precision and recall are the number of words in the automatic translation or reference translation respectively. The translation system with the additional feature function improves recall of pronoun translation by a small, but statistically significant amount. The score captures pronoun translation accuracy only to a certain degree, since it is inherently restricted to the sentence level, only comparing proposed pronoun translations to reference translations. It does not take into account the actual translation of the antecedents of pronouns, which might in fact require different pronoun translations than what the reference translation dictates. Furthermore, this score is not tested with regards to what humans consider a correct translation.

Wong et al. (2011) present work on using lexical cohesion for SMT evaluation for translation into English. They observe that previous SMT evaluation metrics operate sentence by sentence not taking into account document-level concepts of cohesion and coherence. Furthermore they note that any document-level SMT evaluation metric has to be verified against human judgements that are themselves obtained from taking the whole document into consideration. The authors focus on lexical cohesion which is observed as reiteration and collocation in a document. Lexical cohesion is used appropriately by humans without problem. They use the right amount of lexical cohesive devices to produce a cohesive text, while not overusing them (e.g. by reiterating too often). The authors hypothesize that a difference of usage of lexical cohesive devices in human translations vs. automatic translation is observable, since humans know to use lexical cohesion appropriately, but SMT systems do not. They analyse the amount of lexical cohesive devices among content words in both human and automatic translation, where the former shows a higher number of content words used as lexical cohesive devices. This leads to their hypothesis that this difference can be quantified and therefore lexical cohesion can be used as a measure of quality of SMT output. They tackle the lack of document-level human judgements by manually annotating documents according to their coherence. Coherence of a text can be shown if consecutive parts within a sentence or complete sentences themselves can be conjoined

by an appropriate relation label according to RST. For their annotation, they simplify this idea and only consider entire sentences as minimal unit. A score is then given to each sentence depending on whether it can be conjoined with another sentence with an allowed relation (score of one), or not (score of zero), or if the annotator is unsure (score of 0.5). These scores are then averaged over all sentences in the document. The particular relation label is ignored. These annotated document-level scores are compared to sentence-level scores of human adequacy judgements (provided by the Metrics MATR08 task). The comparison shows that only for high sentence-level adequacy judgements, both types of scores correlate. However, for lower adequacy judgements, the document-level scores are more spread out across the scale. This shows that sentence-level judgements do not fully reflect the information obtained through document-level judgements. An experiment is then performed to verify the utility of lexical cohesion as SMT evaluation metric. This is done by measuring the correlation of the metric output with human judgements. A baseline metric is defined as the unigram match of stemmed words between translation and reference (by using the F-measure of precision and recall). The proposed lexical cohesion metric is defined as the ratio of words used as lexical cohesive devices divided by the number of all content words. Since lexical cohesion is a high-level feature, it will not perform well on its own. The authors therefore propose to linearly combine the baseline metric with the lexical cohesion metric. Results show that correlation (measured with the Pearson correlation coefficient) with document-level judgements of the combined metric is higher than with the baseline alone. Just the lexical cohesion metric has positive correlation, but is too weak on its own. Furthermore, when looking at correlation against sentence-level judgements, the same ordering of performance is seen, although with a generally lower correlation for all three metrics versions. The main difference to our work is that the authors propose an evaluation metric based on lexical cohesion, whereas our bilingual coherence model focuses on entity-based coherence. Furthermore, their metric only considers the target-side and takes the degree of lexical cohesive devices used as a measure of quality. Our evaluation metric is based on the coherence patterns observed between the source and target side, thus providing more information to the metric. Similarly to our work, the authors identify the need for document-level human judgements to evaluate discourse-aware SMT evaluation metrics. While they focus on discourse relations between sentences, we focus on entity-based coherence.

Wong and Kit (2012) follow up on their work (Wong et al., 2011) again focussing on lexical cohesion for SMT evaluation. They reiterate that lexical cohesion is one

means to make a text coherent and is represented by content words that are used more than once in a document and that are in one of these relations with each other: synonymy, near-synonymy, superordinate relation, repetition and collocation. For determining most of these relations they rely on WordNet (Fellbaum, 2006). The same statistics of content words and various groups of lexical cohesive devices as in their previous work are presented for a new dataset and source language (Chinese-English). They show similar results in that there are more content words in the human translation, and that most of these additional content words are used as lexical cohesive devices (mostly repetition). They experiment with two ratios used as SMT evaluation metrics of lexical cohesion. The first one, i.e. the ratio of lexical cohesive devices over content words (LC) is the same as in their earlier work. The additional one is defined as the ratio of repetitions over content words (RC). The ratios are computed for human translations and automatic translations and compared. Both RC and LC are more stable among different human translations and more distributed among different SMT translations. Furthermore, both LC and RC are almost always lower in SMT translations than in human translations. The distributions of values are very similar in both data sets. This leads the authors to the hypothesis that LC and RC can be used as quality measures of SMT output. Since lexical cohesion is a high-level concept, they propose to integrate it with low-level evaluation metrics. Instead of using a simple unigram-based F-measure of precision and recall (as done before), the authors take established sentence-level metrics BLEU, METEOR and TER. They are then combined as in their earlier work by taking the weighted average of one sentence-level metric and one document-level metric. The weight is optimized on sentence-level human adequacy judgements from the Arabic-English Metrics MATR08 dataset, while testing is performed on the Chinese-English MTC4 dataset, which is also on sentence-level human adequacy judgements. This is contrary to their earlier work, where they explicitly collect document-level human judgements. Generally a lower weight is given to the LC/RC ratios (between 0.18–0.40). BLEU and TER can be improved by integrating the RC or LC ratio. However, METEOR does not benefit from them. At system level there is a high correlation with human judgements for all metrics and combinations, which drops considerably at document level. The major difference to our work and also their earlier work is that the authors no longer verify their modifications to the evaluation metric against the human judgements collected on document-level.

Guzmán et al. (2014) and Joty et al. (2014) propose an evaluation metric that uses discourse structure in form of RST trees at the sentence level. Two similarity measures

are tested that compare the discourse tree structures of a source and a target sentence. This metric is combined with a range of existing discourse-unaware SMT evaluation metrics. The authors note that there is a logical relationship between sentences following each other. They further emphasize that the same relationship also exists between clauses within a sentence. These logical relationships in coherent texts can for example be exposed with discourse analysis based on RST. The authors hypothesise that semantic and pragmatic information encoded in RST trees is beneficial for both SMT in general and its evaluation in particular. An SMT system will benefit from preserving the source-side discourse structure when translating. The evaluation setting is the focus of their work. In Guzmán et al. (2014), experiments are run on the WMT11 and WMT12 metrics shared task data sets for which only sentence-level human judgements are available. Rather than working on the document level, the authors point out that there are enough long sentences in the data that have a rich discourse structure sentence-internally to show that discourse structure can be beneficial for SMT evaluation. Four different languages are part of the experiments, i.e. French, Czech, German and Spanish into English, respectively, however results are only reported by averaging over all languages. Discourse trees in RST consist of elementary discourse units (EDUs) at the leaf-level. Adjacent EDUs are then connected recursively via discourse relations. Furthermore, nodes in the discourse tree are weighted by importance (i.e. *nucleus* for important ones and *satellite* for additional information). Their discourse-aware SMT evaluation metric is either based on a lexicalized (i.e. with words of the sentence as part of the tree) or unlexicalized (i.e. no words) version of the discourse tree. Tree similarity is then checked with subtree kernels (Collins and Duffy, 2001). Finally, the discourse-aware SMT evaluation metric is combined with a wide range of existing evaluation metrics from the ASIYA toolkit (Giménez and Màrquez, 2010). These metrics range from lexical (including BLEU and METEOR) and syntactic information to semantic information. The effectiveness of the proposed SMT evaluation metric is verified by computing the correlation with human sentence-level judgements. These judgements are taken from the WMT11 and WMT12 news translation shared task, where system outputs have manually been ranked. In total four different kinds of experiments are run to test for correlation with human judgements of the proposed variants of the SMT evaluation metrics. The first experiment evaluates the metrics on *system level*. Both proposed discourse-aware evaluation metrics on their own have a high correlation and improve existing evaluation metrics when combined with them in most of the cases. The lexicalized proposed metric performs better than the un-

lexicalized one. In the second experiment correlation is tested at *segment level*. The unlexicalized metric performs very poorly and the lexicalized one only slightly better. Both metrics on their own are outperformed by most of the existing metrics. However, the lexicalized metric when combined with the existing metrics consistently improves correlation. The metric combinations in the above two experiments are done by an unweighted linear interpolation. In the third and fourth experiment human judgements are used to tune weights for the metric combinations. These experiments are only performed on segment level, both in a cross-validation setup and on two different data sets. Weight tuning is done within a pairwise learning-to-rank framework with human sentence-level judgements as gold rankings. Tuning weights results in a bigger improvement of the combined metrics over individual ones than their untuned counterparts. Overall, experiments show that using discourse structure in SMT evaluation can be beneficial. Contrary to our work, the authors focus only on sentence-internal phenomena. Furthermore, they employ a wide range of existing evaluation metrics to be combined with their proposed metric, while we focus only on information provided by the discourse level. Similarly to our work, they tune weights for the metric combination in a learning-to-rank framework based on human gold rankings. We tune individual components of our SMT evaluation metric based on human gold labels at the document level within the same framework. Furthermore, we also exploit information from the source document with respect to the target-side and do not compare system output to reference translations.

Joty et al. (2014) describe their system submission to the WMT14 metrics task, which is based on the work presented above. They manage to achieve the best results in the task with the tuned evaluation metric consisting of the combination of ASIYA metrics and discourse-aware metrics. They note that especially their unlexicalized metric results in many ties when ranking two systems at segment level against each other. They propose a tie-breaking heuristics proportional to system level scores of their own metric. Ranking ties are also a major problem with our metric and remain an important part that needs to be explored and tackled with discourse-aware SMT evaluation metrics.

Gong et al. (2015) also identify the need for document-level SMT evaluation metrics in order to find out if document-level models for SMT are beneficial. They propose two evaluation metrics. One is based on the overall consistency of a translation with respect to reference translations. The other one is based on the degree of cohesion again comparing it between the automatic and reference translation. The first metric,

capturing the gist consistency, is realized with the help of monolingual topic models. These models provide a set of topics for a document and can be seen to capture the main idea of a document. A monolingual topic model is trained with 120 topics. The document-topic distributions are then inferred from a system output and from the reference translation on document level. The Kullback-Leibler divergence (Kullback and Leibler, 1951) is then used to determine how close these two document-topic distributions are. Based on initial experiments the authors found out that using the full distribution leads to worse performance. They therefore suggest to only take important topics into account, where importance is determined by the probability of a particular topic given the document above a certain threshold. Since multiple reference translations are available, the final gist consistency is the score that is the lowest between the system output and any of these reference translations. The second metric is capturing the cohesion of a document. In particular, the number of matching lexical chains between a system output and a reference translation of a document is taken as a measure of cohesion. Lexical chains are extracted based on a simple matching algorithm. All nouns that have the same stem are grouped into one lexical chain. In order to give credit to partially matching lexical chains between system output and reference translation, the number of matching members of the lexical chain is counted and divided by the total number of members in the reference translation. This score is aggregated over all matching lexical chains. If a lexical chain does not have a counterpart in the reference translation, this is penalized by dividing the above score by the total number of lexical chains in the system output of the document. Again, the best total score is used for a document, comparing one system output against each reference translation. The authors point out that their formulation is in contrast to Wong and Kit (2012) in that the latter work only attempts to capture cohesion on the target-side without comparing it to the reference translation, following the hypothesis that a better translation has more content words that are part of lexical chains. To test whether their evaluation metrics are beneficial for SMT evaluation, they combine each proposed metric with one of the two existing metric BLEU or METEOR. The combination is a simple linear combination with weights tuned on a development set. Experiments are performed for Chinese-English on two different data sets. Each of these data sets contains four reference translations and between three and six different automatic translations from different SMT systems. These system outputs also come with sentence-level human judgements that rated adequacy and fluency. To obtain document-level human judgements, the sentence-level judgements are simply averaged over the document. To

determine whether the proposed metric combinations are helpful, correlation against human judgments is measured with Pearson and Kendall correlation coefficients. All four evaluation combinations result in an increase of correlation with human judgements on both tested data sets with respect to BLEU and METEOR on their own. The metric combination using gist consistency performs better. This is also confirmed with the learnt weights of the linear combination, which are similar for the gist consistency and BLEU or METEOR, but almost zero for the score based on cohesion. One of the major differences to our work is that this work evaluates performance of the proposed SMT evaluation metrics against human judgements that were elicited sentence by sentence. These human judges could therefore not take into account the overall adequacy of the translation of the entire document. An uncommented problem in their result is that the correlation of BLEU with human judgements on one of the data sets is almost zero, suggesting some problem either with BLEU or with the judgements on this data set. Furthermore, it is not explicitly mentioned how lexical chains between system output and reference translations are matched.

Luong and Popescu-Belis (2016) present work on integrating pronoun antecedent information into the decoding process of a phrased-based SMT system for English-French. They also present an automatic evaluation metric, APT, that focuses on evaluating the translation of pronouns. In a manual evaluation they identify problems with this automatic metric, nevertheless showing a correlation with human judgements. Results of the proposed pronoun-aware decoder show that it outperforms a phrase-based baseline in terms of these metrics, but not in terms of BLEU. Work on handling pronoun translation in a full SMT system failed to improve over the baseline of the DiscoMT15 shared task on pronoun translation (Hardmeier et al., 2015). The authors point out that one of the problems might be the quality of coreference resolution systems. These systems are required to identify the antecedent of a pronoun, which bears relevant features (e.g. number and gender). From a large parallel corpus where coreference is automatically resolved, they learn a distribution of pronouns given antecedent features of these uncertain antecedents. First, pronoun-antecedent pairs are automatically extracted on the source side. Via word alignments, these pairs are projected to the target side and features from the target-side antecedent are extracted. A probability distribution on the target side over pronouns given these features is estimated with relative frequencies from a large parallel corpus (IWSLT). This distribution is integrated into the phrase-based decoding process as a back-off translation model, which is only applied if a relevant source pronoun is identified. At decoding time, sentences

with sentence-internal antecedents are first translated without the additional translation model. Then antecedent features are extracted. In a second translation of the same sentence, the additional translation model is now used, based on the previously extracted antecedent features. Pronouns with antecedents in preceding sentences can be processed directly with the additional translation model. A new evaluation metric for assessing the accuracy of the translated pronouns is presented. It is inspired by the ACT metric (Hajlaoui and Popescu-Belis, 2013) that handles evaluation of discourse connectives. The pronoun-focussed metric inspects a source pronoun to find out if it is translated correctly. This is done by following word alignments from the source to both the reference translation and the candidate translation. A pronoun is considered to be correctly translated if the aligned target-side tokens are both identical or if they both belong to the same set of equivalent pronouns. The metric score is then the ratio of correct translations over all pronoun instances in the test set. The translations of pronouns are also evaluated by humans. They are given the source and reference sentence that contains a pronoun with one preceding sentence respectively. Furthermore, they get the output of the translation systems in random order. The source-side pronoun to be evaluated was specially highlighted. The annotators should judge the correctness of the pronoun translation according to the antecedent in the target side (and not according to the source-pronoun). Kappa's inter-annotator agreement is 0.65, based on a subset of overlapping annotations. A higher agreement of 0.94 between the two annotators could be reached after resolving disagreements. Experiments are run on the test set of the DiscoMT15 shared task. Due to coreference or alignment errors only about 50% of the pronouns in the test set are processed by the proposed translation model, the others are handled by the standard translation model. The BLEU score does not show any difference in performance. However, the extended translation system outperforms a phrase-based baseline in terms of the proposed automatic metric. They also evaluate in terms of the pronoun-focussed evaluation metric from Hardmeier and Federico (2010), which provides a similar picture. According to the human annotators, the proposed system is also better. The accuracy scores (automatic and human) have a high correlation. However, the human-provided accuracy scores are much higher (20% absolute) compared to the proposed automatic pronoun metric. This difference is due to the fact that the proposed metric counts pronouns as incorrect if they do not match the reference pronoun, despite the fact that they agree with their actual antecedent in the system output. The major difference in their proposed evaluation metric compared to ours is that it does not take the antecedent of the actual translation into account.

The authors provide evidence that this is problematic, leading to many false negatives. Furthermore, this evaluation metric only focuses on evaluation of pronoun translation, whereas we attempt to capture the entire entity-based coherence of which pronouns are a part.

Guillou and Hardmeier (2016) present a test suite containing 250 hand-selected pronouns for evaluating SMT systems. They also provide an automatic method to compare pronouns from an SMT system output with a reference translation. This test suite is mostly designed for obtaining a deeper understanding of the pronoun translation performance of SMT systems, so non-matching pronouns are referred to human annotators to make a final decision. The data set is the DiscoMT2015 test set (Hardmeier et al., 2015) for English-French. The source side is manually annotated in the same style as the ParCor corpus (Guillou et al., 2014). Referential pronouns are linked to their closest non-pronoun antecedent. They are further annotated by their function. The test suite consist of full documents, however only a selected number of pronouns is used for evaluation. To obtain this selection, pronouns are grouped by their function and according to typical problems in pronoun translation. From each group a number of pronouns is selected that is proportional to the number of ways the respective pronouns might be translated into French. Only the source-side pronouns *it, they* and *you* are considered. An automatic evaluation script compares system pronoun translations with the reference pronoun translation. Word alignments between source and reference translation, and source and system translation is required. If the pronoun is anaphoric, then the antecedent is also checked against the reference translation. Accuracy for each pronoun group and for the entire test suite is produced. In a brief analysis from system outputs of participating systems in the DiscoMT15 pronoun translation shared task, the authors could identify a correlation between better performance of systems on those pronoun groups that were explicitly handled by the systems, and a poorer performance on pronouns not taken care of by the systems. The proposed evaluation script is very similar to the one from Luong and Popescu-Belis (2016). The main difference is that Guillou and Hardmeier (2016) take the reference antecedent into account when computing a match. On the other hand, Luong and Popescu-Belis (2016) allow for greater flexibility when it comes to accepting translation variants. Unlike our work on SMT evaluation, this test suite requires manual annotation of pronouns and their antecedents in the source documents. It furthermore relies on the correct identification of the target-side antecedent based on the gold annotated source-side referent and automatic word alignments. Any errors in this projection would result in the evaluation

script penalizing a system. Furthermore, a noun antecedent in the source document, might not necessarily correspond to a noun antecedent in the target side.

2.4 Entity-based Coherence Modelling

Barzilay and Lapata (2008) present a framework for modelling local entity-based text coherence referred to as the entity grid model (EGM). It is inspired by existing linguistic theories, such as Centering Theory (Grosz et al., 1995) and other entity-based theories, but does not attempt to implement a particular one. Additionally, the emphasis of the framework is on automatic computability of required representations. The basic assumption of the EGM is that entity distributions in locally coherent texts follow regular patterns. These distributions of patterns can then be used to train classifiers for text ordering, evaluation of the coherence of automatically created summaries and readability assessment. A text is first transformed into an abstract *entity grid (EGrid)* representation which records occurrence and absence of entities in sentences. If an entity occurs in a sentence, its syntactic role is specified (S for *subject*, O for *object* and X for *other*). Only one entity mention per entity in a sentence is recorded. The result is a matrix with rows representing sentences and columns representing entities. This representation is then converted into a feature vector which holds distributional properties of n-place entity transitions (e.g. *none-subj* or *subj-obj-none*). These vectors are then used to train classifiers for the above three tasks, which learn to rank more coherent documents higher than less coherent documents (both in terms of artificially created data and according to human rankings) in the first two tasks and learn to predict readability scores for the third task. The first experiment is a sentence ordering task.[3] The sentences of a coherent document have been shuffled to produce documents that are less coherent than the original one. A ranking function is then learnt such that it ranks the most coherent document (the original one) highest and the artificially created ones lower. For evaluation, accuracy is measured, which is the ratio of times a pairwise ranking ranked the original document higher than a shuffled one, divided by the total number of pairwise comparisons. Different model variants are explored (with or without coreference resolution, syntactic roles and salience). Results show the more linguistic information is included in the model, the better it performs. Per-

[3]The sentence ordering task does not try to restore the correct order of a shuffled text, but the task is restricted to ranking a number of shuffled documents lower than the original document in terms of coherence.

formance varies across different domains (newspaper articles vs. accident reports) and the three different parameters vary in usefulness or unimportance. One baseline is outperformed on both domains, whereas a second baseline (which is lexicalized and captures global coherence) is slightly better on one domain. Since complementary information seems to be modelled, the authors suggest to combine both types of models for further benefits. The second task is similar, however this time the shuffled documents are replaced by documents generated by various text summarization systems and graded by humans as to how coherent they are. Furthermore, the original (i.e. coherent) document was supplemented by human generated summaries, which are assumed to be highly coherent as well. In this way the EGM can be evaluated against human judgements of pairwise coherence ratings. As in the previous experiment, using more linguistic information results in better models. However, this time results show that when using the more sophisticated automatic coreference resolution results are worse compared to the simple noun-identity matching coreference resolution version. This is attributed to the fact that machine generated summaries are noisier and pose more difficulties to the automatic coreference resolution system. It is also observed that automatic summaries use pronouns less frequently than humans, so an easier noun matching resolution model is sufficient. The third experiment tested whether the EGM would be useful for improving readability assessment, which means determining how easy or difficult it is to read a particular document. They treat this task as a classification task with binary labels (easy vs. hard to read) and use a Support Vector Machine (SVM) (Joachims, 2002) as classification model. As data articles from the Encyclopedia Britannica (hard) and their equivalent versions written for children (easy) are used. The coherence feature vector was added to baseline features established in Schwarm and Ostendorf (2005). Results show big improvements when coherence features are added. However, coherence features on their own were not sufficient and performed worse than the baseline system.

Guinaudeau and Strube (2013) reformulate the above model from Barzilay and Lapata (2008) and represent the EGrid in a bipartite graph with sentences and entities as the two mutually exclusive node sets. They call it the *entity graph (EGraph)*. An edge between an entity and a sentence records a mention of that entity in that sentence. The syntactic roles are encoded as edge weights where S, O, X are assigned 3, 2 and 1 respectively. The bipartite graph is then projected to a graph with just the sentence nodes (i.e. one-mode projections) where an edge between two sentence nodes is established if an entity is mentioned in both sentences. The EGraph only captures entities

that cross sentence boundaries, since singleton entities do not influence the resulting one-mode projection. Furthermore, entities that are not mentioned in a particular sentence are no longer explicitly represented. Three different one-mode projections are formulated: P_U where an edge with weight 1 exists if at least one entity is mentioned in both sentences; P_W where the edge weight is the count of how many different entities are mentioned in both sentences; and P_{Acc} where the counts are weighted by the syntactic role weights. Instead of learning a ranking function Guinaudeau and Strube (2013) show that they can use a score derived from the one-mode projections directly. They use the average out-degree (AOD) which is defined as the sum of the weighted edges in the one-mode projection divided by the number of sentence nodes.

The above two models inspire our bilingual model of coherence and the SMT evaluation metric based on this model. We take over the abstraction of recording only one entity mention of an entity in a sentence. Furthermore, similarly to the EGraph, we do not record singleton coreference chains or coreference chains with mentions in just one sentence. The major difference is that our setting is bilingual and we attempt to model the patterns of coherence across the two languages. We furthermore define experimental tasks for the bilingual setting with an artificially created corpus with automatically obtained gold labels and on a realistic data set with gold labels extracted from human annotation.

Sim Smith et al. (2015) point out that if discourse-level problems in SMT are to be tackled, labelled data representing coherence violations is required, but it does not yet exist. Coherent data is available everywhere and output of current SMT systems could be considered incoherent. However data of the latter type also exhibits many other errors that are not related to coherence at all. This would make it difficult to assess any advances in coherence modelling for SMT specifically. The authors further note that manually annotating an automatically translated corpus with respect to coherence or errors thereof is a hard task in terms of formalizing the annotation, annotation cost, and time. In monolingual settings, coherence has been modelled in various ways. These models are either evaluated against human judgements of coherence (e.g. judging automatically created summaries). Alternatively, they are evaluated on data that has artificially been made incoherent by shuffling the order of sentences in a document. The proposal in Sim Smith et al. (2015) follows the latter approach and sketches a plan to create a corpus where coherent target-side documents are taken as a starting point. These are then made incoherent by introducing coherence errors based on distributions of typical coherence errors found in existing translations and based on

linguistic insight. In a corpus analysis and based on previous work that tackles issues of discourse in SMT, the authors present the following common errors related to coherence. *Lexical cohesion*, as part of the overall coherence encourages consistent translations of lexical items. However, if lexical chains are wrongly identified, this gives the wrong translation incentives, enforcing consistent translation of unrelated lexical items, which leads to incoherent texts. *Anaphora resolution* is a difficult problem for SMT. Antecedents might occur outside of the sentence context and languages differ in their usage of gender and number agreement between antecedents and referring expressions. This might result in incorrect or missing translations of referring expressions, which has a direct impact on coherence. *Discourse connectives* might be implicit in the source language and therefore missing in the translation, and they can also be ambiguous. Both issues might result in incoherent translations. *Syntax and clause reordering* is also important in translation as too much or to little reordering might render the sentence incoherent. Finally, missing or wrong *negation* on the target side introduces incoherent translations. With a set of operations, i.e. replace, delete, add, and shift, errors can be introduced in the original document and the degree of error introduction can be controlled for. Constructing corpora with the above described errors is left for future work. It therefore remains to be seen how beneficial this proposal is. In our experiments on bilingual coherence modelling we also employ the idea of introducing artificial errors to obtain automatic coherence judgements. However, we base our error introduction on the distribution of errors made by CLPP systems. This means we only focus on the issue of referring expressions. However, we can introduce errors automatically and do not have to manually define the errors that are introduced, while also controlling for the degree of confusions introduced.

Sim Smith et al. (2016) present work on measuring coherence of SMT output using monolingual coherence models. They evaluate the performance of these models on a standard sentence shuffling task, where in a coherent document, the sentences have been randomly shuffled. More importantly, they present a new task where the coherence has to be assessed comparing automatically translated documents against their human reference translation. They first present three existing monolingual coherence models. First the EGrid model (Barzilay and Lapata, 2008) and second the EGraph model (Guinaudeau and Strube, 2013) both presented above. The third model is by Louis and Nenkova (2012) which exploits the assumption that in coherent texts, two adjacent sentences contain similar syntactic patterns. However, patterns between the two sentences are not mapped to each other via alignment information. Sim Smith

et al. (2016) therefore propose an extension to this model where they add this alignment information as latent variable, following the intuition that certain pattern pairs are more likely in adjacent sentences. Experiments are run on the WMT14 test data set for German, French and Russian into English, respectively. For the shuffling task only the human reference translations are used. For each document, one random permutation of sentences is produced. The former documents are considered to be coherent, the latter documents are considered to be incoherent. For the translation task, the human reference translations are used as the coherent documents. As incoherent documents, all outputs of the participating SMT systems from the WMT14 translation task are taken. This follows the assumption that these translations systems produce text that is less coherent than the human reference translation. Three different evaluation scores are computed. The first one counts how often the human reference document is ranked strictly higher than the incoherent documents. The second one also includes in these counts any rankings where the model could not differentiate between human reference and incoherent documents. The third one, only applicable to the translation task, counts how often the human reference is strictly ranked highest over all translations. The results on the shuffling task show that the proposed enriched model of coherence based on Louis and Nenkova (2012) outperforms the original model considerably. The performance is comparable to the EGrid model. Results on the translation task show that the scores are much lower than for the shuffling task, since the former task is more difficult. As before, in the translation task the proposed model always outperforms its original counterpart. The best performing model can score the human reference higher than the SMT output in between 58% and 67% of the cases. There is no clearly best performing model over all languages and all evaluation scores. The evaluation score drops considerably if the human reference has to be ranked as coherent against all SMT translations for a given document. There the best performing models score between 20% and 28%. Factors that contribute to different performances are the quality of the reference translation or the closeness between the source and target language. The major differences to our work are that their models of coherence are only assessing coherence on the target side, disregarding the coherence patterns from the source document. While they investigate different language pairs, the different source languages only indirectly influence their monolingual coherence model scores in that the SMT output tends to be closer to the source document and therefore the translation tends to contain patterns different from human translations. Furthermore, they do not take coreference into account, which is the major source of information for the monolingual

models and our bilingual model.

2.5 Coreference Resolution Evaluation

Vilain et al. (1995) define an evaluation metric for comparing the performance of automatic coreference resolution systems against a gold standard annotation. It is a precision and recall metric that is defined over the equivalence classes which are obtained by grouping all mentions that are coreferent into one set.[4] For recall, the difference between system (i.e. response) and gold (i.e. key) classes is then defined as the minimum number of links that need to be added between mention groups to create the gold class from the system class. To compute precision, the roles of the gold and system equivalence classes are swapped. This metric is hence defined over equivalence classes, and does not compare links between individual entity mentions. For both scores, the equivalence classes of system and gold standard are extracted. For recall, assume one class of the gold standard is defined as S, and $R_1, R_2, \ldots, R_m$ are all the equivalence classes defined by the system response. Intersecting S with all these R_i partitions S into sets of mentions, each being the result of the intersection with a specific R_i. Each remaining element in S that was not affected by any intersection forms a singleton set in the partition (i.e. a set with just one mention). For recall, these singleton sets contain the mentions that are annotated in the gold standard, but which were not identified by the system, as belonging to the coreference chains denoted by S. All these partitions together are referred to as the set of partitions $p(S)$. Recall for one gold equivalence class S is then defined as $\frac{|S|-|p(S)|}{|S|-1}$. The numerator is the minimum number of links that is necessary to add to the system response (i.e. to the partition of S with respect to the response equivalence classes R_i) in order to create the equivalence class S. The denominator is the minimum number of links between gold standard mentions in S that is necessary to create the equivalence class S. To extend that to an entire test set, one can sum over all gold equivalence classes as follows $\frac{\sum|S_i|-|p(S_i)|}{\sum|S_i|-1}$. As mentioned above, to compute precision, the sets which define S and R_i are simply swapped between system and gold sets. This metric is referred to as the *MUC score* in the literature.

Bagga and Baldwin (1998) identify two shortcomings with the MUC6 scorer (Vilain et al., 1995). First of all it does not give any credit for correctly identifying single-

[4] Vilain et al. (1995) use terminology in a different way than we do here: They call noun phrases *entities*, we call them *mentions* or *entity mentions*. One coreference chain forms an equivalence class, which we also call *entity*, i.e. the entity to which all members of the chain (or equivalence class) refer. Hence, our *entity* is *not* to be confused with the *entity* in the described paper.

ton entities, since the equivalence classes are expected to contain at least two mentions (the precision and recall formulae would be undefined, i.e. division by zero, if singleton entities were allowed). Second, all types of coreference resolution errors are considered to be equivalent. According to the authors this is counterintuitive. For example, if two long coreference chains are wrongly linked by the system, this is a more severe error than wrongly linking a long chain with a short one (i.e. less entities are wrongly linked together).[5] Instead of comparing how much system entities partition gold entities (or vice versa), which is the way the MUC6 scorer is defined, the B-cubed metric is mention-specific and computes precision and recall for each mention with respect to all other mentions in an equivalence class. Precision and recall are therefore defined for each entity mention m_i. For a specific m_i the number of correct entity mentions in a system output chain containing m_i is counted. For precision, this count is divided by the total number of mentions in that system output chain. For recall, this count is divided by the total number of mentions in the gold chain. The final precision and recall scores are then defined as the weighted sum over all mention-specific precision and recall scores in a document. This metric is referred to as the B^3 or *B-cubed score* in the literature.

Luo (2005) point out that entity mentions can be used for matching more than once in the above evaluation metric, thus artificially inflating the performance. They therefore propose CEAF, the constraint entity-alignment F-measure. It aligns entities from the gold standard and the system output with the constraint that each entity in the gold standard can only be aligned at most once with one in the system output, and vice versa. The alignment is optimized such that it maximizes entity similarity, while following the constraint that entities can only be aligned once. For finding this optimal entity alignment, the Kuhn-Munkres (KM) algorithm is used. The above setup requires a definition of similarity between two entities. Luo (2005) provide two suggestions. The first one is a mention-based similarity measure, which simply counts how many entity mentions from a system chain and a reference chain match. In other words, this reflects the ratio of mentions that are in the correct chain. The alignment that produces the maximum of this score over all system and reference chains is searched. The found score is then divided by the total number of mentions in all system output chains (for precision) and divided by the total number of mentions in all reference chains (for recall). The second similarity measure is an entity-based similarity measure, which

[5] Bagga and Baldwin (1998) use the same naming convention as in (Vilain et al., 1995), which is different from ours. See footnote 4.

also counts how many entity mentions from the system chain and the reference chain match, but in addition normalizes this count by the number of entity mentions within system and reference chain. In other words, this reflects the ratio of correct entities. Again, the alignment that produces the maximum of this score is searched. The found score is then divided by the total number of system chains (for precision) and the total number of reference chains (for recall). The F-measure of precision and recall is then taken as the final evaluation score. It automatically penalizes systems that generate too many chains (which lowers precision) and systems that generate to few chains (which lowers recall). When using the first similarity measure, this produces mention-based CEAF (or *CEAF-m*), and the second one produces entity-based CEAF (or *CEAF-e*).

These evaluation metrics all compare gold coreference chains against coreference chains hypothesized by a coreference resolution system. From a high-level perspective this setup is related to our setup in the SMT evaluation metric based on bilingual coherence in that it is based on comparing entities (i.e. coreference chains) from the source document with the ones occurring in the target document. From one angle this could be viewed as the source-side entities being the gold standard coreference chains, and the target-side entities being the coreference chains hypothesized by a coreference resolution system. In fact Miculicich Werlen and Popescu-Belis (2017) take this view in their experiments. They first project the target-side coreference chains to the source-side and then compute the above three presented coreference evaluation metrics. They then use this score to rerank translation hypotheses that maximize the scores. Their work however relies on gold-standard annotations of coreference chains on the source side.

This is the major difference to our work. To define a scalable SMT evaluation metric, we wanted to base it on automatic coreference resolution systems (in both languages from the language pair involved). However, these systems make errors themselves such as wrongly linking mentions into one coreference chain, or missing a link. We therefore can no longer assume that we have a gold standard on the source side. The coreference resolution evaluation metrics, however, were designed to maximize the matching between gold standard and system output, so they are not directly applicable.

Our SMT evaluation metric is nevertheless inspired by these metrics. First of all, we also use the KM algorithm to find the optimal one-to-one alignment between source-side and target-side entities. However, instead of inflexibly punishing coreference chains that are too short or too long on the target-side (with respect to the source),

we provide this information as a feature in a feature vector, whose weights then can be tuned. In that way the errors of the coreference resolution systems can be taken into account.

The CEAF metric offers two versions, one which focuses on the performance of each mention, and one that compares entities as a whole. In our bilingual EGM for SMT evaluation we also have an entity perspective and a mention perspective. In the first one, we record how many entities have matching counterparts and how many remain unalinged, and in the second one, we count how many mentions are inserted, deleted or have a matching counterpart in both languages. Contrary to the CEAF metric, we integrate both views into our feature vector, so that they can both contribute to the assessment of bilingual coherence.

Another major difference between these coreference evaluation metrics is that we do not take each single entity mention into account, when aligning the entities. Inspired by the monolingual EGM and the salience of entity mentions, we only take the most salient entity mention per sentence into consideration. This also provided us with an abstraction that is useful for abstracting away from minor language-specific details. Furthermore, we do not take singleton chains into account, since they do not contribute in the coherence of a document.

Finally, none of the above three papers explain in detail, how exactly it is determined that a gold entity mention and a system entity mention are the same. The implicitly assumed setup is most likely that gold markables, i.e. spans of potentially referential or non-referential mentions, are known to the coreference resolution system. In that case it is then trivial to determine equality, since the participating entity mentions would always have the same span in gold annotation and system output. Other possibilities are also conceivable, such that mentions are deemed to be equal, if their semantic head words are identical, irrespective of any surrounding words in the markable span. Another definition could take the word overlap of both spans into account. Any of these approaches do not work directly in the cross-lingual setup. The least requirement would be to have word alignments, so that markable span correspondences could be mapped from one language to another. This is the approach that Miculicich Werlen and Popescu-Belis (2017) take, where they then rely on heuristics to determine the markable span mapped from the target-side onto the gold annotations of the source side. In our work, this is not directly applicable and necessary, since each coreference chain only records one mention within a sentence. Furthermore, we do not want to prescribe that the markable spans across two languages have to be exactly the

same. This should be verified in a parallel corpus annotated for coreference in both languages.

Chapter 3

Pronoun Translation for SMT with CLPP

In this chapter, we focus on the problem of automatic pronoun translation, as an instance of discourse-level phenomena that SMT systems have to deal with. We first present the two shared tasks on cross-lingual pronoun prediction (CLPP) which were set up in order to explore and better understand the problem of pronoun translation and to establish a basis for researchers to work on it (Section 3.1). We then present both CLPP systems we submitted to each of the shared tasks (Wetzel et al., 2015; Wetzel, 2016). The first one handles the English-French language pair (Section 3.2) and experiments show that the target-side features are more important than source-side features, confirming linguistic knowledge. More specifically, they also show that having access to the antecedent of a referential pronoun and the grammatical features of the antecedent is beneficial for performance. Our second system validates that the taken approach also generalizes to English-German and is still suitable with the impoverished target-side representation from the CLPP16 shared task (Section 3.3). In that section, we also explore the use of a feature that learns to predict the empty word showing its benefit, and experiment on integrating sequence information into the prediction with negative results. Finally, we discuss the impact of and draw conclusions from both our systems and the conducted experiments (Section 3.4).

3.1 Pronoun Translation and the CLPP Shared Tasks

Translation of pronouns is a non-trivial task due to ambiguities in the source language (event pronouns, referential and non-referential uses) and due to diverging usage of

pronouns between two languages (e.g. morphological differences including gender and number and differing agreement constraints between a pronoun and its antecedent, or pro-drop languages that leave subject pronouns implicit, cf. Section 2.1.2). In the recent past there has been work on analysing these differences and there are various approaches to tackle the problem for SMT (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012; Weiner, 2014; Hardmeier et al., 2014; Guillou et al., 2014).

The DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015) called for contributions to tackle this problem for English-French.[1] It consists of two sub-tasks, i.e. a full translation task where evaluation focuses on the performance of the translation of pronouns, and a CLPP task which worked with a fixed human-authored target-side, enabling automatic evaluation. We focus on the latter task. A second instance of this shared task was offered at WMT 2016 (Guillou et al., 2016), this time only offering the CLPP shared task.[2] In addition to the English-French language pair, it introduced data sets for English-German, as well as the inverse translation directions from French and German into English. Throughout this thesis we will refer to these tasks as the CLPP15 and CLPP16 shared tasks.

3.1.1 Shared Task Setup

In the CLPP shared task pronoun translation is treated as a classification task. Given a set of source-side pronouns, a classification into a closed set of target-side pronouns (or classes) has to be made for each of these source-side pronouns. Training and test data consists of full parallel documents with aligned sentences. In each document the entire source-side is given. Each parallel document also comes with word-alignments. On the target-side pronouns were removed and replaced by special REPLACE tokens. These mark the position where a particular target-side pronoun occurs. Furthermore, the REPLACE token provides a reference to the position of the source pronoun it corresponds to. The training data additionally provides the original pronoun that each REPLACE token substituted, as well as the official class the pronoun is assigned to. Example (1) shows one pronoun instance *it* taken out of a document with gold label *elle* as follows:

[1] https://web.archive.org/web/20151202210858/http://www.idiap.ch/workshop/DiscoMT/shared-task

[2] http://www.statmt.org/wmt16/pronoun-task.html

English source pronouns	it, they
French target pronouns CLPP15	ce, elle, elles, il, ils, ça, cela, on, OTHER
French target pronouns CLPP16	ce, elle, elles, il, ils, cela, on, OTHER
German target pronouns CLPP16	er, sie, es, man, OTHER

Table 3.1: Overview of source- and target-side pronouns involved in the CLPP shared tasks for English-French and English-German in 2015 and 2016.

(1) *Source*: He said he 'd been listening to the symphony and *it* was absolutely glorious music [...]
Target: Il raconta qu' il avait écouté une symphonie et qu' REPLACE_{10} était fabuleuse [...]
Gold label: *elle* – CLPP15 development set

Not all pronouns on the source and target-side are of interest in the shared task. Focus is put on a small set of potentially ambiguous pronouns on the source side, and a manageable set of pronouns on the target side (so as not to provide too many class labels for learning classifiers). The set of English source pronouns for all tasks are the same, i.e. *it* and *they*. The set of target side pronouns of the first shared task for English-French is the following: *ce, elle, elles, il, ils, ça, cela, on*, OTHER. For the second shared task the two pronouns *ça* and *cela* were merged into one class. The English-German target-side pronouns are *er, sie, es, man*, OTHER. An overview is provided in Table 3.1.

The OTHER class groups together any translation of the source pronouns that does not fall into one of the defined classes, i.e. less frequent translations. It also includes cases, where the source-side pronoun does not have a target-side counterpart according to the provided word alignments. We refer to this special case as NONE. Furthermore, it also captures alignment errors, where the source pronoun should have been aligned to an existing target side pronoun.

The target side of both CLPP shared tasks consist of human reference translations. This makes it possible to study and tackle the problem of pronoun translation independently of further SMT processing. It abstracts away from additional difficulties that arise with noisy automatic translations. One of the major differences in the CLPP16 shared task compared to the one from 2015 is the target-side data. It comes in the form of lemmatized tokens with their POS tag, instead of the full word forms. This makes

the task more challenging, since agreement features of words surrounding a pronoun are no longer available. For example all determiners are mapped to one generic form irrespective of their gender or number. Guillou et al. (2016) also argue that it makes the task more realistic, when considering SMT as the driving goal. SMT systems do not necessarily produce the correct target-side surface word forms and approaches to pronoun translation should not rely on error-free translations of the relevant context. This change therefore attempts to obtain models that can better handle noisier or underspecified input.

The shared task training data consists of the Europarl corpus (Koehn, 2005), the NewsCommentary corpus[3] and the IWSLT corpus (Cettolo et al., 2012), which consists of TED talks.[4] The development and test data are also TED talks, which generally are long documents (between 48 and 246 sentences) with a higher frequency of pronouns than news article data commonly used in SMT experiments (cf. Ruiz and Federico, 2014; Guillou, 2016, Chapter 4).

3.1.2 Shared Task Baseline

The baseline system for the shared task (Hardmeier et al., 2015, Section 4) predicts pronouns based on a LM. The idea behind it is to choose for each sentence the pronoun out of the set of class labels that maximizes sentence probability according to the LM. The LM is a 5-gram model with Kneser-Ney smoothing. It is trained using KenLM (Heafield, 2011) and the data sets from the shared task: IWSLT, Europarl, NewsCommentary and news data from WMT2007-2013. The trained Kneser-Ney LM is also provided as part of the shared task data.

In detail, the baseline works as follows. If there is only one REPLACE token in a target sentence, then the baseline first substitutes this token with one of the class labels, and computes the probability of the entire sentence with the LM. This is done for each class label that represents an actual pronoun.

The OTHER class is treated slightly differently, since it cannot be directly substituted into the sentence. The LM has not seen this token during training, and would therefore penalize the overall sentence probability. Instead, the most frequently aligned target-side words to source-side pronouns (excluding the above class labels), are substituted.

[3] http://opus.lingfil.uu.se/News-Commentary.php

[4] https://www.ted.com/

Furthermore, a prediction is made computing the sentence probability, when substituting the REPLACE token with the empty string (referred to as NONE). Since n-gram LMs generally give higher probabilities to shorter sentences, a tunable penalty is added to each NONE prediction, discouraging the baseline to predict NONE too frequently.

The class label, the most frequently aligned word, or the NONE token that resulted in the sentence with the highest probability is then taken as the prediction of the baseline for that pronoun instance. In the latter two cases, the prediction is first mapped to the OTHER class.

If there are multiple pronoun instances in a sentence, the best combination of REPLACE token substitutions is searched. Due to combinatorial explosion the search space can get very large if there are many REPLACE tokens in one sentence. Therefore, the search space is pruned if it exceeds 5000 sentence queries and only the 200 sentences with highest probability are retained.

The LM is purely oriented on the sentence level and does not take a wider context into account. It therefore has the same shortcomings as full SMT systems. Furthermore, it only considers the target-side when making the pronoun predictions.

3.2 A Maximum Entropy Classifier for CLPP

In this section we describe our submission to the DiscoMT 2015 shared task on CLPP for English to French. Our approach builds on a MaxEnt classifier that incorporates features based on the source pronoun and local source- and target-side contexts. Additional, discourse context is taken into account by extracting features from the target-side noun referent (i.e. the *antecedent*) of a target-side pronoun.

A MaxEnt classifier can model multinomial dependent variables (discrete class labels) given a set of independent variables (i.e. observations). Each observation is represented by a set of m features extracted from the observation. The m features can provide overlapping evidence, hence do not have to be independent of each other. The model consists of a function $f(x_i, y_i) \rightarrow \mathbb{R}^{m+1}$ that maps the i-th observation x and associated label y to a real valued vector. It also has a weight vector $\vec{\theta}$ of corresponding size, which contains the model parameters that are learned from the training data. The model is of the form

$$p(y|x) = \frac{exp\ \vec{\theta} \cdot f(x,y)}{Z(x)}$$

where $Z(x)$ is a normalizing factor ensuring valid probabilities.

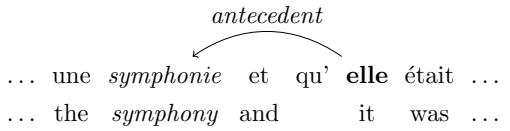


Figure 3.1: Antecedent of a pronoun within local context, which is also captured by a 5-gram language model.

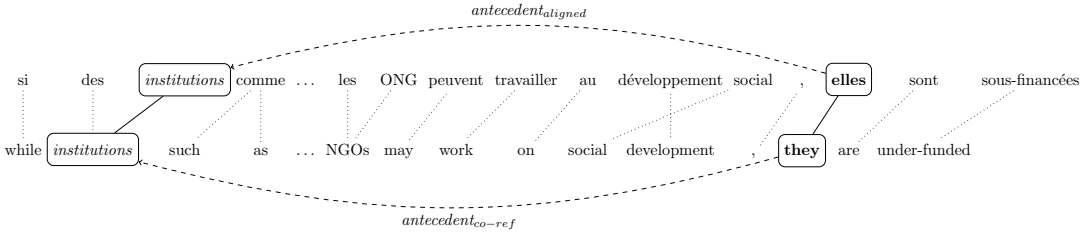


Figure 3.2: The lower sentence is in the source language (English), the upper on is in the target language (French). The $antecedent_{coref}$ of *they* on the English sentence is determined with a coreference resolution system. The target-side $antecedent_{aligned}$ is obtained by following the word alignment links. In the shared task, the target pronoun *elles* is the pronoun that has to be predicted.

3.2.1 Features

Local Context

The local context around the source pronoun and target pronoun can contain the antecedent (cf. Figure 3.1) or other relevant information, such as the inflection of a verb which can provide evidence for the gender or number of the target-side pronoun. Therefore, we include the tokens that are within a symmetric window of size 3 around the pronoun. We integrate this information as bag-of-words, but separate the feature space by source and target side vocabulary and whether the word occurs before or after the pronoun. Special BOS and EOS markers are included for contexts at the beginning or end of sentence, respectively. We neither remove stopwords nor normalize the tokens.

We also include as features, the *POS tags* in a 3-word window to each side of source and target pronouns. This gives some abstraction from the lexical surface form. For the source side we use the POS tags from Stanford CoreNLP (Manning et al., 2014) mapped to universal POS tags (Petrov et al., 2012). For the target side we use

coarse-grained POS tags obtained with Morfette (Chrupała et al., 2008).[5]

Language Model Prediction

LMs provide a probability of a sequence of words and are trained on large monolingual corpora. They are used in SMT as a model to encourage fluency, i.e. producing typical target-language sentences.

We include the prediction of a target-side LM as a feature for the classifier. A 5-gram LM is queried by providing the preceding four context words followed by one of the eight target-side pronouns that the class labels represent. The pronoun that has the highest prediction probability is the feature value that we include in the feature vector. The ninth class OTHER requires special treatment, since it represents all other tokens that are aligned to a source-side pronoun. It does not itself appear in the LM training data. To get an accurate prediction probability for this aggregate class one would have to iterate over the entire vocabulary V (excluding the other eight pronouns) and find the most likely token. Since this would require a huge amount of LM queries overall ($|V|\times$ number of training instances) we approximated this search by taking the 40 most frequent tokens that are observed in the training data in the position which is labelled as OTHER. The highest prediction probability is then used to compete with the probabilities of the other explicit classes. We use the trained LM model provided by the shared task.

Target-side Antecedent Information

The *closest target-side noun antecedent* of the pronoun determines the grammatical features the pronoun has to agree with, i.e. *number* and *gender* in both French and German. Furthermore, if an antecedent exists, this marks the pronoun as referential, therefore discouraging translations of non-referential pronouns. We first apply the deterministic Stanford coreference resolution system, Stanford DCoref, (Lee et al., 2013) to the source side to determine the coreference chains in each document of the training data. We then project these chains to the target side via word-alignments (cf. Figure 3.2). The motivation to obtain target-side coreference chains in that way rather than computing them on the target side directly is three-fold. First, the target side of the training data is missing most of the target-side pronouns since it is the task to predict them. Therefore, relevant parts of coreference chains are missing and

[5] https://github.com/gchrupala/morfette

the place-holders for these pronouns might introduce additional noise to the target-side coreference resolution system. Secondly, we have a SMT scenario in mind as an application for cross-lingual pronoun prediction. Applying a coreference system to SMT output of already translated parts of the document is subjecting the resolution system to noisier data than it was originally developed for. Thirdly, resources and tools for automatic coreference resolution are more easily available for English than for French. In fact up till now there is no end-to-end coreference resolution system available for French.

Given the projected target-side coreference chains in a document, we greedily search for the closest noun token in the chain in the preceding context for each pronoun instance. The reason why we do not just search for the closest noun-antecedent on the source side and then take its projection is that nouns do not necessarily have to align to nouns, but could be aligned to NULL, pronouns, or other words. This found entity mention is included in the training data for the classifier as lexical feature. In addition, we extract morphological features from the noun (i.e. number and gender) by automatically analysing the target-side sentences with Morfette.[6] In cases where the pronoun was not assigned to a coreference chain, a special indicator feature was used. In addition, the word alignment can align one source token to multiple target tokens. We searched for the first noun in the aligned tokens and considered this to be the representative head antecedent of the given pronoun. If no noun could be found with this method, we resorted to taking the best representative antecedent of the source chain as determined by the Stanford coreference system and took the aligned token as the relevant target-side antecedent. In this case *null* alignments are also possible and a special indicator feature is used for that.

Pleonastic Pronouns

Pleonastic pronouns are a class of pronouns that do not have a referent in the discourse (i.e. they are non-referential) and they act as dummy subjects in constructions such as "*It* is raining". Their surface form in English is indistinguishable from referential forms. Nada (Bergsma and Yarowsky, 2011) is a tool that provides a probability estimate for *it* pronouns whether they are non-referential.[7] We include these estimates as

[6] Morfette's performance is quite robust and can handle sentences that contain *REPLACE_xx* tokens, which are the placeholders for target-side pronouns that have to be predicted. A comparison of the performance on the original sentences and the sentences with the *REPLACE_xx* tokens showed only minor differences.

[7] `https://code.google.com/p/nada-nonref-pronoun-detector/`

an additional feature. This should provide information especially for the French class labels that can be used as pleonastic pronouns, e.g. "*il* pleut (it is raining)" or "*ça* fait mal (it hurts)". Bergsma and Yarowsky (2011) suggest to use a threshold on the computed probability to make a discrete decision whether a pronoun is referential or not, however we use the probability estimate directly as a feature.

In addition, the rule-based detection of pleonastic pronouns is only basic in the Stanford DCoref system (cf. Lee et al., 2013, Appendix B). However since they do not have a referent, they cannot be part of a coreference chain. Therefore, we expect this feature to also counteract wrong decisions by the coreference resolution system to a certain degree. Since Nada only provides estimates for *it*, we do not have such a feature for pleonastic uses of the other source pronoun of the task *they*.

3.2.2 Experiments and Evaluation

Data

The shared task provides three corpora that can be used for training. The Europarl7 corpus, the NewsCommentary9 corpus and the IWSLT14 corpus which are transcripts of planned speech, i.e. TED talks. Only the latter two corpora come with natural document boundaries. Since these boundaries are necessary for coreference resolution, we did not use the Europarl corpus. The test data contains 1105 pronoun classification instances within a total of 2093 sentences in twelve TED talk documents.

Classifiers

We use Mallet (McCallum, 2002) to train the MaxEnt classifier.[8] The variance for regularizing the weights is set to 1 (default setting).

We trained classifiers in two different setups. The first setup provides all our extracted features as training data to one MaxEnt classifier, including the source pronoun as additional feature for each training instance (from now on referred to as the ALLINONE system). The second setup splits the training data into the two source pronoun cases (*it* and *they*) and trains a separate classifier for each of them (POSTCOMBINED system). At test time we use the specific classifiers according to the source pronoun and combine individual predictions to form the final set of predictions for the entire document.

[8] http://mallet.cs.umass.edu/

	Mac-F1	Acc
BASELINE	58.40_{1}	66.30_{8}
ALLINONE	57.07_{3}	72.31_{5}
POSTCOMBINED	54.96_{7}	71.40_{7}

Table 3.2: Official performance on the test data. Ranks according to each metric are given in subscripts out of 14 submitted systems (including multiple submissions per submitter and the baseline).

Evaluation Metrics

The official evaluation metric for the shared task is the macro-averaged F-score over all prediction classes (Mac-F1) defined by the evaluation script as follows

$$
\begin{aligned}
P_c &= \frac{\text{\# of correct predictions of pronoun class } c}{\text{\# of predictions of pronoun class } c} \\
R_c &= \frac{\text{\# of correct predictions of pronoun class } c}{\text{\# of gold pronouns in class } c} \\
F1_c &= 2 \cdot \frac{P_c \cdot R_c}{P_c + R_c} \\
P_{macro} &= \frac{1}{|C|} \sum_{c \in C} P_c \\
R_{macro} &= \frac{1}{|C|} \sum_{c \in C} R_c \\
F1_{macro} &= \frac{1}{|C|} \sum_{c \in C} F1_c
\end{aligned}
$$

where C is the set of target-side classes. Since this metric favours systems that perform equally well on all classes (i.e. all precision or recall values are weighted equally when averaged), the task puts emphasis on handling low-frequency classes well instead of only getting the frequent classes right. In addition to scores from the official metric we also report overall accuracy (Acc), i.e. the ratio between the correctly predicted classes and all test instances.

Results

Table 3.2 shows the official results on the test set together with the respective ranks out of 14 submitted systems. Table 3.3 and Table 3.4 provide the per-class precision,

	Prec	Recall	F1
ce	77.78	87.50	82.35
cela	25.00	18.52	21.28
elle	51.79	34.94	41.73
elles	85.00	33.33	47.89
il	50.00	59.62	54.39
ils	76.84	91.25	83.43
on	63.64	37.84	47.46
ça	62.69	41.18	49.70
OTHER	80.95	90.48	85.45
Macro-averaged	63.74	54.96	**57.07**
Accuracy			**72.31**

Table 3.3: Performance of **ALLINONE** classifier.

recall and F1, overall accuracy, and overall macro-averaged F-score. Table 3.5 shows results of our feature ablation experiments.

Generally, in terms of the official score (macro-averaged F1) one can see that our system has a high rank at position three (including the baseline), however it is still below the baseline. This is true for all the other submitted systems. In terms of accuracy, however, we are well above the baseline performance (albeit only at rank 5).

The worst performing class is *cela*. However, separating the two pronouns *cela* and *ça* is open for discussion, as it is more a matter of style than function, and it is often not consistently used according to common conventions. Merging these two classes is expected to result in an overall better system without reducing the accuracy of a translation.

3.2.3 Discussion

Confusion Matrices Table 3.6 and Table 3.7 present confusion matrices on the test set. Divergences from strong diagonal values in both tables derive in part from gender-choice errors (pronoun is predicted as $\{il,ils\}$, but should have been $\{elle,elles\}$ and

	Prec	Recall	F1
ce	78.05	86.96	82.26
cela	9.52	7.41	8.33
elle	49.06	31.33	38.24
elles	80.00	31.37	45.07
il	51.54	64.42	57.26
ils	75.79	90.00	82.29
on	61.90	35.14	44.83
ça	64.29	44.12	52.33
OTHER	80.00	88.52	84.04
Macro-averaged	61.13	53.25	**54.96**
Accuracy			**71.40**

Table 3.4: Performance of **POSTCOMBINED** classifier.

vice versa). On the other hand, the grammatical number of the personal pronouns is almost perfectly predicted in all cases. The OTHER class causes quite a few confusions among all pronouns, which is not surprising since it aggregates a heterogeneous set of possible source pronoun translations. We expect that a more detailed distinction in this group will result in better systems in general.

Feature Ablation In order to investigate the usefulness of the different types of features, we performed a feature ablation (cf. Table 3.5). When removing all features that are related to the antecedent of the target pronoun we need to predict, i.e. the antecedent itself and its number and gender (*all w/o antecedent features*), we observe a considerable drop in performance for both evaluation metrics. This is according to our expectations, since number and gender are strong cues for most of the pronoun classes. The antecedent token itself (*all w/o number/gender*) also provides enough information to the classifier to make a positive impact on the results.

When removing all features related to the target side we can observe a consistent drop in performance over all sets and classifiers.[9] This result shows the strong influ-

[9]Features related to the target side are the LM, the target side context windows (lexical tokens and

	ALLINONE		POSTCOMBINED	
	Mac-F1	Acc	Mac-F1	Acc
all features	57.07	72.31	54.96	71.40
all w/o antecedent features	51.59	70.14	54.15	71.13
all w/o nada	50.86	69.86	54.84	71.40
all w/o number/gender	54.62	71.67	54.33	71.40
all w/o language model	54.83	71.13	55.32	71.59
only source-side features	34.81	55.20	34.41	54.84
only target-side features	55.05	71.49	54.82	71.31

Table 3.5: Feature ablation for both types of classifiers.

classified as →	ce	cela	elle	elles	il	ils	on	ça	OTHER	*Total*
ce	161	·	1	1	11	·	·	3	7	184
cela	·	5	2	·	4	·	·	9	7	27
elle	8	1	29	·	21	3	2	5	14	83
elles	2	·	·	17	·	28	·	·	4	51
il	12	1	12	·	62	1	4	2	10	104
ils	1	·	·	1	·	146	·	·	12	160
on	2	·	3	1	5	4	14	2	6	37
ça	6	12	7	·	18	·	1	42	16	102
OTHER	15	1	2	·	3	8	1	4	323	357
Total	207	20	56	20	124	190	22	67	399	1105

Table 3.6: Confusion matrix for the **ALLINONE** classifier. Row labels are gold labels and column labels are labels as they were classified. Dots represent zeros.

classified as →	ce	cela	elle	elles	il	ils	on	ça	OTHER	*Total*
ce	160	·	2	·	11	1	·	3	7	184
cela	·	2	1	1	5	·	·	8	10	27
elle	10	·	26	·	23	3	3	6	12	83
elles	2	·	1	16	·	28	·	·	4	51
il	9	1	10	1	67	1	2	2	11	104
ils	·	·	·	2	·	144	·	1	13	160
on	2	·	5	·	6	4	13	2	5	37
ça	5	14	6	·	14	·	1	45	17	102
OTHER	17	4	2	·	4	9	2	3	316	357
Total	205	21	53	20	130	190	21	70	395	1105

Table 3.7: Confusion matrix for the **POSTCOMBINED** classifier Row labels are gold labels and column labels are labels as they were classified. Dots represent zeros.

ence the target language has on the translation of a source pronoun. Removing the source-side features does not have a strong impact on the results, which is consistent again over all settings. Both results taken together strongly indicate that the target-side features are much more important than the source-side features.

Classifier Types The overall results show a consistently better performance for the ALLINONE classifier compared to the POSTCOMBINED one. One reason for this might be that splitting the training data into two separate sets for the POSTCOMBINED setting results in much smaller training data sizes for each of the individual classifiers. Our feature ablation results show that particular features are useful for the former classifier, but useless or even harmful for the latter. This instability might be due to the fact that the POSTCOMBINED classifier has to learn from much smaller data sets. Incorporating more training data from the Europarl corpus could alleviate this problem and would make it possible to determine whether these differences persist. However, the additional data comes from a different domain and is of a different genre than the test data, which might cause problems elsewhere.

POS tags), the antecedent of the target pronoun (lexical token and grammatical features).

Language Model The mixed results for the usefulness of the LM features (i.e. performance increase without the LM feature for the POSTCOMBINED system) prompt for a further investigation of how to integrate the LM. We base the LM predictions on the preceding n-gram of the target pronoun. However, it is also conceivable for this task to query the LM with n-grams that are within a window of tokens containing the target pronoun. Furthermore, there is a small mismatch between the trained LM which has been trained on truecased data and the preceding tokens we have from the shared task data where the case was not modified. If this difference is eliminated we expect more accurate LM predictions, which should in turn provide more accurate features for the classifiers.

Additionally, our LM feature currently predicts OTHER with a fairly high frequency of around 80% (followed by *il* with around 15%). This might be another reason why some classifiers work better without this feature, since this distribution does not match the observed distribution of target pronouns in the training data. The over-prediction might be counterbalanced by querying a smaller number of proxy words (i.e. 40 in our experiments).

3.2.4 Post Shared-Task Improvements

Since the baseline had such a strong performance on the test set, we experimented after the official shared task deadline with combining our classifiers with it. We replace our LM feature with the predictions from the official baseline. The results are shown in Figure 3.3. A big increase in both macro-averaged recall (6.16-9.13% absolute) and accuracy (3.53-4.53% absolute) can be observed for both of our systems. This combination brings our system above the baseline for both evaluation metrics, and therefore also to rank 1. It remains to be shown, however, how much the other participating systems would benefit from such a combination.

3.3 Feature Extension and Language Transfer

To improve our CLPP system, we experiment with extending the feature set. We add n-gram features, a more informed LM feature that is closer to the shared task baseline and explicitly learn to predict the empty string (NONE).

Furthermore, triggered by the additional language pair in the CLPP16 shared task, we apply the extended MaxEnt classifier to both English-French and English-German,

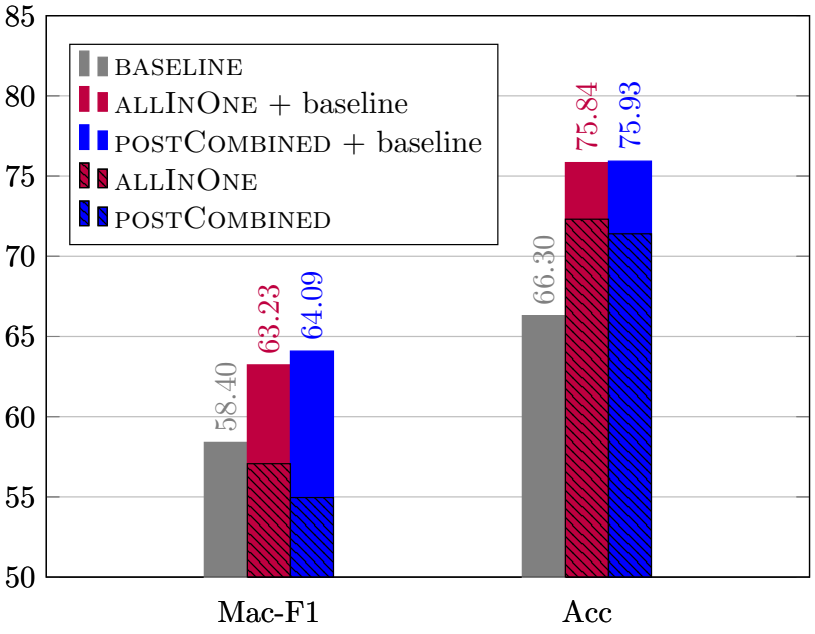


Figure 3.3: The baseline and the shaded areas correspond to the official shared task results from Table 3.2. The non-shaded area shows the improvement of our classifiers, when including the official baseline predictions as additional feature.

thereby testing it on a new language pair.

Additionally, the data available in the second CLPP shared task provides the target-side only in lemmatized and POS-tagged form, which more closely relates to a realistic SMT setting. In participating, we can experiment, whether our classifier is robust enough to also learn from a weaker signal.

Finally, we also experimented with a stronger sequence model, i.e. Conditional Random Fields (CRFs), attempting to exploit the sequential nature of coreference chains, where pronouns take part in.

In the following sections we describe our extensions and experiments in terms of our submission to the second CLPP shared task.

3.3.1 Features

In this section we motivate and describe the types of features we extract for learning the MaxEnt classifier and the CRF models.

Local Context

For each training instance, i.e. for each source pronoun for which we want a prediction, we extract a bag of words consisting of the ± 3 tokens around the source pronoun. Additionally, we extract the tokens in the ± 3 context window of the aligned target pronoun. The source-side feature consists of tokens in their full form, whereas the target-side feature uses the lemmatized tokens.

Additionally, we extract POS tags for these tokens. For the source side we automatically obtain POS tags with StanfordCoreNLP (Lee et al., 2013). For the target side the POS tags are provided as part of the training and test data.

A common strategy to improve linear classifiers is to include combinations of features so that the classifier can tune additional weights if feature combinations provide information that they cannot provide on their own. Therefore, we experiment with combining the above context window features within each type. In addition to the unigram values, we extract n-gram combination of these values by concatenating adjacent tokens or POS tags.

All of the above features are extracted both from the source and the target side.

Pleonastic Pronouns

We include a probability estimate as feature value of a source-side *it* that it is a pleonastic pronoun. This feature remains unchanged compared to Section 3.2.1.

Language Model Prediction

In our first CLPP system we incorporated a LM feature based on the preceding 5-gram context of a target-side pronoun (cf. Section 3.2.1), by utilising the conditional probability $P(classLabel_5|w_1,w_2,w_3,w_4)$, where *classLabel* is one of the class labels from the closed set of target classes, or the OTHER class, and w are the preceding words. This ignored any information following the pronoun, which could as well be indicative of the correct prediction. Therefore, we expand the feature to provide a rating for the entire sentence, i.e. $P(\langle s\rangle, w_1, ..., classLabel, ..., w_n, \langle /s\rangle)$, where n is the sentence length, and $\langle s\rangle$ and $\langle /s\rangle$ are sentence boundary markers.

The class label that produces the highest scoring sentence according to the LM is then used as a feature value in our classifier. To obtain such a prediction for the class labels that correspond to pronouns we can directly substitute the target-side pronoun placeholder with each class label when querying the LM.

The OTHER class requires special treatment, since it does not occur as such in the LM training data. We approximate the probability for this class in the same way as described in (Wetzel et al., 2015). We first collect frequencies of words that are tagged as OTHER from the training data. Then we query the LM with the top-n words as substitute for the placeholder. The highest scoring word within that group then competes as representative for OTHER against the probabilities of the rest of the class labels.

Target-side Antecedent Information

The antecedent feature proved useful in (Wetzel et al., 2015). Both in German and French, the pronoun has to agree in gender and number with its antecedent. Intuitively, if we know the closest target-side antecedent of a referential target-side pronoun, we have access to information such as grammatical gender and number of the referent. Furthermore, the fact that an antecedent exists provides information as well, since it separates referential from non-referential cases.

We use the same antecedent detection method as in our previous CLPP system. We perform antecedent detection with the help of source-side coreference chains. We

Corpus	en-de	en-fr
NC9	63.72	25.12
IWSLT15	68.55	31.25
TEDdev	60.00	34.31

Table 3.8: Percentage of NONE within the OTHER class.

follow the source-side chain that contains the source pronoun of interest in reverse order (i.e. towards the beginning of the document) and check if the token that is aligned to the source-side mention head is a noun. If it is not, the search proceeds. We take the closest noun that we can find on the target side.

Since the target side only contains lemma information, where all gender- or number-specific information has been removed from nouns (or merged to the same token for e.g. determiners), we cannot apply a morphological tagger to give us this information for the found antecedent. Therefore, we resort to a simpler method and look up the most frequent gender for a given lemma in a lexicon. We only experiment with this feature on the English-German task.

Predicting NONE

Source pronouns do not necessarily have a counterpart in the target language. These cases are mapped to the OTHER class in the training and test data and occur very frequently within this class (cf. Table 3.8). Though not part of the official set of classes, they are still specifically marked with a NONE maker. If we know that a source pronoun does not have a translation, then this might be useful in an SMT scenario, where for example a feature function could score phrases higher that do not contain target-side pronouns. For CLPP our expectation is that it should help to improve prediction performance for the very heterogeneous OTHER class.

For training the classifiers we first map all NONE cases from the OTHER class label to NONE, thus introducing a separate class label. For the final predictions we map any NONE predictions back to OTHER before evaluation.

Sequence length	en-de	en-fr
1	74.45	74.13
2	15.34	15.43
3	5.34	5.43
4	2.21	2.29
5	1.08	1.11

Table 3.9: Distribution of pronoun sequence lengths up to 5 in the English-German and English-French training data (IWSLT15 and NewsCommentary9) for the ALLINONE setup in percent.

3.3.2 Pronoun Prediction as a Sequence Labelling Task

The MaxEnt classifier makes the assumption that the translation of a pronoun is only dependent on the source and target contexts and the antecedent it refers to (for referential pronouns). This ignores the fact that pronouns are part of a longer chain of coreferring expressions, which in turn may be other pronouns.

Therefore, we first prepare the training and test data such that all pronoun instances that belong to the same coreference chain form one training or testing sequence. We then train a linear-chain CRF with the features as given above to predict an optimal sequence of target pronouns for a given sequence of source-side pronouns, rather than making each prediction independently of the other pronouns. This way, typical patterns of pronoun sequences can be learnt, which might help with the prediction. Table 3.9 gives the distribution of sequence lengths, i.e. the number of pronoun instances per coreference chain. The slight difference between English-French and English-German comes from the fact that there is a slight mismatch between the parallel documents contained in each language pair variant.

We expect this additional information of typical pronoun sequences within one coreference chain to be useful for the prediction task, since it allows for optimizing the individual predictions in a chain on a global document level.

Gender	Frequency
Masculine	20878
Feminine	21221
Neuter	12894
Total	54993

Table 3.10: Number of nouns with gender information in the raw Zmorge lexicon for German.

3.3.3 Experiments

We first describe the experimental setup of our systems, then briefly describe the data we used and provide information about feature and parameter settings. Finally, we report our results on development and test data.

Systems

We use Mallet (McCallum, 2002) for training the MaxEnt classifiers and CRF models. For the MaxEnt classifier we use the default settings. For the CRF we train *three-quarter* order models (i.e. one weight for each ⟨feature, label⟩ pair, and one for each ⟨current label, previous label⟩ pair) and only allow label transitions that have been observed in the training data.

As before, we have two setups in all experiments. The POSTCOMBINED setup, where we split the training and test data for each source pronoun into separate sets, train separate classifiers and combine the predictions after classification. And the ALLINONE setup, where we do not split the data.

The systems marked with *initial* consist of the local context window features, the pleonastic pronoun feature, the LM feature and the antecedent information (without gender information). We use *fGender* to refer to the gender feature, *3-gram window* to refer to the n-grams from the local context window and *fNone* to refer to the NONE-prediction feature. Systems marked with *sequence* are the CRF models. We submit the best performing system according to the official macro-averaged recall measure on the development set for each language pair as primary test set submission.

Baseline

The official BASELINE uses LM predictions similarly to our LM feature. Additionally, it attempts to find the optimal predictions for a sentence, if there are multiple pronouns that have to be predicted. It has a NULL penalty parameter that determines the influence of not predicting a pronoun at all, which for this shared task is tuned. The underlying LM is trained on lemmatized text.

Data

For training, we only extract information from the IWSLT15 and NewsCommentary (NC9) corpus. We do not employ the provided Europarl corpus, as it does not come with predefined document boundaries other than parliamentary sessions of a complete day. For development, we use the TEDdev set. For the final submission on the official test set we include TEDdev in the training data.

Features and parameters

For the LM feature, we take the provided trained models from the shared task, which are 5-gram modified Kneser-Ney LMs that work on lemmatized text. We use KenLM (Heafield, 2011) for obtaining probabilities. As proxy for the OTHER class we use the top 35 words for German, and the top 70 for French. The threshold values are not tuned, but set to include mostly other pronouns that are not part of the official set of class labels.

For gender detection of German antecedents we use the lexicon from Zmorge (Sennrich and Kunz, 2014).[10] Gender distribution of nouns is given in Table 3.10. When a noun has multiple genders in the lexicon, we take the most frequent one for that noun.

The different parameters such as context window size were taken from our findings of the previous year (Wetzel et al., 2015). The n-grams of the context window are extracted for $n = [1..3]$ including beginning- and end-of-sentence markers if necessary.

Results

The results on the development set are given in Table 3.11 for English-German and in Table 3.12 for English-French.

The *initial* systems in each language pair perform much better than the baseline, which is especially noticeable in English-French. Adding the gender feature to the

[10] `kitt.ifi.uzh.ch/kitt/zmorge`, zmorge-lexicon-20150315

	Mac-R	Acc
BASELINE	34.35	42.81
ALLINONE-initial	39.24	56.14
+ fGender	40.00	57.37
+ fGender, 3-gram window	41.21	57.72
+ fGender, 3-gram window, fNone	40.86	58.77
ALLINONE-sequence-initial	35.67	54.91

Table 3.11: System performance in percent for English-German on the development data set.

	Mac-R	Acc
BASELINE	40.63	49.73
ALLINONE-initial	52.25	69.98
+ 3-gram window	54.68	73.36
+ 3-gram window, fNone	57.34	74.25
ALLINONE-sequence-initial	49.27	64.65

Table 3.12: System performance in percent for English-French on the development data set.

English-German classifier shows improvements in performance, thereby confirming the usefulness of adding gender information.

The additional feature that predicts NONE as possible translation is helpful for the English-French pair. Results on English-German showed a decrease in performance with respect to macro-averaged recall. This decrease is surprising, especially considering the much larger frequency of NONE in the German data set (cf. Table 3.8). In terms of accuracy the NONE feature is also beneficial for English-German.

Including the n-gram features from the context window is very good for performance in both language pairs according to both metrics.

The final results including the ranks on the official test set of the shared task are given in Table 3.13.

	en-de		en-fr	
	Mac-R	Acc	Mac-R	Acc
ALLINONE	48.72_5	66.32_6	61.62_4	71.31_3
POSTCOMBINED	47.75	64.75	59.83	68.63
BASELINE-1	n/a	n/a	50.85	53.35
BASELINE-2	47.86	54.31	n/a	n/a

Table 3.13: Official shared task results. Ranks of our primary submission are given in subscripts with a total of nine submissions for each language pair.

3.3.4 Discussion

General Performance In general, performance is considerably lower for English-German compared to English-French, despite the former having a much smaller set of class labels to choose from. One reason for that might be that in English-German, the OTHER class is even more heterogeneous than in French, and taking apart this class to the same degree as in the English-French data sets might be beneficial.

Development vs. Test Data Performance between development and test sets varies greatly despite similar class label distributions (except for a much smaller amount of OTHER instances in the English-French test set). To a certain degree this is expected, however the big changes in performance suggest that there are other differences in the data sets which are worth exploring.

Baseline Training a MaxEnt classifier where we substitute our LM feature with predictions from the shared task baseline performed slightly worse. This suggests that a simpler LM feature is sufficient when included in the classifier, and that joint prediction of multiple target pronouns within one sentence is not necessary. However, we did not tune the NULL penalty of the baseline model.

Furthermore, the lemmatization of the French data merges singular and plural forms of *il* into one lemma, similarly for *elle*. The baseline which uses the LM trained on the lemmatized data is therefore never able to predict the plural forms of these two pronouns, resulting in zero precision and recall for these classes. This is confirmed by the corresponding confusion matrix. This might also have an indirect impact on the

	er	sie	es	man	OTHER	*Total*
er	4/4	2/2	3/8	·	6/1	15
sie	3/2	73/100	11/15	3/·	34/7	124
es	2/·	9/4	61/85	2/·	27/12	101
man	·	·/1	2/4	1/1	5/2	8
OTHER	2/1	11/17	7/16	·	115/101	135
Total	11/7	95/124	84/128	6/1	187/123	383

(a) English-German

	ce	elle	elles	il	ils	cela	on	OTHER	*Total*
ce	58/60	·	·	6/6	·/1	1/·	·	3/1	68
elle	2/2	10/9	2/·	5/8	·/1	2/3	·	2/·	23
elles	2/·	2/·	3/6	·	15/17	1/·	·/1	2/1	25
il	5/6	1/6	·	43/43	2/1	4/3	2/2	4/·	61
ils	·	·	9/7	·	54/63	·	·	8/1	71
cela	·	3/1	·	8/7	·	15/20	1/1	4/2	31
on	·	·	·	·/1	2/4	·	6/4	1/·	9
OTHER	1/3	1/·	·	4/7	1/·	1/1	·/2	77/72	85
Total	68/71	17/16	14/13	66/72	74/87	24/27	9/10	101/77	373

(b) English-French

Table 3.14: Confusion matrices for the ALLINONE classifier on the test set. Row labels are gold labels and column labels are labels as they were classified. Dots represent zeros. Numbers to the left of the slash represent our shared task submissions, numbers to the right are for the results when we removed the LM feature from these submissions.

	Mac-R	Acc
ALLINONE	48.72	66.32
− fAntecedent	46.24	64.23
− fLM	55.76	75.98
− fLM, + fNone	60.17	77.28

(a) English-German

	Mac-R	Acc
ALLINONE	61.62	71.31
− fAntecedent	61.89	71.85
− fLM	63.03	74.26

(b) English-French

Table 3.15: Feature ablation results of our submitted ALLINONE systems on the test set.

performance of our classifiers, since they use LM prediction as a feature.

Confusion Matrices The confusion matrix for English-German in Table 3.14a (counts to the left of the slash) shows that OTHER is over-predicted (i.e. the sum of the OTHER-column is much larger than the number of gold labels for that class), which might explain the overall lower performance of the system. Furthermore, *es* and *sie* are confused by our classifier. For English-French in Table 3.14b (counts to the left of the slash) one can observe that the biggest confusion is between gender in plural pronouns (i.e. *elles* and *ils*). This might be because we did not include any explicit gender information as feature. As above, the OTHER class is also very confused over all cases, however it is not too over-predicted.

Classifier Types Similarly to our findings from last year, the POSTCOMBINED setup scored consistently worse on the test sets (and only once slightly better on the development set). This provides evidence, that splitting the training data according to source pronouns is counterproductive. Furthermore, it might even be worse for the inverse prediction tasks available at the shared task, since there are a lot more source pronouns, hence making the available data even sparser.

Feature Ablation Feature ablation experiments on the test set shown in Table 3.15 revealed that the antecedent feature is helpful for English-German, but not for English-French. One possible explanation for this might be that we do not have gender information of the antecedent in French and only adding the antecedent itself might not be sufficient.

Additional ablation experiments on the test set showed that our LM feature in fact

hurts performance. Removing this feature gives a boost in performance, which brings our systems to the second place (first for accuracy) for English-German and to the third place (second for accuracy) for English-French. This contradicts findings from experiments we conducted for last year's shared task, where adding baseline predictions, which are very similar to our LM feature, greatly improved results. An explanation for this behaviour could be that the LM this year was trained on lemmatized text and therefore performs much worse than when trained on original data. Confusion matrices for these results are given in Table 3.14 (numbers to the right of the slash). For both language pairs we are now under-predicting OTHER, however gaining accuracy on the classes representing pronouns.

Furthermore, removing the LM feature brings the English-German system to the same level of performance as the English-French system. This provides some evidence that the same set of features can be used to make predictions into both of these languages.

3.4 General Discussion

Generally, our CLPP experiments show that a small set of features already results in a good performance. In both CLPP15 and CLPP16 shared tasks we manage to achieve results much better than the LM-based baseline (albeit in the CLPP15 shared task we achieved this only after modifications to our official submission). Our best-performing classifiers (other than our official submissions) result in the best rank at the CLPP15 shared task, and among the top two (English-German) and three (English-French) performing systems in the second shared task. Taken together this shows that we have competitive systems for CLPP at hand. These state-of-the-art results also make our findings more reliable, since they are based on well-performing systems. Furthermore, we will reuse our CLPP systems in Chapter 6 to generate discourse-aware translations in a post-editing setup based on an NMT system. These system translations are then used to as a basis for verification of our discourse-aware SMT evaluation metric.

The LM-based baseline is very strong in the first shared task, i.e. it outperforms all submitted systems. This strong baseline can be exploited by combining the baseline predictions with the other features of our classifiers, resulting in an increase in performance. One of the reasons why the baseline is so strong is that the LM predictions of the baseline are based on human reference translations. Therefore, reliable

cues are provided in the surrounding context of the target-side pronoun. However, in a more realistic setting (i.e. one which has a full SMT system in mind), the surrounding context no longer provides ground truth cues, and a diminishing performance of the LM baseline is expected. This hypothesis is supported by the fact that the baseline no longer performs so well on the CLPP16 shared task, which no longer provides full human reference translations, but only lemmatized tokens on the target side. This setting is more realistic in terms of the full SMT scenario, in that it does not make any commitment to linguistic cues of surrounding words that agree in gender and number. Furthermore, including the baseline in our classifier for the CLPP16 shared task hurts performance, additionally suggesting that other features provide more reliable information. The detrimental effect of the LM predictions or of features derived from these baseline predictions is also observed in a feature ablation study in Stymne (2016).

We applied our CLPP system to two language pairs, i.e. English to French and English to German. Experiments confirm a similar performance on both language pairs (when considering our best performing systems), showing that our set of features captures relevant information in both settings. Most of our features require only few resources that are specific to the target side, i.e. either a morphological analyser and POS tagger (for the CLPP15 shared task) or a simple lexicon with gender information (for the CLPP16 shared task). However, since we require a coreference resolution system on the source language, it is not trivial to train CLPP systems in the inverse direction. An end-to-end coreference resolution system for German exists (Tuggener, 2016, CorZu), however, for French there is currently no such end-to-end system available.

The feature that captures non-local dependencies (i.e. the antecedent feature) proved to be useful across both shared tasks and languages. The only exception is the English-French system on the CLPP16 shared task, where we see a slight decrease of 0.27% (absolute) for macro-averaged recall and 0.54% (absolute) for accuracy, when the feature is used. However, in this particular setting, we only have the antecedent token as feature and we did not extract gender or number from the token as in all other settings. So the missing gender and number might make the feature less useful, especially when the target side is lemmatized as in the CLPP16 shared task. The gender and number information from the antecedent proved very useful in our feature ablation experiments of the CLPP15 shared task, which adds another piece of evidence to this hypothesis.

We included a feature into the classifier that learns to predict when a pronoun in the source should *not* be translated in the target side. This is represented with the

special NONE class. With regards to evaluation in the two CLPP shared tasks, the NONE prediction was not part of the official set of class labels. However, this feature not only provided a performance improvement with respect to the OTHER class under which it is officially grouped. But it also allows us to actually make predictions that instead of a pronoun, the target side counterpart of a source pronoun should be the empty string. This is potentially useful in a full SMT setting, providing a more informed translation choice. We reuse this feature for prediction of the empty word in one post-editing system variant in Chapter 6.

In all experiments we only made use of a part of the provided training data (i.e. the IWSLT and NewsCommentary corpora) and we did not consider the much larger Europarl corpus. This suggests that a small amount of training data is sufficient for training a classifier with good performance. The main reason for excluding the Europarl corpus is that it does not come with clear document boundaries, which are important to have for the coreference resolution system used in our antecedent feature. Without clear document boundaries mentions could be linked to entities from items on the parliamentary agenda that have nothing in common, or mentions could be missing in the beginning of the document. Providing a plausible separation of the parliamentary sessions into self-contained documents would be an interesting task for future work, to make this larger resource available for document-level SMT in general. Note however, that the LM used in the baseline and in one of our features was trained on a much larger set of data that included the Europarl corpus.

The CLPP shared tasks focus on TED talks as test set, since they generally have a higher distribution of pronouns than other domains or genres (Ruiz and Federico, 2014; Guillou, 2016, Chapter 4). This is a good basis for encouraging research on pronoun translation, since especially problematic low frequency cases (e.g. *elles*, the feminine 3rd person plural pronoun in French) have a higher chance of occurring, if there are more pronouns in general. One next step would be to show that CLPP systems can also be applied to other genres or domains, such as the more streamlined domain of written newstexts which is commonly used as test set for general SMT systems at the WMT translation task. It remains to be shown, whether CLPP systems can be successfully applied to such a new domain without any adaptation, or what the additional requirements are. We address this issue to a certain degree in Section 4.5 by comparing the performance of our CLPP systems on the CLPP16 test set (i.e TED talks) against a newstext test set. Results there show a decrease in performance according to macro-average recall, but a slight increase in accuracy.

Our sequence learning experiments showed to be unfruitful with the same set of features from the MaxEnt classifier and we did not further investigate why this is the case. Intuitively, there should be a benefit in performance from optimizing CLPP predictions globally. However, maybe the gender and number of pronouns is only really dependent on the closest noun antecedent, and no generalizations can be made if they belong to the same chain. I.e. nouns with different grammatical gender can be used to refer to the same entity.

There is still a gap to fill between existing CLPP systems from the two shared tasks and being able to apply them in an actual SMT scenario. On the one hand, the shared tasks only focussed an a very small set of source and target pronouns (3rd person singular and plural subject position pronouns on the source, and only a few most frequently aligned pronouns on the target side), albeit ones that represent the more difficult cases. Expanding these sets to other pronouns would increase the instances, where the CLPP system can make a particular translation choice. In addition to that, the target-side data of the shared task is so far still based on human reference translations. The second shared task has seen a step towards a more realistic setting, where the full word forms were no longer available, but were lemmatized. This is a step in the right direction, since SMT output, where we want to apply the CLPP systems in the end, is much noisier and less reliable that human reference translations. We address this issue in Chapter 4.

Chapter 4

A Corpus of Document-level Pronoun Annotations in SMT Output

In this chapter, we present an English-German parallel corpus with an automatically translated target side created by a state-of-the-art NMT system and manual gold annotations of pronouns in this translation. We first motivate the necessity and benefits of such a corpus emphasising that in addition to its value in evaluating CLPP systems in a realistic setting, it can also serve as a proxy of human coherence judgments (Section 4.1). Then, we give details on how we processed and constructed the corpus (Section 4.2). We present the annotation interface used to elicit human annotations (Section 4.3) and explain how we conducted the actual annotation. We provide an analysis of the resulting annotations (Section 4.4) showing an overall high inter-annotator agreement, low error rate and only a small number of unresolvable ambiguities. Finally, we run experiments with our existing CLPP systems leveraging this new resource (Section 4.5). We explore the effect on CLPP systems when changing the domain and genre which causes a decrease in performance with respect to macro-average recall, but an increase for accuracy. We further explore the performance difference between human and automatic translations. This experiment does not provide a clear picture, neither observing a consistent increase nor decrease of performance. Furthermore, we compare the pronoun prediction performance of the NMT system revealing a better performance than the CLPP systems.

4.1 Motivation

Both CLPP shared tasks have human reference translations on the target side of the training and test data. While this was useful as a way to study and experiment with pronoun translation without the additional noise introduced by SMT output, the major goal is to apply these CLPP systems to such data, so that they can be used in a full SMT scenario. The shared task organizers went a step towards that direction in the second shared tasks, where they removed morphological information from words by lemmatizing them. This matches the SMT scenario more closely, since the morphological information in SMT output no longer represents ground truth, but adds uncertainty in terms of whether the words are translated in their correct form.

However, the current setup still does not enable experimentation to find out whether CLPP systems work well also on SMT output. We therefore create a corpus with automatic translations on the target-side produced by a state-of-the-art NMT system. From this corpus, we remove the same set of target-side pronouns based on the same set of source-side pronouns as in the CLPP16 shared task (cf. Section 3.1.1), enabling experiments to predict pronouns in context of fully automatically translated data.

One of the major advantages of the original two shared task setups is that since the target-side documents come from human translations, they also automatically provide the gold translations for each pronoun instance. This makes it possible to train and evaluate CLPP system performance automatically without the big overhead of manual annotation. If the target-side document consist of automatic translations this automatic supply of gold labels is no longer given.

However, to enable experimentation of CLPP on actual SMT output, manual annotation is unavoidable. We therefore elicit manual annotations where each pronoun instance that has been removed from the automatically translated target side of a document has to be reinserted by a human with one of the available class labels (i.e. target-side pronouns) from the official sets of the CLPP16 shared task. With these gold annotations in place, we can then provide a test data set for CLPP systems, to find out how well they perform on realistic data.

These manual annotations also allow us to directly assess the performance of the state-of-the-art NMT system with respect to pronoun translation. Each pronoun that was removed and substituted with a REPLACE token to produce the CLPP test set, can also be seen as a pronoun prediction of the NMT system. These can then also be compared to the gold pronoun annotations.

One additional benefit of having an automatically translated target side is that the baseline translations of pronouns from the SMT system can be integrated in the CLPP system as additional feature. This would, for example, make it possible to learn whether the baseline translation already covers certain pronouns well, so that these should no longer be changed by a subsequent CLPP system.

Finally, this annotated corpus can be seen as a proxy for manually judging the coherence of a document. Pronouns are part of coreference chains and therefore contribute to the coherence of a document. Hence, if we translate more pronouns correctly, this should increase the coherence of the resulting document. A translation that matches the pronouns of the annotated corpora more often, should be the preferred translation, over one where the pronouns match less frequently. This provides us with relative rankings between two translation systems on the document level. These human-based rankings in turn can then be used to verify if an SMT evaluation metric operating on the document level, correlates with these human-based rankings. We come back to this in our later chapter where we present our discourse-aware SMT evaluation metric (cf. Chapter 6).

4.2 Corpus Construction

In this section we first present the data we use as a basis for the annotations. Then we mention how we obtain automatic translations. Following this, we elaborate on our strategy to determine the final list of possible annotation choices. Finally, we explain how we process and prepare the data for the annotation process.

4.2.1 Data

The data we use is from well-known test sets that are used in the WMT news translation task for English-German. We chose to use one of the WMT news translation task test sets, since they are publicly available and commonly used in SMT evaluation. This will make our corpus more useful for comparison and further experimentation. Furthermore, we have access to a state-of-the-art NMT system that performs best on this data set. On the other hand, the two CLPP shared tasks used TED talks as development and test data. The major reason for this was the higher frequency of pronouns in TED talks than in news text (Hardmeier et al., 2015). Despite using a different data set, we still want to work with documents that have a similarly high frequency. Additionally,

	en	de	min-sent	average-sent	max-sent	# of docs
WMT12 news test	0.86	1.98	5	30.33	147	99
WMT13 news test	1.15	2.13	37	57.69	99	52
WMT14 news test	0.85	1.81	4	18.31	83	164
WMT15 news test	0.99	2.03	5	26.78	264	81
WMT16 news test	1.01	2.11	2	19.35	77	155
IWSLT15	1.68	3.24	1	121.97	399	1592

Table 4.1: Several statistics from the WMT news test corpora between 2012 and 2016, and the IWSLT15 corpus. The first two columns represent the pronoun/token ratio in percent averaged over the entire corpus. The minimum, average and maximum sentence count is provided to determine document length. The final column shows the total number of documents in each corpus.

the documents in the corpus should not be too long, since they have to be annotated manually. In longer documents, annotators are expected to be less focused (i.e. making more errors) towards the end.

We therefore only consider one of the WMT test sets to limit the annotation requirements and to allow to choose the data set that has a pronoun distribution that is as close as possible to the CLPP shared task test sets. In order to find a suitable one, we collected a set of statistics as presented in Table 4.1. For comparison, we also include the statistics for the IWSLT corpus, which consists of TED talks and was part of the training data in the CLPP shared tasks. For the statistics, we first tokenized and truecased the documents using the Moses pre-processing scripts via the EMS.[1] For the English pronoun/token ratio, we count the number of occurrences of the two pronouns *it* and *they* and divide by the number of tokens in the corpus. Similarly, for the German pronoun/token ratio, we count the number of the pronouns *er, sie, Sie, es, man*. These two sets are the same set of pronouns that are used in the CLPP shared task 2016.

All requirements taken together lead to the *WMT16 news test set*, since it provides one of the highest pronoun/token ratios in both languages while containing comparatively short documents (i.e. on average 19.35 sentences). The WMT16 news test set therefore forms the basis for our annotations.

[1] `http://www.statmt.org/moses/?n=FactoredTraining.EMS`

4.2.2 NMT System

The automatic translations that form the basis for annotating the target side are obtained with the state-of-the-art NMT system Nematus (Sennrich et al., 2016a). This system was the best-performing one at the WMT16 translation task with a BLEU score of 34.8 (uncased) and 34.2 (cased) for that language pair and direction. It follows the encoder-decoder architecture with attention mechanism (Bahdanau et al., 2014). For handling unknown words and restricted vocabulary size, it uses byte-pair encoding (BPE) (Sennrich et al., 2016b), which is a compression-based algorithm that splits words into smaller subunits. In addition to the standard parallel data of about 4 million sentence pairs, a sample of the same size of the monolingual target-side data is back-translated into English with a German-English system, and added as additional training data. Furthermore, it uses ensemble decoding with the last four models stored during training after a fixed set of epochs. We reuse the pre-trained ensemble models for English-German[2] to translate the WMT16 news test set in order to obtain automatic translations for our annotation task. We verified the translation and obtained the same BLEU scores as reported in Sennrich et al. (2016a).

4.2.3 Annotation Choices

Annotators are presented with a complete parallel document from the automatically translated WMT16 news test set. For each removed target-side pronoun instance, annotators are given a closed set of choices from which they have to choose. In a pilot study, we experimented with three different sets of pronouns, to determine the final choices we offer the annotators. The sets differ in the source and target pronouns. The source pronouns determine where an annotation location (i.e. pronoun instance) is (cf. Section 4.2.4). The target pronouns determine the choice the annotator has when annotating. The three sets are as follows:

p001 source pronouns: *it, they*
target pronouns: *er, sie, es, man*, NONE, OTHER

p002 source pronouns: *it, they*
target pronouns: *er, sie (singular), sie (plural), es, man*, NONE, OTHER

p003 source pronouns: *i, you, he, she, it, we, they, me, him, her, us, them, mine, yours, his, hers, ours, theirs*

[2] http://data.statmt.org/rsennrich/wmt16_systems/en-de/

	p001/p002	p003
total # of docs	155	155
# of pronoun instances (total)	669	2181
# of pronoun instances (average per doc)	4.35	14.07
# of docs without pronoun instances	32	10

Table 4.2: Statistics about pronoun instances in the WMT16 news test corpus for the different source- and target-side pronoun sets.

	p001	p002	p003
average duration per instance	27.05s	36.33s	18.45s
total estimated duration	5.0268h	6.7513h	11.1776h

Table 4.3: Observed and estimated annotation times (in seconds and hours, respectively) for the WMT16 news test set for all three annotation variants.

> target pronouns: *ich, du, er, sie, es, wir, ihr, mir, dir, ihm, uns, euch, ihnen, mich, dich, ihn, man*, NONE, OTHER

In Table 4.2 we show an overview of how many pronoun instances the different pronoun sets produce in the WMT16 news test dataset, which in turn require an annotation choice.

We annotated a small set of documents for each of the three annotation variants. We chose around five different documents for each variant. This avoids getting to know the documents, which would result in less accurate estimates. The average annotation duration per pronoun instance and the total estimated annotation time for the entire corpus are given in Table 4.3. The duration estimates are the lower bound, since we are familiar with the task and can therefore make faster decisions. In addition to the duration estimates for the entire WMT16 news test set, the time for understanding the guidelines has to be added. Reading the guidelines (cf. Section 4.4.1) slowly takes about 3 minutes, so for someone new to the task it should not take longer than 10-20 minutes to fully understand them.

After the pilot study, we observed three main issues discussed below. For the first two pronoun lists (*p001* and *p002*), and the documents we annotated there was almost always only one choice that seemed plausible. From this we do not expect a great

variety of annotations from different annotators. The pronoun list *p003* also contains pronouns that are very easy to annotate. For example, *I* can be directly translated to *ich* without thoroughly reading the entire sentence or even surrounding context. On the other hand, and contrary to the previous two lists, it contains hard cases. For example, *you*, which could be translated as of the pronouns *man, du, Sie, sie* and is – when the *you* is a generic-you – a matter of stylistic choice.

In rare occasions the translation was too bad to make a decision. This was even the case when consulting the surrounding source and/or target context. This motivated the addition of an UNDECIDABLE option to the set of target-side pronouns.

In many cases, when choosing OTHER, we added the pronoun to the available comment field. However, we also added additional comments. This makes it harder to post-process the annotations, so this motivated adding an extra field in the comment dialogue, where such a pronoun can be added.

Rather than using NONE and OTHER directly as possible pronoun choices, we use the substitutions NO WORD and ANOTHER WORD respectively. These are more straight forward for an annotator to understand.

For the actual annotation task we decided to use the second pronoun list (*p002*). This list provides a good compromise with respect to annotation time and potential benefit that we can obtain from the annotations. Finally, this list is very close to the official set of pronouns used in the second CLPP shared task, which makes it easier to reuse these annotations with existing CLPP systems.

4.2.4 Preparation for Annotation and as CLPP Test Corpus

For the annotation and also for preparing the parallel corpus so that it can be used with CLPP systems, we require word alignments. This is used to link the source pronouns with their position on the target side. And in order to determine gold labels in case our target side is taken from human reference translations. The NMT system does not directly provide high quality word alignments for the translated text. We therefore use MGiza[3] with the *grow-diag-final-and* heuristic via the EMS[4] to compute these word alignments. In order to make them more reliable by providing more training data, we computed word alignments for a concatenation of the WMT12-16 news test set and the Europarl7 corpus, IWSLT15 corpus and NewsCommentary11 corpus. Unlike in the data preparation for the CLPP 2016 shared task (Guillou et al., 2016, Section 3.3.1),

[3] https://github.com/moses-smt/mgiza

[4] http://www.statmt.org/moses/?n=FactoredTraining.EMS

we do not experiment with different settings to obtain word alignments. We choose the standard word alignment settings used in phrase-based SMT system training.

In order to identify the target-side pronouns for which we want human annotations, we follow the procedure used to prepare data for the CLPP shared tasks (Guillou et al., 2016, Section 3.3.2). Given word alignments and a set of source pronouns, we follow the word alignments for each of them and take the aligned token as the target-side pronoun instance we want to obtain annotations for. Several heuristics are necessary to cope with null alignments and 1:m alignments as detailed below:

1. A source-side pronoun has exactly one word alignment link to the target side (e.g. "[it] → [es]"). In this case the target-side pronoun will be substituted with a REPLACE token. If it is not equal to one of the defined class labels, it will be assigned to the OTHER class.

2. A source-side pronoun has more than one alignment link to the target-side. There are three cases:

   (a) There is one, and only one class label in the list of aligned target-side tokens (e.g. "[they] → [Mehr, sie]"). In that case this target-side word will form the pronoun instance and the other tokens will be ignored.

   (b) There is more than one class label in the list of aligned target-side tokens (e.g. "[they] → [Sie, es]"). In this case one of the pronouns will be chosen at random.

   (c) None of the linked target-side words are pronouns from the set of class labels (e.g. "[they] → [Mehr, da]). In this case the shortest word will be taken and assigned to the OTHER class.

3. A source-side pronoun does not have any alignment links (e.g. "[they] → []"). In this case a position for the special NONE class has to be found. This is done by expanding the source-side context around the source-side pronoun token by token (alternating between left and right, starting with the left context first). As soon as a source-side token with an alignment link is found, the REPLACE token for the NONE class is inserted after (before) the found link for left (right) context. Whenever an explicit NONE link is inserted, all other affected word alignment ids are updated.

The second CLPP shared task emphasizes that the main focus of CLPP in the task is the prediction of subject position pronouns (Guillou et al., 2016, Section 3.3.3). They therefore filter out pronoun instances that were obtained in the way described above, that are not in the subject position according to a source-side dependency parser. For our corpus, we do *not* perform this additional filtering step.

For the corpus used for annotation, the target side consists of automatic translations. The procedure above assumes that the word alignments can be reliably used to identify the correct position (i.e. the position where a manually annotated pronoun should occur) on the target side even in noisy automatic translations. During annotation it is implicitly verified by the annotators whether the location of the annotation gap for target-side pronouns was at a correct position. Only one observation of a wrong position was mentioned as a comment by the annotators, which provides some evidence that the identified target-side locations were at least acceptable to make an annotation choice. However, we did not explicitly elicit any comments in this regard.

4.3 Annotation Interface

The annotator, once logged into their account, is presented with a button to start a new annotation task together with the total number of annotations that still need to be done. A screenshot is shown in Figure 4.1.

Each annotation task consists of the annotation guidelines at the top part of the screen. Following the guidelines is the parallel document where the annotations have to be performed, and a comment text area, with a button to submit the annotation. These three parts are shown in Figure 4.2.

In each annotation task the annotator is presented with the entire source and target document where pronouns of interest have been removed and replaced by a drop-down list of possible choices. The drop-down list does not have any pronoun selected by default and the annotator has to make a choice for each list. An example of the drop-down list is shown in Figure 4.3. The sentences from the source- and target- side document are vertically aligned so that corresponding sentences can be found quickly.

If the annotator chooses ANOTHER WORD in the drop-down list, then they are encouraged to write down the word that they have in mind. However, this is not enforced. The word can be entered into a special text field that is shown once a button next to the drop-down box is clicked. With the same button, a comment can also be left for each pronoun instance. A screenshot of these two form fields is shown in Figure 4.4.

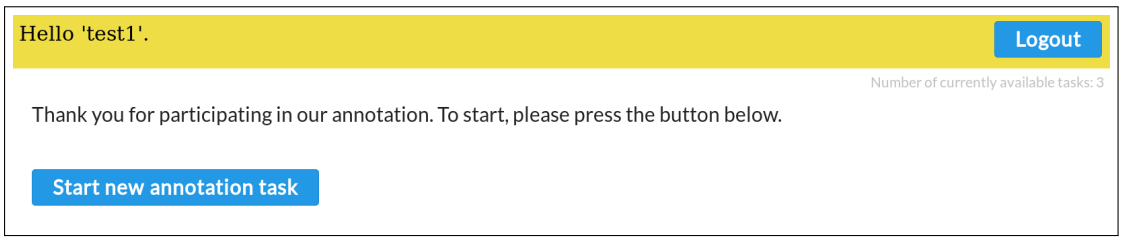


Figure 4.1: Welcome screen that is shown, when the annotator first logs into their account.

The annotation interface is programmed in the JavaServer Faces framework using the Apache MyFaces implementation.[5] The front end uses PrimeFaces.[6] The data is stored in an SQLite data base. The interface is accessed via a web browser. The server on which the application is run is Apache Tomcat.[7]

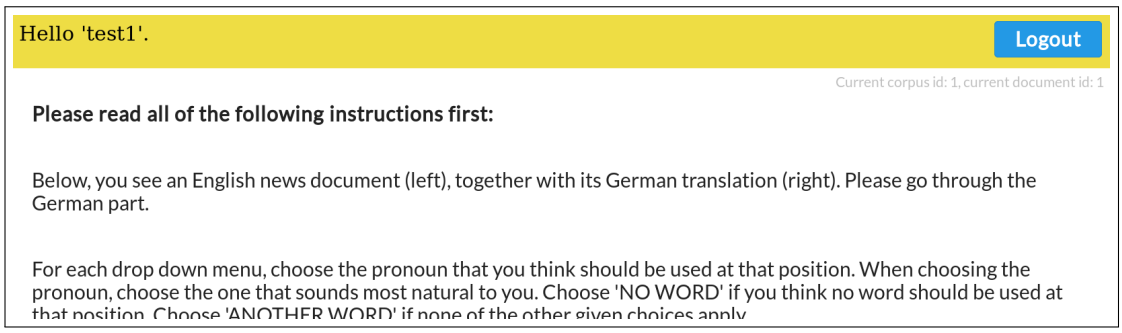
4.4 Annotation Process and Analysis

4.4.1 Annotation Guidelines

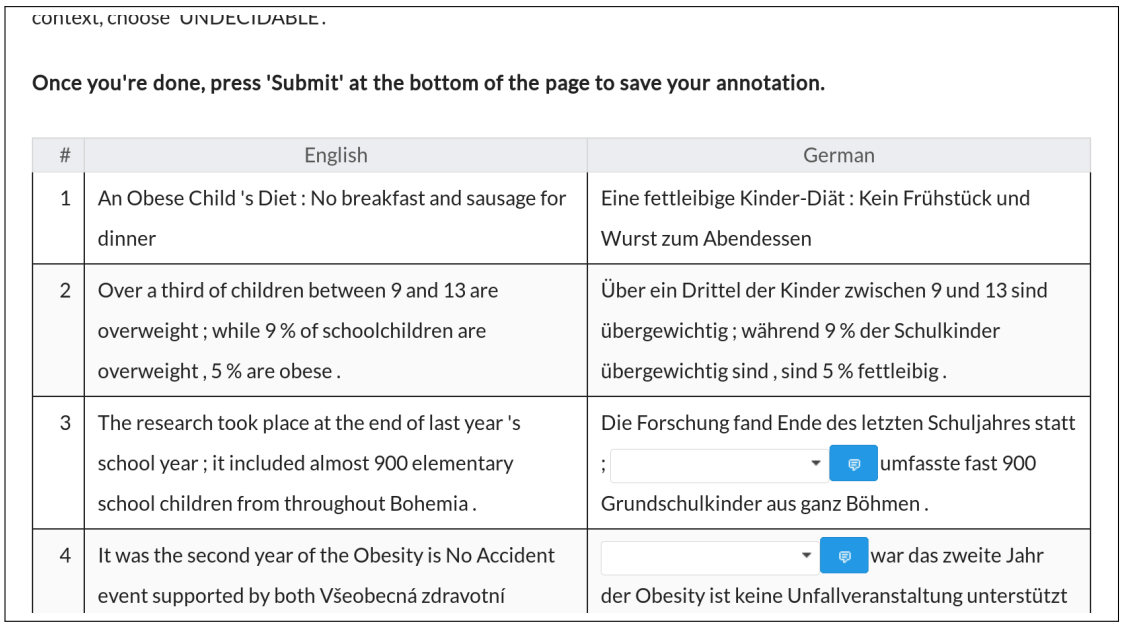
Each annotation task showed the following guidelines before showing the actual document that had to be annotated:

> **Please read all of the following instructions first:**
>
> Below, you see an English news document (left), together with its German translation (right). Please go through the German part.
>
> For each drop down menu, choose the pronoun that you think should be used at that position. When choosing the pronoun, choose the one that sounds most natural to you. Choose 'NO WORD' if you think no word should be used at that position. Choose 'ANOTHER WORD' if none of the other given choices apply.
>
> Each drop down menu has a blue button associated with it. There you can leave an optional comment. And if pronouns other than your first choice are also possible, or if you chose 'ANOTHER WORD' write down the word you would expect.

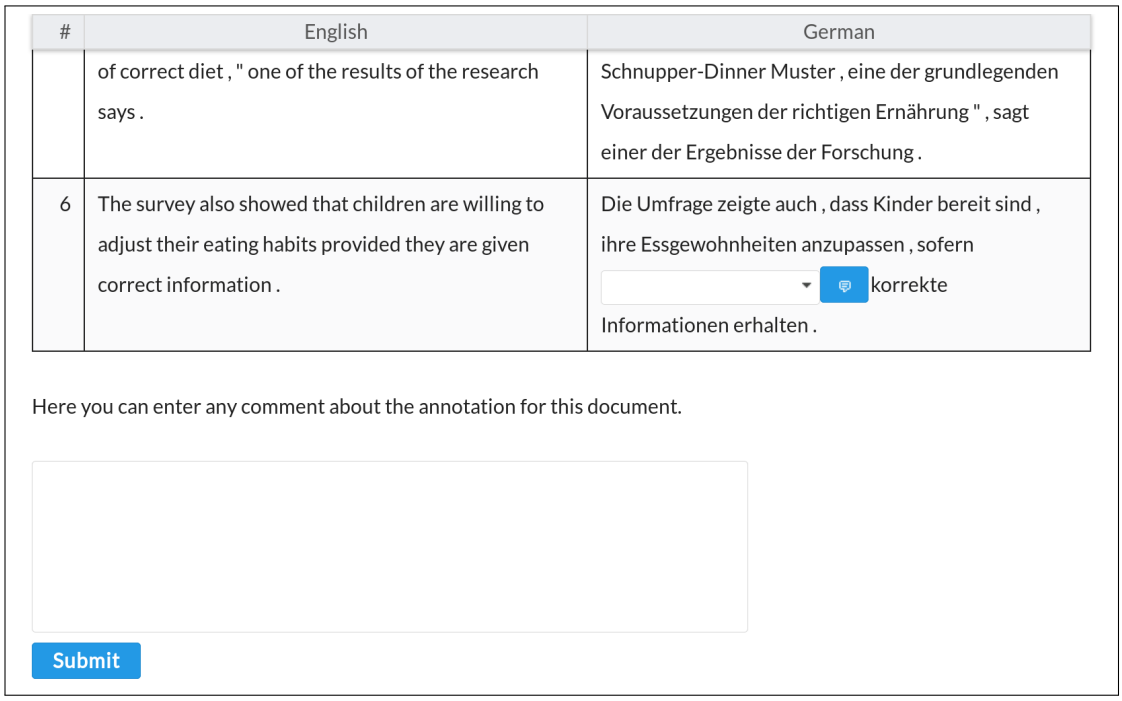
[5] https://myfaces.apache.org/
[6] https://www.primefaces.org/
[7] https://tomcat.apache.org/



(a) Partial view of the top part of the annotation screen showing the annotation guidelines.



(b) A part of the parallel document shown in the center part of the annotation screen.



(c) Comment text area and submit button shown at the bottom part of the annotation screen.

Figure 4.2: The three parts of the annotation screen that represents one annotation task.

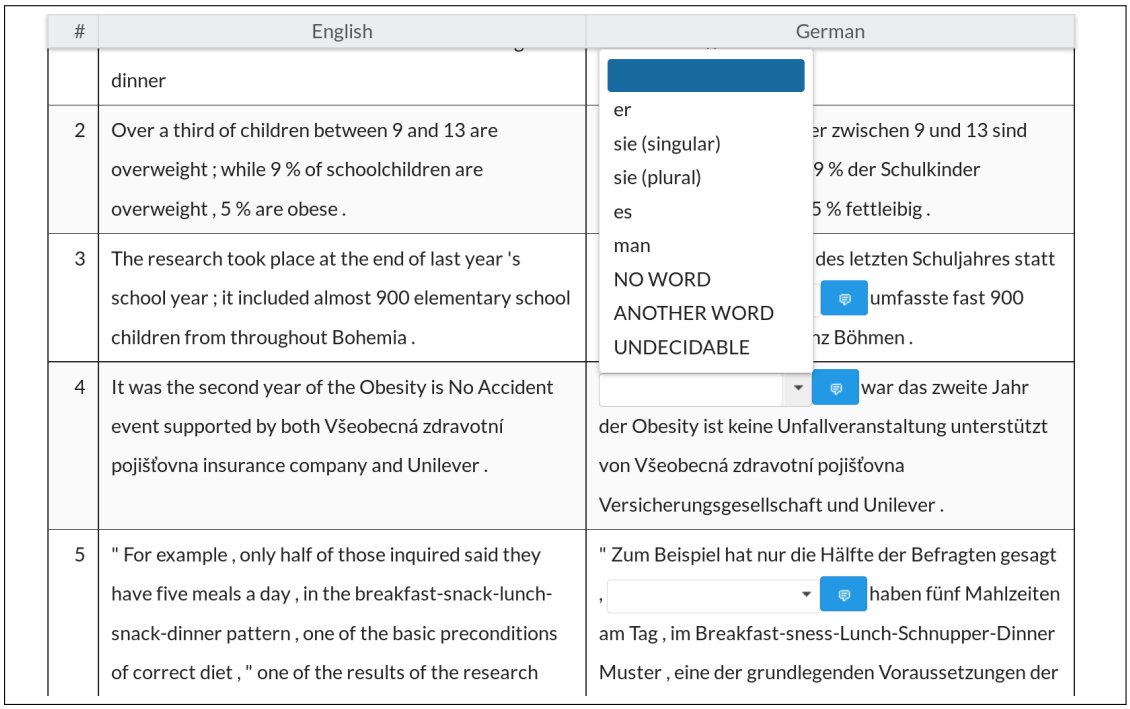


Figure 4.3: Screenshot: annotation screen – pronoun list.

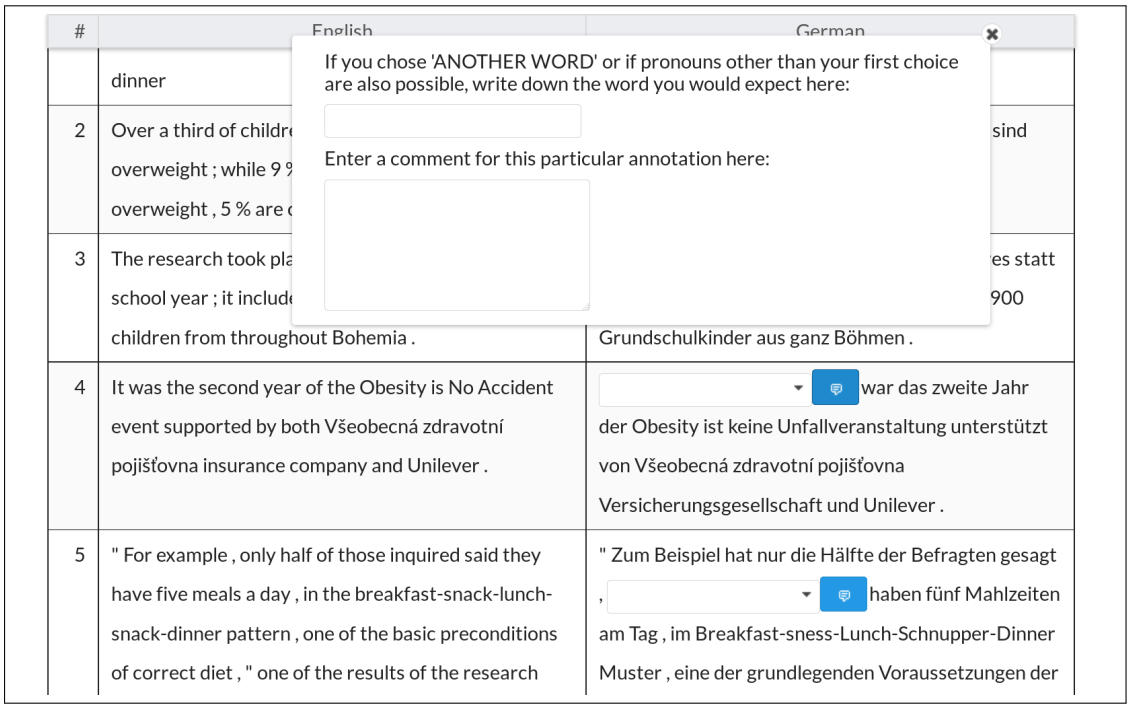


Figure 4.4: Screenshot: annotation screen – comment dialogue.

If the German translation of a sentence is too hard to understand on its own, you can refer to the original sentence in English (left) at your convenience.

You do not have to read the entire document, however you might have to read sentences preceding/following a drop down menu to understand the context and make an informed choice.

You must make a choice for each drop down menu.

If and only if you cannot make a choice even with the English sentence and surrounding English or German document context, choose 'UNDECIDABLE'.

Once you're done, press 'Submit' at the bottom of the page to save your annotation.

4.4.2 Inter-Annotator Agreement and Comparison

For determining inter-annotator agreement and to find out if the annotations contain a lot or only few ambiguities, which would result in lower agreement, we collected annotations for a small set of documents from the corpus with the pronoun list *p002*. Furthermore, a high inter-annotator agreement means that annotations from a single annotator are reliable enough to be used on their own. Two annotators (one was the author) annotated the same 15 documents each (equalling to 110 pronoun instances).

For inter-annotator agreement, we computed Cohen's Kappa,[8] which is 0.801. This can be considered a high inter-annotator agreement. Table 4.4 shows the confusion matrix between annotator 1 and 2 which shows us on which labels the annotators agreed and disagreed, and how frequently that happened. In the 110 pronoun instances that were doubly annotated, the pronoun *man* was never selected. The low frequency of *man* is also observed in the CLPP shared task corpora. Most of the pronouns are annotated as *es*, then *sie*. This also coincides with earlier observed frequencies in human translations. There are three UNDECIDABLE cases, two on which both annotators agreed. This gives some evidence, that the automatic translation – together with the source-side – can be understood by humans (i.e. they managed to make annotation choices), hence being of at least decent quality. Most of the annotations for *sie* are for the plural form. There are only very few confusions (i.e. disagreements) between the two classes *sie (singular)* and *sie (plural)*. This suggests that the same surface

[8] computed with the R package 'irr'

	er	sie (singular)	sie (plural)	es	man	NO WORD	ANOTHER WORD	UNDECIDABLE
er	3	·	·	·	·	·	·	·
sie (singular)	·	2	3	1	·	·	·	·
sie (plural)	·	1	21	1	·	·	·	·
es	4	·	·	57	·	·	·	·
man	·	·	·	·	·	·	·	·
NO WORD	·	·	·	·	·	3	2	·
ANOTHER WORD	·	·	·	1	·	·	8	1
UNDECIDABLE	·	·	·	·	·	·	·	2

Table 4.4: Confusion matrix showing agreements (diagonal values) and disagreements (off-diagonal values) between annotator 1 and 2. Rows represent labels annotated by annotator 1, columns represent labels annotated by annotator 2. Dots stand for zero.

form can be reliably disambiguated by humans. Furthermore, the two classes will not be treated differently in later experiments, so the small number of confusions will not have a negative impact.

We analysed the cases where annotators disagreed and categorized the type of disagreements. In the following table, we provide a detailed analysis of these disagreements:

ID	English sentence	German sentence	Comments
3	The research took place at the end of last year 's school year ; it included almost 900 elementary school children from throughout Bohemia .	Die Forschung fand Ende des letzten Schuljahres statt ; **sie (plural)/sie (singular)** umfasste fast 900 Grundschulkinder aus ganz Böhmen .	→ **annotation error**: annotator 1 wrongly assigned plural version of "sie"; ⇒ resolving to "sie (singular)"
	Conservative MP Andrew Bingham also told BBC bosses that the public service radio station appeared to be " dumbing down " and is becoming increasingly indistinguishable from its commercial rival .	Der konservative Abgeordnete Andrew Bingham sagte auch den BBC-Chefs , dass der öffentlich-rechtliche Radiosender " duzen " erschien und von seinem kommerziellen Rivalen zunehmend ununterscheidbar sei .	
8	" Radio 3 seems to be - I don 't like to use the word " dumbing down " - but it seems to be turning into Classic FM , " he said .	" Radio 3 scheint zu sein - ich mag es nicht , das Wort " dumbing down " zu verwenden - aber **es/er** scheint sich in Classic FM zu verwandeln " , sagte er .	→ **ambiguity**: both seem plausible (the first annotator treated it either as an event reference or as referring to the abstract concept "the radio" = "das Radio (neutr.)", the second annotator refers to "the radio station" = "der Radiosender (masc.)"); ⇒ resolving to "es" (since it is more natural)
16	Earlier this year , BBC Radio 3 controller Alan Davey argued that it has to work harder to engage audiences than it did in the past , because Britons are less educated about classical music .	Anfang des Jahres hatte BBC Radio 3 Controller Alan Davey argumentiert , dass **es/er** härter arbeiten muss , um das Publikum zu engagieren als **es/er** in der Vergangenheit , weil die Briten weniger über klassische Musik gebildet sind .	→ **ambiguity**: same argument as above; ⇒ resolving to "es" (since it is more natural; and it avoids creating antecedent ambiguities)

ID	English sentence	German sentence	Comments
160	A survey of 80 economists polled by Reuters found a little over half who only last week thought the Fed would go for it , now think it will hold fire a bit longer and keep rates at the current 0-0.25 percent range .	Eine Umfrage von 80 Ökonomen , die von Reuters befragt wurden , fand etwas mehr als die Hälfte , die erst letzte Woche dachte , die Fed würde UNDECIDABLE sorgen , jetzt denken , **sie (singular)/es** wird ein bisschen länger warten und die Zinsen im aktuellen 0-0,25 Prozent halten .	→ **annotation error**: error with second annotator; ⇒ resolving to ”sie (singular)”
	However the Fed cannot ignore the less rosy global outlook .	Allerdings kann die Fed die weniger rosigen globalen Aussichten nicht ignorieren .	
165	It has warned markets to be ready for a hike but indications are they also believe the odds are against such a move .	**sie (singular)/sie (plural)** hat die Märkte gewarnt , für eine Wanderung bereit zu sein , aber **sie (singular)/sie (plural)** glauben auch , dass die Chancen gegen einen solchen Schritt sind .	→ **annotation error**: both errors with second annotator; ”it” refers to ”the FED” which is a singular entity; note: this is a nice example where only context can disambiguate the ”it”; ⇒ resolving to ”sie (singular)” in both cases
168	After shooting and killing his girlfriend in Mississippi on Monday morning - and before he shot and killed his colleague later that day - Shannon Lamb wrote a note to say that he was ” sorry ” for the first murder and wished he ” could take it back , ” authorities revealed Tuesday .	Nach Schüssen und Tötung seiner Freundin in Mississippi am Montagmorgen - und bevor er seinen Kollegen später an diesem Tag schoss und tötete - schrieb Shannon Lamb eine Notiz , um zu sagen , dass er für den ersten Mord ” Entschuldigung ” sei und wünschte , er ” könnte **ANOTHER WORD(ihn)/es(ihn)** zurücknehmen ” , verrieten die Behörden am Dienstag .	→ **ambiguity**: both annotators provided one correct pronoun ”ihn” as their alternative pronoun choice. The second annotator chose to translate ”it” as event reference, however the first solution ”ihn” sounds more natural; ⇒ resolving to ”ANOTHER WORD”

ID	English sentence	German sentence	Comments
175	When police got to the scene they found the body of Amy Prentiss , 41 , and a handwritten note from Lamb , 45 , that said : " I am so sorry I wish I could take it back .	Als die Polizei auf die Bühne kam , fanden **sie (singular)/sie (plural)** den Körper von Amy Prentiss , 41 , und eine handschriftliche Notiz von Lamb , 45 , der sagte : " Es tut mir so leid , ich wünschte , ich könnte es zurücknehmen .	→ **annotation error** (debatable; due to agreement conflict): "die Polizei" is usually singular in German, therefore the first annotator is correct; however the second annotator chose the plural version most likely because it agrees with the translated verb; ⇒ resolving to "sie (singular)"
	It is awful - a boy who comes to Manchester United at 18 , has it very difficult and then plays fantastically and then this happens .	es ist furchtbar - ein Junge , der mit 18 Jahren zu Manchester United kommt , hat es sehr schwierig und spielt dann fantastisch und dann passiert das .	
219	When it was in the dressing room he had an oxygen mask on .	Als **es/er** in der Umkleidekabine war , hatte er eine Sauerstoffmaske auf .	→ **not applicable** (due to strange source sentence): the source sentence sounds strange; annotator 1 is consistent with the source, the choice by annotator 2 yields a better translation; ⇒ resolving to "er" (most likely this was the intended meaning)
239	I feel very bad about it , I am so sorry .	Ich fühle mich sehr schlecht **NO WORD/ANOTHER WORD (deswegen)** , es tut mir so leid .	→ **annotation error**: annotator 1 is wrong; annotator 2 is right; ⇒ resolving to "ANOTHER WORD"
326	As with his equivocation over Nato - Tom Watson is adamant that JC won 't campaign to quit - foreign diplomats will be obliged to try to make sense of it all for their masters at home .	Wie bei seiner Equipe über die Nato - Tom Watson setzt darauf , dass JC nicht Wahlkampf machen wird - ausländische Diplomaten werden verpflichtet , zu versuchen , für ihre Meister zu Hause Sinn **NO WORD/ANOTHER WORD (daraus)** zu machen .	→ **annotation error**: choice of annotator 2 is more consistent with the source sentence; choice of annotator 1 is still acceptable, but results in a worse translation, so annotator 2 should be favoured; ⇒ resolving to "ANOTHER WORD"

ID	English sentence	German sentence	Comments
332	PHE published the " landmark " report last month , describing it as a " comprehensive review of the evidence . "	Phe veröffentlichte im vergangenen Monat den " richtungsweisenden " Bericht , der **ANOTHER WORD (ihn)/UNDECIDABLE** als eine " umfassende Überprüfung der Beweise " bezeichnete .	→ **annotation error**: choice of annotator 1 results in an understandable, yet quite bad translation; annotator 2 could have made a choice here; ⇒ resolving to "ANOTHER WORD"
352	PHE has a clear duty to inform the public about what the evidence shows and what it does not show , especially when there was so much public confusion about the relative dangers compared to tobacco .	Phe hat die klare Pflicht , die Öffentlichkeit darüber zu informieren , was die Beweise zeigen und was **sie (plural)/es** nicht zeigt , vor allem , wenn es so viel öffentliche Verwirrung über die relativen Gefahren im Vergleich zum Tabak gab .	→ **annotation error** (due to agreement conflict): annotator 1 is correct, but his/her choice results in agreement error with the following verb; ⇒ resolving to "sie (plural)"

Table 4.5: Sentences (and preceding context if necessary) where the annotators disagreed. The last column categorises the disagreements and provides a decision as to what the most likely resolved annotation should be.

annotation error	9
ambiguity	4
not applicable	1
total	14

Table 4.6: Counts of the types of annotation differences between annotator 1 and 2.

Table 4.6 shows the counts of the types of annotation differences for both annotators together. Most of the annotator differences were due to annotation errors. However, seen over all the 110 pronouns annotated, the combined error (disregarding the ambiguous cases and the one *not applicable* case) of both annotators is 8.18%. Individually, annotator 1 made 3 errors (2.7%), annotator 2 made 6 errors (5.5%).

None of the *sie (singular)/sie (plural)* confusions are due to ambiguity, which further emphasizes the earlier conclusion, that humans can successfully handle and resolve ambiguous cases.

Three of the *er/es* confusions are due to different gender depending on what the referent is assumed to be referring to. The antecedent is not ambiguous, only the particular noun that is used to refer to it, i.e. *the radio station* (*der Radiosender*, masculine) vs. *the radio* (*das Radio*, neutral).

We also compared the *alternative pronouns* that annotators could provide, e.g. when choosing ANOTHER WORD or when more than one pronoun is plausible. There were 15 such cases. Among these, there were 8, where only one annotator mentioned an alternative pronoun, and 7 where both mentioned an alternative pronoun. The annotators provided identical alternative pronouns in those 7 cases.

The analysis in this section showed that we can expect a low amount of ambiguity and error in the annotation process. Together with the high inter-annotator agreement obtained on a subset of documents, we believe that annotation quality will be acceptable if we use one annotator for the remaining documents.

4.4.3 Full Annotation

In Table 4.7 we show the frequencies of how often a pronoun was used by a human annotator. The table compares annotations from annotator 1 and 2. The figures show a similar count for each pronoun across both annotators. Again (as already shown in the 15 documents which both annotators annotated), *es* is the most frequent pronoun, followed by *sie*. Also, *man* is the least frequent pronoun. Furthermore, the table also shows frequencies after merging annotations of both annotators while resolving differences as specified in Table 4.5.

We also recorded the amount of time each annotator spent annotating documents. In total annotator 1 spent 3.45 hours on the annotation, and annotator 2 spent 4.11 hours. Only the time spent between starting a new annotation task and clicking the submit button is measured.

4.5 Experiments

With the data prepared in the above described way, and with the human annotations in place, we run a set of experiments. Ultimately, we want to test how well our CLPP systems perform on the noisier SMT data set we created above.

First of all, we want to test how our trained CLPP systems perform on the CLPP 2016 test data set where we marked pronoun instances with our own implementation.

	raw counts			relative counts		
annotator	1	2	*m*	1	2	*m*
er	11	17	22	3.12	3.95	3.27
sie (singular)	15	21	34	4.25	4.88	5.05
sie (plural)	100	133	208	28.33	30.93	30.91
es	183	210	332	51.84	48.84	49.33
man	2	1	3	0.57	0.23	0.45
NO WORD	11	16	22	3.12	3.72	3.27
ANOTHER WORD	23	21	36	6.52	4.88	5.35
UNDECIDABLE	8	11	16	2.27	2.56	2.38
Total	353	430	673	100	100	100

Table 4.7: The statistics of different annotation choices for annotator 1 and 2. The total number is not equal, since none of the annotators annotated the entire corpus. Annotator *m* refers to the merged annotations. The relative counts are given in percent.

There are some differences to the officially described method. The biggest difference is that we do not perform subject filtering (Guillou et al., 2016, Section 3.3.3), thus having a bigger proportion of OTHER class labels. Any difference in performance in this experiment compared to the official shared task results will be due to the differences in implementation. This experiment acts as a sanity check of our implementation and to determine comparable performance scores for the following experiments.

In the second experiment, we want to see how well our systems perform on a different data set, i.e. on the WMT16 news test set, which formed the basis for our annotated corpus above. This helps us to identify what effect the changed data set has on the CLPP systems. In the official shared task setting, the target side consisted of human reference translations (lemmatized and POS-tagged). We therefore run in this experiment our trained CLPP systems also on the human reference translations of the WMT16 news test set, in order to keep the conditions as close as possible for better comparability.

Finally, in the third experiment we will test how well our trained CLPP systems perform on data that contains the automatically translated target-side. It is designed to understand how well CLPP systems perform in a more realistic setting that is closer

data set name	target-side created by	gold class labels
CLPP16test-official	human translator	automatic extraction
CLPP16test-ownimpl	human translator	automatic extraction
WMT16test-human-target	human translator	automatic extraction
WMT16test-auto-target	NMT system	manual annotation

Table 4.8: Overview of the test data sets used for the experiments that test our CLPP systems.

to the SMT scenario. Furthermore, here we compare the performance of pronoun translation of the NMT system against the pronoun predictions of our CLPP systems. This is the experiment where we can fully make use of our annotated corpus, since now we have gold labels available for each pronoun instance.

4.5.1 Data

For the first experiment, we compare the performance of our CLPP systems on two different data sets. The first one is the official CLPP 2016 shared task test set (henceforth: *CLPP16test-official*) that contains transcribed TED talks. The second one contains the same documents, however we marked pronoun instances with our own implementation (henceforth: *CLPP16test-ownimpl*) as described in Section 4.2.4. Both data sets have human reference translations. For the second experiment, we prepare the test set based entirely on the WMT16 news test set, which means the target side documents are human reference translations. We mark pronoun instances with our implementation and refer to this data set as *WMT16test-human-target*. The data for the third experiment comes completely from our annotated corpus, i.e. automatically translated documents from the WMT16 news test set. The gold labels for marked pronoun instances come from the manual annotations. We refer to this data set as *WMT16test-auto-target*. An overview showing the differences between the used data sets for the three experiments is given in Table 4.8.

The data in the above corpus for human annotation is unlemmatized (so that humans can understand and read the documents). However, for applying our trained CLPP systems, and to be as close as possible to the setting of the CLPP16 shared task, we further process the target-side data. All of the above data sets have their target-side lemmatized and POS-tagged. For the data sets we created ourselves, we used the Tree-

Tagger[9] (Schmid, 1994) for both lemmatization and POS-tagging. The POS-tags are then mapped to universal POS-tags (Petrov et al., 2012). This is in accordance with the CLPP16 shared task.

For all data sets without word alignments we obtain them in the same way as for our annotated corpus above (cf. Section 4.2.4). We ran MGiza with the *grow-diag-final-and* heuristics on a concatenation of Europarl7, NewsCommentary11, IWSLT15 and the respective data set.

4.5.2 CLPP Systems

We experiment with three different settings of our English-German CLPP system. All of them reuse the respectively trained models as described in Section 3.3.3 from the ALLINONE setup with the following variations in terms of feature usage.

None of the systems use the LM feature (*fLM*) since it was shown in the feature ablation study that it hurts performance. The first system (henceforth: CLPPPLAIN) was trained and is tested without the NONE feature (*fNone*). The second system (henceforth: CLPPNONEFEAT) was trained and is tested with *fNone*, while mapping all of the NONE predictions back to OTHER before evaluation. Finally, the third system (henceforth: CLPPNONEFEAT&PREDICT) was also trained and is also tested with *fNone* and the only difference between the latter two systems CLPPNONEFEAT and CLPPNONEFEAT&PREDICT is that in the latter system the NONE predictions are *not* mapped to OTHER. In those two systems all the predictions except NONE and OTHER are identical. Furthermore, this also means that CLPPNONEFEAT&PREDICT is the only system of the three CLPP systems that actually has NONE predictions in the final prediction output.

4.5.3 Results

Experiment I: Testing our Pronoun Instance Extraction Implementation

Results from the first experiment are shown in Table 4.9. When comparing between the performance of the official data set and the one preprocessed with our implementation, one can see a small drop for both metrics. For the first two systems, macro-averaged recall decreases between 5.18% and 6.02% (absolute) and accuracy decreases between

[9] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/; version 3.2.1; with Java interface from https://reckart.github.io/tt4j/

3.21% and 4.55% (absolute). The CLPPNONEFEAT&PREDICT system that has the additional NONE class label, the recall has a similar drop of 5.50% (absolute) as the first two systems, however there is a bigger drop for accuracy (12.06% absolute). Comparison of the confusion matrices shows that the main reason for that is a considerably worse prediction of the NONE cases.

One of the reasons for the overall drop in performance might be that we do not perform any filtering of pronoun instances according to *subject* relations of the source-side dependency tree. This is what was done in the CLPP 2016 shared task, considerably reducing the number of pronoun instances from the OTHER class. Therefore, in our data set, we have a larger OTHER class label proportion (135 vs. 166 pronoun instances with the OTHER gold label). Furthermore, the models we reused in this experiment were trained on data that was preprocessed with the shared task tools. The same differences mentioned above therefore also apply to the training data, resulting in a small mismatch between training and test data. Retraining the CLPP systems with the training data preprocessed with our implementation might increase performance on the test set, but is left for future work. Table 4.10 provides the full class label distributions with and without subject filtering in order to quantify the impact it has on the resulting test sets. It shows that the filtered *CLPP16test-ownimpl* data set more closely matches the one from *CLPP16test-official*, with OTHER and *es* seeing the biggest change. Finally, we obtained word alignments using the standard heuristics used in training SMT systems, whereas the shared task used precision-oriented word alignment heuristics.

Also note that since the CLPPNONEFEAT&PREDICT system has NONE predictions in the final output, we also adjusted the evaluation script to treat the NONE class separately from the OTHER class. This produces one more class than for the other systems, and results from this system are therefore not directly comparable to the results of the other two systems.

This experiment verified that our implementation of pronoun instance extraction works comparably to the official implementation, so that we can also apply it to other data sets in the following experiments. This experiment was mainly performed as a sanity check of our pronoun extraction implementation and to rule out compounding factors for the following experiments.

Experiment II: Performance on a Difference Domain and Genre

Results from the second experiment are shown in Table 4.11. We compare the results of each CLPP system with its performance in the previous experiment in Table 4.9.

	Mac-R	Acc
our official task submission	48.72	66.32
CLPPPLAIN	55.76	75.98
CLPPNONEFEAT	60.17	77.28
CLPPNONEFEAT&PREDICT	54.48	72.06

(a) Results on *CLPP16test-official*.

	Mac-R	Acc
CLPPPLAIN	50.58	71.43
CLPPNONEFEAT	54.15	74.07
CLPPNONEFEAT&PREDICT	48.98	60.00

(b) Results on *CLPP16test-ownimpl*.

Table 4.9: Experiment I: Macro-averaged recall and accuracy of our pre-trained CLPP systems on the CLPP 2016 shared task test set (both official data set, and data set preprocessed with our implementation).

	CLPP16test-official full	*CLPP16test-ownimpl* full	*CLPP16test-ownimpl* with subj filtering
er	15	16	16
sie	124	132	125
es	101	130	108
man	8	11	11
OTHER	135	166	119
Total	383	455	379

Table 4.10: Class label distributions from the *CLPP16test-official* and *CLPP16test-ownimpl* test sets. For the latter, we provide figures for the full data set without subject filtering, and for the reduced data set with subject filtering.

	Mac-R	Acc
CLPPPLAIN	49.41	73.69
CLPPNONEFEAT	48.07	73.84
CLPPNONEFEAT&PREDICT	43.45	60.69

Table 4.11: Experiment II: Macro-averaged recall and accuracy of our pre-trained CLPP systems on the WMT16 news test set with human reference translations (*WMT16test-human-target*).

This tests what influence the change of domain and genre of the test set has on the pre-trained CLPP systems. With respect to macro-average recall, one can observe that the performance drops by between 1.17% to 5.53% (absolute). With respect to accuracy, however, the results improve for two systems by 0.69% and 2.26% (absolute), while slightly decreasing for one system by 0.23% (absolute).

The two biggest influencing factors in this change of performance are the different genre and domain of the new test set. The CLPP16 test set contains TED talks, which are from transcribed speech, whereas the WMT16 test sets contain written newspaper articles. The other contributing factor is the difference in pronoun distributions. Unlike for the CLPP test data sets, where the organizers attempted to select documents with a larger number of rare pronouns (Guillou et al., 2016, Section 3.2), we did not impose this requirement on our test set in order to provide a more generic one (i.e. the full WMT16 news test set rather than a subset of it).

Experiment III: Performance on Automatically Translated Data

Results from the third experiment are shown in Table 4.12. We compare these results against the ones from the previous experiment given in Table 4.11, in order to identify what effect the automatically translated target-side has on CLPP performance compared to translations from human translators. Changes in performance are marked with arrows indicating the direction of change. The performance of our CLPP systems does not provide a consistent picture as to whether they perform better or worse on the data set with the automatically translated target-side. Two systems in fact increase their performance by between 0.94% and 4.94% (absolute) for macro-averaged recall, while one system decreases its performance by 4.77% (absolute) for the same measure. In terms of accuracy, one system decreases its performance by 1.19% (absolute)

	Mac-R	Acc
CLPPPLAIN	54.35↑	72.50↓
CLPPNONEFEAT	48.93↑	75.78↑
CLPPNONEFEAT&PREDICT	38.68↓	70.25↑
NMTBASELINEPLAIN	60.71	84.75

Table 4.12: Experiment III: Macro-averaged recall and accuracy of our pre-trained CLPP systems and the NMT baseline translation predictions on the WMT16 news test set (*WMT16-test-auto-target*, the only one for which we have human annotations) with automatic translations on the target side, where the pronoun instance translation are taken from the human annotations. Arrows represent increase or decrease of performance with respect to the results in Table 4.11.

and two systems increase their performance by between 1.94% and 9.56% (absolute). Only one system (CLPPNONEFEAT) continuously improves on the data set with the automatically translated target-side.

We also show the pronoun prediction performance of the NMT baseline system in Table 4.12. Comparing the performance of our CLPP systems against these predictions, we can observe that in fact the NMT system has a much higher performance. This result is unexpected, since the NMT system does not have access to context that goes beyond sentence boundaries. It thus has the same deficiencies as phrase-based SMT systems (cf. Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012). Our earlier experiments with the CLPP systems (cf. Chapter 3) showed that integrating non-local information increased performance of pronoun prediction. One major difference is that the NMT system was trained on a much larger data set and has thus better estimates. However, it remains to be investigated in more detail as to why the performance of the NMT system is better in terms of pronouns, and whether this holds in general or only on that specific data set. The confusion matrix in Table 4.13 reveals that the main confusions of the NMT system are between OTHER and *es*, and *sie* and *es*.

In order to better understand the high performance of the NMT system in predicting target-side pronouns, we conduct a more detailed analysis. Our hypothesis is that the NMT system performs very well on pronouns that are either non-referential or have their antecedent within the same sentence, and performance decreases for pronouns

classified as →	er	sie	es	man	OTHER	*Total*
er	2	9	11	·	·	22
sie	·	221	13	·	5	239
es	1	21	281	·	28	331
man	·	1	1	1	·	3
OTHER	·	1	10	1	62	74
Total	3	253	316	2	95	669

Table 4.13: Confusion matrix for the NMTBASELINEPLAIN predictions on the *WMT16-test-auto-target* test set. Row labels are gold labels and column labels are labels as they were classified. Dots represent zeros.

	Mac-R	Acc
NMTBASELINEPLAIN: *sentence-internal antecedent*	50.00	78.43
NMTBASELINEPLAIN: *cross-sentential antecedent*	33.33	87.08
NMTBASELINEPLAIN: *non-referential*	49.04	84.35

Table 4.14: Experiment III: Macro-averaged recall and accuracy of the NMT baseline translation predictions on the WMT16 news test set (*WMT16-test-auto-target*), separated by pronoun instances (a) with antecedent within the sentence, (b) with antecedent outside of the sentence, (c) that are non-referential.

with cross-sentential antecedents. We expect the former case due to the larger context and size of training data the NMT system has access to compared to our CLPP systems. The latter is expected because the CLPP system has access to selected information beyond the sentence context, which the NMT system does not.

To confirm our hypothesis we split the dataset into three subsets for each of the three cases (a) a pronoun instance has an antecedent within the sentence, (b) a pronoun instance has an antecedent outside of the sentence, (c) a pronoun instance is non-referential. The results are shown in Table 4.14. The NMT system has the lowest recall by a large margin of 15.71% (absolute) on the subset with pronoun instances that have cross-sentential antecedents compared to the other subsets. On the other hand it has highest accuracy on the same subset with a margin of 2.73% (absolute) compared to the other subsets. This suggests that on this subset it is very good at predicting the majority class and thus achieving a high accuracy, but very bad at the minority classes thus lowering recall. The confusion matrix in Table 4.15 confirms this. The model never predicts the classes *er, es* and *man* correctly (with *sie* as the majority class) on this subset. The results on the split test set confirm our hypothesis that the NMT performs much worse on pronouns with cross-sentential antecedents.

4.6 Conclusions

In this chapter, we presented an annotation procedure, a tool and a corpus for obtaining gold translations of pronouns in automatically translated documents. We showed that this is a manageable task both in terms of the required number of annotators and how much time is required for the annotation. This was shown with a high inter-annotator agreement. Furthermore, it was shown by analysing the annotation differences and finding out that only very few were due to ambiguities. These conclusions are based on the assumption that the SMT system used for creating the target side has a high (sentence-level) performance to start with. This is a reasonable assumption given the recent advances in SMT and NMT performance. Handling cross-sentence discourse phenomena requires a decent performance on lower-level phenomena.

In the experiment section, we put our manually annotated corpus to use in order to test how CLPP systems perform on realistic data, i.e. on automatically translated documents. Results could not provide a clear picture, as some systems were decreasing in performance according to one of the two measures we used. However, some systems were also increasing in performance. This shows that it might be necessary

classified as →	er	sie	es	man	OTHER	*Total*
er	2	2	·	·	·	4
sie	·	37	·	·	·	37
es	·	6	·	·	3	9
man	·	·	·	·	·	·
OTHER	·	·	·	·	1	1
Total	2	45	·	·	4	51

(a) Pronoun instances with sentence-internal antecedents.

classified as →	er	sie	es	man	OTHER	*Total*
er	·	6	·	·	·	6
sie	·	180	·	·	·	180
es	1	15	·	·	3	19
man	·	1	·	·	·	1
OTHER	·	1	·	·	2	3
Total	1	203	·	·	5	209

(b) Pronoun instances with cross-sentential antecedents.

classified as →	er	sie	es	man	OTHER	*Total*
er	·	1	11	·	·	12
sie	·	4	13	·	5	22
es	·	·	281	·	22	303
man	·	·	1	1	·	2
OTHER	·	·	10	1	59	70
Total	·	5	316	2	86	409

(c) Non-referential pronoun instances.

Table 4.15: Confusion matrices for the NMTBASELINEPLAIN predictions on the *WMT16-test-auto-target* test set that has been split into three subsets. Row labels are gold labels and column labels are labels as they were classified. Dots represent zeros.

to evaluate CLPP performance with a specific goal in mind in order to get the highest benefit out of these systems. Macro-average recall weighs performance on each target-side pronoun class equally. This means that prediction changes affecting a few rare classes influences macro-average recall considerably. This was deliberately chosen by the shared task organizers to encourage system submissions that perform well on the difficult cases. In a downstream task, however, it might be more desirable to have a high accuracy, which simply counts the number of correct pronoun predictions irrespectively of class frequency.

We direct our attention to evaluating the performance of CLPP systems in an extrinsic setting in Chapter 6. There we embed CLPP system predictions in the final output of an NMT system via post-editing. We then evaluate the performance of the complete translation with an evaluation metric based on full coreference chains. The CLPP system performance is therefore indirectly evaluated since pronouns form a part in coreference chains and wrong translations might break these chains.

With our corpus, we could identify that the performance in terms of pronoun prediction obtained from the state-of-the-art NMT system is unexpectedly high and exceeds the performance of our CLPP systems. For further insight it might be necessary to investigate how much influence the subject filtering and word alignment optimization has in the obtained results. Especially the former one has a direct influence on the class distribution and creates an artificial difference between training and test data. Retraining our CLPP systems on training data without subject filtering might provide a generally higher performance and possibly manage to exceed performance of the NMT system. Furthermore, the CLPP system was trained on data with artificially reduced information, i.e. on a lemmatized target-side. This provides the NMT system with an advantage, since it was trained on natural parallel text. This might partially explain the good performance of the NMT system. Adding back in the full word forms in the training data for the CLPP systems might boost their performance and might be part of a new instance of a redefined CLPP shared task.

This corpus enables the research community on CLPP and pronoun translation to evaluate their methods and analyse performance on data that is more realistic than what was previously done. The target side is closer to the data obtained during automatic translation compared to the test sets from the CLPP16 shared task where information was artificially removed by lemmatizing the target side of a human translation. With this corpus weaknesses of state-of-the-art CLPP systems can be exhibited, that remained hidden from the lemmatized human-authored tests sets. This corpus further-

more makes it possible to integrate suggested pronoun translations from the state-of-the-art NMT system as additional feature of a CLPP system. This way, the model has the potential to learn when a baseline pronoun translation can be relied on, or when a different pronoun prediction should be used.

Chapter 5

Bilingual Models of Coherence based on Entity Grid Models

In the previous chapters we focussed on pronoun translation in particular. It has been observed in previous work that pronouns cannot just be evaluated on their own (e.g. by comparing them to the pronoun in the reference translation). A wider context including antecedents of referential pronouns has to be taken into account. Furthermore, pronouns are part of the wider discourse phenomenon of entity-based coherence. As members of coreference chains they contribute to producing a coherent document, by connecting multiple occurrences of the same entity across the entire document.

There exists a wide range of work on entity-based coherence modelling in the monolingual setting (e.g. Barzilay and Lapata (2005); Filippova and Strube (2007); Barzilay and Lapata (2008); Cheung and Penn (2010); Elsner and Charniak (2011); Guinaudeau and Strube (2013); Tien Nguyen and Joty (2017)) used for tasks such as evaluating automatic summaries or reconstructing the order of documents where sentences are no longer in their original order. Alternatively, they have been used as an additional feature for judging the readability of texts. In this chapter we want to bring these monolingual models of coherence to the bilingual setting. We do this with the goal in mind to define an SMT evaluation metric that captures the coherence of a translation with respect to the source document. The translation setting enables us to not only model coherence of one language, but to exploit the coherence given by the source-side document and the meaning-preserving relation between source- and target-side coherence to define a richer and more informed model. We therefore design a bilingual coherence model that takes the relation between source- and target-side coherence into account, aiming at evaluating SMT systems based on their coherence.

However, before we apply this bilingual model of coherence to fully automatically translated data, we take the same approach of the CLPP shared tasks to define the problem in a simpler, more manageable way, which lets us experiment in a more controlled setting. The second advantage, one that also the CLPP shared task organisers exploited, is that in our simpler definition of the problem we can obtain automatic gold labels for coherence which is one of the major requirements for our experiments. For testing our approach, we focus on one particular document-level problem that SMT systems encounter. Pronoun translation in particular can have cross-sentence dependencies and has already been identified as a challenging problem in the CLPP shared task (Hardmeier et al., 2015; Guillou et al., 2016). Translation of pronouns affects the coherence of the target-side document, and introduces incoherence when translating to the wrong target-side pronoun. We consider typical errors CLPP systems make as a proxy of errors a pronoun-aware SMT system is likely to produce. This allows us to study the problem in a more controlled environment, free from noise and variation introduced by a fully automatic translation.

We first give an overview of the monolingual coherence model (i.e. the entity grid model (EGM)) which inspires our bilingual version (Section 5.1). We then investigate how the monolingual model behaves cross-lingually both based on gold standard and automatic entity extraction (Sections 5.2.2 and 5.2.3). We then present and analyse an automatic entity alignment method (Section 5.2.4) which forms the basis of our bilingual EGM (Section 5.2.5). Following this, we apply our bilingual model to ranking tasks in order to evaluate its performance on distinguishing gradually more and more confused data from the original human-authored translation and on distinguishing among the different confusions and test its generalization capabilities to unseen confusions (Section 5.3). We discuss the findings (Section 5.4) and give a conclusion (Section 5.5).

5.1 Monolingual Coherence Modelling with the EGM

Barzilay and Lapata (2008) present a monolingual model of entity-based coherence. In this model, a document is represented by a matrix where rows represent sentences and columns represent entities. Each cell in the matrix records whether a particular entity is mentioned in that sentence and optionally which syntactic role it fills. Distinctions about syntactic roles are made between SUBJECT (S), OBJECT (O) and OTHER (X). Multiple mentions of the same entity in one sentence are only recorded once with

the highest ranking syntactic role according to this ordering: S > O > X. The matrix also records the absence of an entity in a sentence. This entire matrix is called the *EGrid*. Once extracted from a document it is then converted into a feature vector which records adjacent entity mention transitions of length two or longer as relative frequencies. Ranking functions for these feature vectors are learnt for three different experiments in a supervised setup. The first experiment applies the model to distinguish documents with shuffled sentences from the original documents. The second experiment verifies that the ranking the model produces also correlates with human judgements of coherence. The final experiment combines additional features typically used for readability judgements with the EGM and assesses whether a text is easy or difficult to read.

Guinaudeau and Strube (2013) reformulate the above model and represent the EGrid in a bipartite graph with sentences and entities as the two mutually exclusive node sets. They call it the *EGraph*. An edge between an entity and a sentence records a mention of that entity in that sentence. The syntactic roles are encoded as edge weights where S, O, X are assigned 3, 2 and 1 respectively. The bipartite graph is then projected to a graph with just the sentence nodes (i.e. one-mode projection) where an edge between two sentence nodes is established if an entity is mentioned in both sentences. Furthermore, sentence nodes in the one-mode projected graph are connected via directed edges, which can only point from an earlier sentence node to a later one in the naturally occurring order of the sentences. The EGraph only captures entities that cross sentence boundaries, since singleton entities and entities that are mentioned strictly within one sentence do not influence the resulting one-mode projection. Furthermore, entities that are not mentioned in a particular sentence are no longer explicitly represented. Three different one-mode projections are formulated: P_U where an edge with weight 1 exists if at least one entity is mentioned in both sentences; P_W where the edge weight is the count of how many different entities are mentioned in both sentences; and P_{Acc} where the counts are weighted by the syntactic role weights.

$$
\begin{aligned}
P_U &: w(s_i, s_j) = min(1, |E_{ij}|) \\
P_W &: w(s_i, s_j) = |E_{ij}| \\
P_{Acc} &: w(s_i, s_j) = \sum_{e \in E_{ij}} w(e, s_i) \cdot w(e, s_j)
\end{aligned}
$$

where s are sentence nodes, e are entity nodes, E_{ij} is the set of entities occurring in sentences s_i and s_j and $w(e,s)$ represents an edge weight between two nodes in the bipartite graph, and $w(s_i,s_j)$ represents an edge weight between two sentence nodes in the one-mode projected graph. This is done for all sentence node pairs s_i and s_j, where $i < j$ (i.e. sentence s_i occurs before s_j).

Instead of learning a ranking function Guinaudeau and Strube (2013) show that they can use a score derived from the one-mode projections directly. They use the AOD which is defined as the sum of the weighted edges in the one-mode projection divided by the number of sentence nodes. The intuition behind this score is that if entities are mentioned more often across the document, there are more outgoing edges with a larger weight in a sentence node, resulting in a larger AOD. A larger AOD is the result of a more coherent document.

5.2 Bilingual Coherence Modelling

With cross-lingual coherence in mind our hypothesis is that there is a strong correlation between source- and target-side coherence. In the following analyses we verify that this hypothesis holds.

5.2.1 Monolingual Model Details

We reimplement the EGraph model and for obtaining the required entities, we use Stanford DCoref (Lee et al., 2013) for English and CorZu[1] (Klenner and Tuggener, 2011) for German documents. To obtain syntactic roles recorded in the EGraph, we obtain dependency parse trees for both the English and German documents with Stanford CoreNLP[2] and ParZu,[3] respectively. Noun phrases can span over several tokens, but generally one token is considered the semantic head of the phrase. This token also determines the syntactic role of the phrase. We therefore obtain the semantic head of a noun phrase span by searching the token whose syntactic head according to the dependency tree points to a token outside of the noun phrase span.

[1] http://www.cl.uzh.ch/en/research/completed-research/coreferenceresolution.html, v1.1

[2] github.com/stanfordnlp/CoreNLP

[3] github.com/rsennrich/ParZu

	P_U		P_W		P_{Acc}	
	en	de	en	de	en	de
AOD	0.3769	0.4125	0.3836	0.4210	2.5361	2.6257
Pearson	0.8911		0.8912		0.9332	

Table 5.1: The AOD (averaged over the ten documents) and Pearson correlation of the one-mode projection graph of three different types of projections (cf. Section 5.1) on the manually annotated TED part of the ParCor corpus.

5.2.2 Analysis of EGMs based on Manual Annotations

In this section we analyse the output of the EGMs on English and German documents individually and examine correspondences across the two languages. To get an idea of the upper bound of the performance of the monolingual EGMs, this experiment is based on manual annotations of entities. For this we need a parallel corpus (for English-German) with fully annotated coreference chains, however no such resource was available at the time of writing. The ParCor corpus (Guillou et al., 2014) contains annotations approximating full coreference chains. This corpus provides partial coreference chains where pronouns were manually linked to their closest noun antecedent. We therefore use the TED part (10 long documents) of this corpus for our analysis. We extract the one-mode projection of the EGraph on source- and target-side documents separately and compute the AOD for each document and language.

For cross-lingual comparison we compute the Pearson correlation between the AOD of the English documents and the AOD of the German documents to find out whether the source- and target-side coherence as measured by the separate EGMs are related. As noted in Section 5.1, Guinaudeau and Strube (2013) have shown that the AOD computed from the EGraph is an adequate measure of monolingual entity-based coherence. Our hypothesis is that the translation of a document should exhibit a similar coherence, since it is expected to preserve the meaning of the source document. The respective AOD scores are therefore expected to correlated with each other. Results are given in Table 5.1. The AOD averaged over all 10 documents is very similar in both languages. Furthermore, the Pearson correlation is very high (between 89% and 93%) which suggests that the coherence as measured by the AOD of a source- and target-side document pair is strongly connected, thus confirming our expectations.

	P_U		P_W		P_{Acc}	
	en	de	en	de	en	de
AOD	10.7684	3.8320	11.2735	4.0701	78.2543	24.6206
Pearson	0.6409		0.6357		0.7093	

Table 5.2: The AOD (averaged over the ten documents) and Pearson correlation of the one-mode projection graph of three different types of projections (cf. Section 5.1) on the automatically annotated TED part of the ParCor corpus.

5.2.3 Analysis of EGMs based on Automatic Annotations

The experiment in the previous section is based on manually annotated, but incomplete coreference chains. To get an idea of performance under complete, but automatically annotated coreference chains, we compute the AOD with entities resolved by the above mentioned coreference resolution systems (cf. Section 5.2.1). Results are shown in Table 5.2. The Pearson correlation between source and target coherence is lower than with manual partial coreference annotations (i.e. decreasing by 22.39 and 25.55 percentage points), but still high enough (i.e. between 64% and 71%) to support our hypothesis that there is a strong correlation between source- and target-side coherence. The averaged AOD is considerably different between English and German, with the latter having much lower values. One reason for the larger difference between source- and target-side AOD could be due to the difference in performance of the coreference resolution systems.

To get an intuition of how well the automatic coreference resolution systems work on the genre and domain of the ParCor corpus (i.e. TED talks), we manually completed the partial ParCor annotations to full coreference chains for two documents in both languages and computed the standard evaluation metrics MUC, B^3 and CEAF (cf. Section 2.5) using the reference implementation of the CoNLL shared task (Pradhan et al., 2014).[4] Results are given in Table 5.3. MUC shows a good F1 performance between 51.08% and 67.51% with a balanced precision and recall. For the other metrics, precision is generally a lot higher than recall by between 23% and 52% (absolute). All metrics show a similar performance when comparing source- and target side. Commonly the average of the F1 scores of MUC, B^3 and CEAF-e are reported as a final score. On the two source documents the average F1 score is 38.73, and on the two

[4] http://conll.github.io/reference-coreference-scorers/, v8.01

	MUC	B^3	CEAF-m	CEAF-e
docId	R, P, F1	R, P, F1	R, P, F1	R, P, F1
src009	66.66, 68.37, 67.51	29.89, 65.15, 40.98	31.08, 67.83, 42.63	6.97, 51.48, 12.28
src010	62.50, 51.28, 56.33	30.00, 53.44, 38.43	25.26, 57.66, 35.14	9.83, 58.98, 16.85
tgt009	60.71, 69.67, 64.88	25.12, 60.59, 35.51	28.81, 53.79, 37.52	10.84, 46.67, 17.59
tgt010	45.19, 58.75, 51.08	26.77, 64.08, 37.77	26.49, 66.66, 37.91	13.06, 64.67, 21.73

Table 5.3: Recall, Precision, F1 of standard coreference resolution evaluation metrics for two documents in English (source-side) and German (target-side). This evaluation includes singleton entities, and is based on automatically detected mention spans.

target documents it is 37.98. The results are far from perfect, but also only look at two documents. We speculate that this can also be attributed to the mismatch of our annotated entity mentions (and their spans) and the automatically extracted ones. General evaluation of the two coreference resolution systems as given in the respective papers show a much better performance on larger gold standard corpora. The average F1 measure for Stanford DCoref is 54.62 on the English CoNLL2011 dev set, and CorZu has an average F1 measure of 69.2 on the German SemEval 2010 data set. We give these scores as indication of the performance range on standard tests sets, however they are not directly comparable, since they are based on different data sets.

Furthermore, the EGraph ignores singleton entities, i.e. entities that occur only once, since the one-mode projection only establishes a connection between two sentences if entities are mentioned in both. The MUC metric in turn is the only metric that does not take singleton entities into account. This means, that the most relevant metric is also the one with the highest performance.

5.2.4 Entity Alignment

We manually inspected the EGraphs of source- and target-side documents by plotting them in a shared space with a single set of sentence nodes. This inspection suggested that there is a close correspondence between source- and target-side entities and that they can be aligned via the shared sentence node space of the two language-specific graphs. An excerpt of a source-target document pair is shown in Figure 5.1 with entities represented as *e* nodes in the source (src) or target (tgt) side, and sentences as *s* nodes. Edges are drawn between the two if the entity is mentioned in the sentence. The

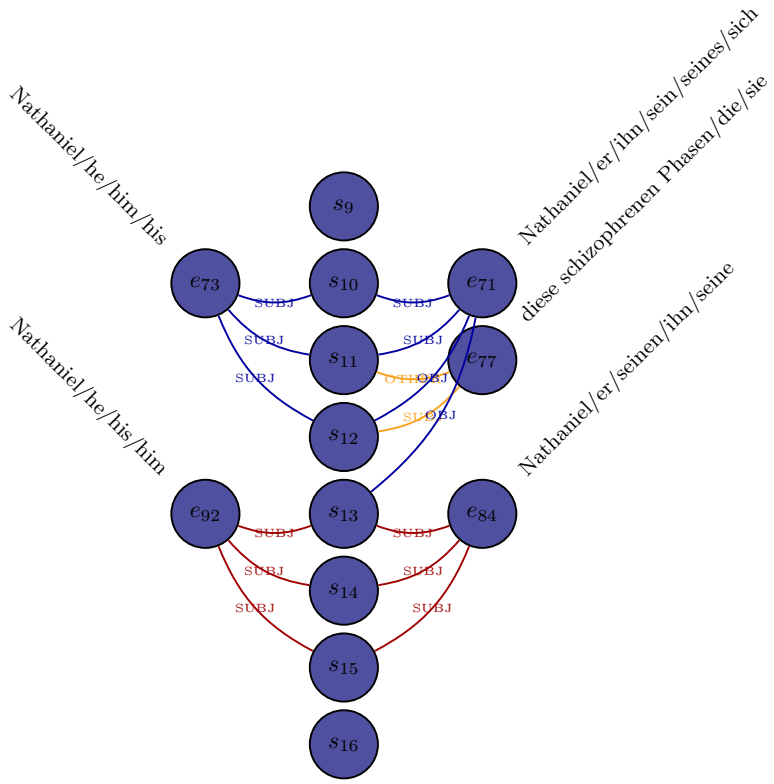


Figure 5.1: An excerpt of a source-side (left) and target-side (right) EGraph where the sentence nodes are shared between the two graphs. Extracted from the original ParCor document 009.

numbers provide an identifier for each node, but do not correspond to each other across the node sets. It can be seen that entity e^{src}_{73} (*Nathaniel, he, ...*) most likely corresponds with e^{tgt}_{71} (*Nathaniel, er, ...*), since they are both mentioned in sentences 10–12; and similarly entity e^{src}_{92} corresponds to e^{tgt}_{84}, since they both occur in sentences 13–15. It is highly unlikely that entity e^{tgt}_{77} (*schizophrenen Phasen, die, ...*) corresponds to any of the entities on the source side, since there is only little overlap. Furthermore note that the graph is based on the original ParCor annotations, which only provide partial coreference chains. With full coreference chains, the two separate entity nodes e^{src}_{73} and e^{src}_{92} (and similarly for the target-side) would be represented in one entity node.

In order to verify this intuition from manual inspection and to systematically find the entities in source and target side that correspond to each other we automatically align source- and target-side entities. We use the KM algorithm to obtain the optimal cross-lingual one-to-one entity alignment for a given document. The algorithm operates on a cost matrix with rows and columns as source and target entities respectively.

This matrix records the cost of aligning a particular source and target entity. We determine the cost in each cell by first counting how often the respective source and target entity is mentioned in the same sentences. Since the algorithm is defined to obtain a matching with minimal cost, we invert the counts by subtracting them from the maximum count. Furthermore, the algorithm requires a square matrix, so we add dummy entities with the maximum cost (so that they are most unlikely to get chosen), if we do not have an equal number of entities in source and target language. The above steps are illustrated in Table 5.4.

We manually inspected the aligned entities by looking at the source- and target-side entity mentions of each entity pair. The results on a subset of the ParCor corpus (documents 006–010) are summarized in Tables 5.5 and 5.6 for EGraphs based on manual ParCor annotations (i.e. partial coreference chains) and automatically resolved coreference chains, respectively. We distinguished three groups: (a) *corr*: the entity alignment is correct, (b) *incorr*: the alignment is wrong, (c) *context unk*: with just looking at the mentions it is not possible to judge whether the alignment is correct or incorrect (e.g. if a coreference chain only consists of pronouns). Automatic coreference chains may contain unrelated entity mentions, so we gave partial credit if an inconsistently resolved coreference chain contained at least some matching entity mentions between the aligned source and target entity (*partial*).

The automatic alignment of entities from EGraphs based on manual annotations performs well with 69% correct alignments. This gives justification that our method of entity alignment generally works. For the EGraphs based on automatic annotations, the performance drops considerably, but still 39% of the paired entities are correctly aligned.

In addition to the entity alignment procedure described above, we also experiment with two different versions. First of all, the KM algorithm can align a source- and target-side entity even though none of the entity mentions in either source or target occur in the same sentence (i.e. the cell in the cost matrix will have the highest cost). This is due to the fact that the algorithm does not leave any entity unaligned. We therefore check all aligned entity pairs and remove the alignment if none of the entity mentions in the source or target side occur in the same sentence (henceforth WITHPOSTKM-FILTERING). The second version takes word alignment into account, when creating the cost matrix. Word alignments are obtained with MGiza.[5] If a source and a target entity mention in a particular sentence is not connected via word alignment links, then

[5] https://github.com/moses-smt/mgiza

	e_0^{src}	e_1^{src}	e_2^{src}	e_3^{src}
s_0	X	X	-	-
s_1	X	-	X	-
s_2	-	-	X	X
s_3	-	X	-	X

(a) EGrid of source document.

	e_0^{tgt}	e_1^{tgt}	e_2^{tgt}
s_0	X	-	-
s_1	X	X	-
s_2	-	X	X
s_3	X	-	X

(b) EGrid of target document.

	e_0^{tgt}	e_1^{tgt}	e_2^{tgt}	$dummy_0^{tgt}$
e_0^{src}	2	1	0	0
e_1^{src}	2	0	1	0
e_2^{src}	1	2	1	0
e_3^{src}	1	1	2	0

(c) Weight matrix is created by counting the number of shared sentences for each source-target entity pairing. This is finally converted into a cost matrix by subtracting all weights by the maximum weight (here 2) to find minimum cost pairings.

	e_0^{src}-e_0^{tgt}	e_2^{src}-e_1^{tgt}	e_3^{src}-e_2^{tgt}	e_1^{src}
s_0	XX	–	–	X
s_1	XX	XX	–	-
s_2	–	XX	XX	-
s_3	-X	–	XX	X

(d) Optimal pairing of source and target entities found by the KM algorithm.

Table 5.4: Steps involved to find the optimal matching between source- and target-side entities. Sentences are represented by s and entities by e.

docId	corr	incorr	context unk
006	60.00	13.33	26.67
007	60.00	40.00	0.00
008	63.41	14.63	21.95
009	88.89	11.11	0.00
010	73.33	0.00	26.67
average	69.13	15.81	15.06

Table 5.5: Performance in percent of the automatic source-target entity alignment where the EGraphs are based on manual ParCor annotations. We ignore singleton entities when aligning.

docId	corr	incorr	context unk	partial
006	32.31	64.62	0.00	3.08
007	44.44	55.56	0.00	0.00
008	43.90	51.22	0.00	4.88
009	33.33	36.67	0.00	30.00
010	42.42	48.48	6.06	3.03
average	39.28	51.35	1.21	8.20

Table 5.6: Performance in percent of the automatic source-target entity alignment where the EGraphs are based on automatically resolved coreference chains. We ignore singleton entities when aligning.

this entity mention pair does not contribute to the weight in the cost matrix creation (henceforth WITHINFORMEDKMCOSTMATRIX).

5.2.5 Bilingual Model of Source-Target Coherence Interaction

The method from the previous section to automatically establish an entity alignment between a source- and target-side document allows us to define our bilingual entity model. It models the interaction between source- and target-side coherence, and is based on the hypothesis that there is a strong correlation between source- and target-side entities.

We experiment with two different versions. The first model provides a coarse-grained summary of the document similarly to the monolingual EGraph model. We create a new merged EGraph by exploiting the automatic alignment between source and target entities. The sentence nodes remain the same as for the individual EGraphs and for each sentence and aligned entity pair, we create an edge if that entity is mentioned in that sentence in both languages. The edge weight is computed by multiplying the edge weights of the monolingual edges of the involved entity mentions. If entity mentions do not have a counterpart in the other language, then no edge is established. Similarly, entities that were not aligned with the KM algorithm (because there were more entities in the source than in the target document, or vice versa) will not be in the new EGraph. For the aligned entities as determined in the example from Table 5.4, the resulting merged EGraph is shown in Table 5.7. The three variants of the one-mode projection can now be computed based on the new EGraph just as before and the respective AOD scores represent three variants of the coarse-grained version of our model.

The above version of the bilingual entity model is coarse grained because it only provides a one-score summary of bilingual coherence. The second version of our bilingual entity model looks at cross-lingual patterns in more detail and hence provides a more fine-grained picture. We extract patterns on the entity level and on the entity mention level. In the first group counts are collected of *aligned entities*, *unaligned source entities* and *unaligned target entities*. In the second group counts are collected only based on aligned entities by counting the number of *inserted mentions* and *deleted mentions* (i.e. an entity mention exists in the source sentence, but not in the target sentence, and vice versa respectively), and the number of *preserved mentions* (i.e. the entity is mentioned in both source and target sentence). Additionally, we count the

	$e^{merged}_{0,0}$	$e^{merged}_{2,1}$	$e^{merged}_{3,2}$
s_0	1	0	0
s_1	1	1	0
s_2	0	1	1
s_3	0	0	1

Table 5.7: The merged EGraph from the entity alignment shown in Table 5.4 (with X from that table taking on the weight 1 as defined before). The bipartite graph is represented as a matrix, where zero means no edge exists between an s node and an e node.

number of source and target entities. These eight counts represent the features of the fine-grained model in addition to the scores from the coarse-grained version.

We experiment with three different fine-grained versions. For the first (LEVELNORM) we normalize the counts within the entity level by dividing them by the sum of counts, and similarly for the mention level counts. The second version (NONORM) uses the counts directly. We experiment with this unnormalized version to find out what effect normalization has on the model. In the third version (SRCNORM) we divide the first two scores within the entity level group by the number of source entities and leave the other counts unnormalized. This normalization ensures the same denominator if comparing different translations of the same document, since the source-side remains the same. In Table 5.8 we summarize the involved features, their values and the different normalizations for our fine-grained model.

5.3 Experiments on Artificially Confused Data as Translation Proxy

To test our model, we focus on the problem of pronoun translation. This problem was also the main aspect of the CLPP shared task (Hardmeier et al., 2015; Guillou et al., 2016) to which we submitted systems (cf. Chapter 3). Participating systems were required to predict a translation of the pronouns *it* or *they* in subject position by choosing from a closed set of target-side pronouns given the source and target document and word alignments. The closed set consists of the most frequent pronoun translations and an OTHER class which groups together the remaining, less frequent translations.

group	no.	feature	levelNorm	noNorm	srcNorm
1	1	AOD of merged EGM: P_U	1	1	1
	2	AOD of merged EGM: P_W	1	1	1
	3	AOD of merged EGM: P_{Acc}	1	1	1
2	4	number of mention insertions	$\sum_{group2}$	1	1
	5	number of mention deletions	$\sum_{group2}$	1	1
	6	number of mention preservations	$\sum_{group2}$	1	1
3	7	number of aligned entities	$\sum_{group3}$	1	S
	8	number of unaligned source entities	$\sum_{group3}$	1	S
	9	number of unaligned target entities	$\sum_{group3}$	1	1
4	10	number of entities in source document	1	1	1
	11	number of entities in target document	1	1	1

Table 5.8: The values in the last three columns denote the denominator which is used for normalizing the feature value from the "feature" column. $\sum_{group2}$ = sum of counts in group 2. $\sum_{group3}$ sum of counts in group 3. S = number of entities in the source document.

We follow the underlying idea of the CLPP shared task to first experiment with our model in a controlled setting. Therefore, in order to investigate the performance of our model without the additional noise and variation introduced by automatically translated data (e.g. parsing and coreference resolution might not work well on the output of an SMT system), we work on an artificially created corpus, created from parallel documents where the target-side is taken from the human reference translation. We artificially introduce coherence errors in the data by confusing target-side pronouns to varying degrees. We base the confusions on typical errors CLPP systems make to get a more realistic confusion than just a uniformly random one. We achieve this by sampling from the distribution of pronoun confusions from the confusion matrices of CLPP systems. We take these confusions as proxy of discourse-related errors that a full SMT system with otherwise perfect translation performance would make.

The second motivation for working on artificially confused corpora is the fact that we automatically obtain labels for training and testing purposes with regards to coherence judgements. This is based on our assumption that if we confuse more pronouns, the coherence of the documents is also further away from the source document.

We conduct three experiments to test various aspects of our model. The first exper-

iment investigates whether the model can distinguish between a confused target-side document and the original version with respect to the source document. Our expectation here is that the coherence of a confused document exhibits systematic differences with respect to the source document that cannot be found in the original document with respect to the source document. The second experiment determines whether our model is capable of distinguishing among data sets with different amounts of confusions. With this we test if our model is generally suitable to rank different translation hypotheses of the same source document in terms of coherence. The third experiment tests whether we can reuse a learnt ranking function on a different confused corpus. This would be evidence in favour of the fact that the learnt ranking function is general enough to capture also differences between unseen confusions and the original corpus.

For these experiments, we learn ranking functions with SVM^{rank} (Joachims, 2006).[6] The output is a ranking that puts those documents at a higher rank that are closer to the coherence of the source document. The SVM ranker learns a function $h(\vec{x})$ such that for all input vector pairs $\vec{x}_i$ and $\vec{x}_j$ where the input vector $\vec{x}_i$ has a higher rank y than $\vec{x}_j$, the following condition holds:

$$h(\vec{x}_i) > h(\vec{x}_j) \Leftrightarrow y_i > y_j$$

The model parameters are learnt by minimizing the number of wrong ranking orders for a given training set of input pairs $\vec{x}_i$ and $\vec{x}_j$ with gold rankings $\vec{y}_i$ and $\vec{y}_j$ respectively.

5.3.1 Corpus

We take the IWSLT15 corpus for English-German as a basis for our experiments.[7] It contains transcribed documents from TED talks and is therefore from the same genre of the TED part of the ParCor corpus we explored in our earlier experiments. It contains 1592 source-target document pairs in total with 122 sentences per document on average.

5.3.2 Creation of Confused Data Sets

We prepare two different confused corpora. One is based on the confusion matrix obtained from applying the TurkuNLP system (Luotolahti et al., 2016, TURKUNLPTEST)

[6] http://www.cs.cornell.edu/People/tj/svm_light/svm_rank.html
[7] https://wit3.fbk.eu/mt.php?release=2015-01

System	Mac-R	Acc
TURKUNLPTEST	64.41	71.54
UEDINDEV	41.21	57.72

Table 5.9: Macro-averaged recall (Mac-R) and accuracy (Acc) of the two CLPP systems whose confusion matrices are used as a basis for introducing coherence errors.

to the CLPP16 shared task test set and the other one from applying our CLPP16 submission, i.e. the UEdin system (Wetzel, 2016, UEDINDEV) to the development set. Performance of these systems is given in Table 5.9. The first one performed better in the CLPP shared task both in terms of macro-averaged recall and accuracy. Since it makes fewer errors, the resulting confused corpus will have fewer changes from the original corpus.

For both corpus versions we remove the rows and columns from the confusion matrix representing the OTHER class. We do this because we cannot replace a pronoun with OTHER, since it does not stand for a particular pronoun and it would create unnatural corpora with the word *OTHER* in them. This leaves us with the entries for *er, sie, es* and *man*. We convert the counts in the confusion matrices to distributions from which we can sample. We then use these distributions to replace all target-side pronouns from the confusion matrix with randomly chosen ones. We replace pronouns at the beginning of the sentence with their first character converted to upper case. For all other occurrences we use lower case. One important difference to the CLPP shared task setup is that we replace target-side pronouns irrespectively of the source-side pronouns, i.e. we do not only replace target-side pronouns if their source-side counterpart is the pronoun *it* or *they* in subject position.

We experiment with two confusion matrix representations. The first one includes the values on the diagonal of the matrix, i.e. the correct predictions (henceforth: CMWITHDIAG). With this confusion matrix representation, pronouns can be replaced with themselves. This happens more often if the CLPP system has a higher accuracy. For the other representation we use confusion matrices, where the diagonal values are set to zero, i.e. each confusion results in a pronoun that is never itself. We refer to this version as CMWITHOUTDIAG.

type of confusion	model variant	accuracy
UedinDev, CmWithDiag	AOD of P_U	55.33
UedinDev, CmWithDiag	AOD of P_W	57.33
UedinDev, CmWithDiag	AOD of P_{Acc}	70.67
TurkunlpTest, CmWithDiag	AOD of P_U	52.00
TurkunlpTest, CmWithDiag	AOD of P_W	54.33
TurkunlpTest, CmWithDiag	AOD of P_{Acc}	68.67

Table 5.10: Performance of the coarse-grained model in terms of accuracy in percent on the single train/test split of 1442/150 documents.

5.3.3 Experiment I: Binary Ranking

In this experiment we test the general ability of our model to distinguish the *original document pair* from a document pair with the confused target-side pronouns (*confused document pair*). We frame this problem as a ranking problem, where the original document pairs should be ranked higher than their confused counterparts. We report results in terms of accuracy (i.e. the number of correct rankings divided by the number of all rankings). We run experiments on a single train/test split of 1442/150 documents and additionally perform a 10-fold cross-validation. The data consists of tuples of feature vectors, where the first element in the tuple represents the original document pair, and the second one represents the confused document pair. We learn separate ranking functions for each confused corpus. Our expectations are that the ranking performs slightly better for the confused corpus based on the UEDINDEV confusion matrix, than the TURKUNLPTEST one, since the former contains more confused pronouns and hence is more distinct from the original corpus.

The results for the coarse-grained model are given in Table 5.10 for the single train/test split. The performance of the first two variants is quite low (i.e. between 52.00-57.33% accuracy). The accuracy at 70.67% and 68.67% of the third variant is considerably better than that. The results for the fine-grained model are given in Table 5.11 for the single train/test split, and in Table 5.12 for the cross-validation experiments. The accuracy is much higher than with any of the coarse-grained models and very high in general (i.e. 88.00-96.00%). These results hold both for the single split experiment and the 10-fold cross-validation. This is a first piece of evidence that the

type of confusion	model variant	levelNorm	noNorm	srcNorm
UedinDev, CmWithoutDiag	both KmExtensions	92.00	94.00	95.33
UedinDev, CmWithoutDiag	only withPostKmFiltering	91.33	96.00	95.33
UedinDev, CmWithDiag	both KmExtensions	93.33	88.67	91.33
TurkunlpTest, CmWithDiag	both KmExtensions	88.00	84.67	88.00

Table 5.11: Performance of the fine-grained model in terms of accuracy in percent on single train/test split of 1442/150 documents.

type of confusion	model variant	levelNorm	noNorm	srcNorm
UedinDev, CmWithoutDiag	both KmExtensions	90.63±1.37	91.63±1.97	91.95±1.87
UedinDev, CmWithoutDiag	only withPostKmFiltering	90.38±1.36	92.07±2.59	92.14±2.48
UedinDev, CmWithDiag	both KmExtensions	90.69±1.62	89.75±2.08	90.75±1.97
TurkunlpTest, CmWithDiag	both KmExtensions	87.99±2.02	86.92±2.53	87.55±2.48

Table 5.12: Performance of the fine-grained model in terms of accuracy in percent averaged over results from a 10-fold cross-validation setup. Standard deviation is given after ±.

bilingual coherence model is a good model to judge consistent cross-lingual coherence. Furthermore, our expectations seem to hold that the TURKUNLPTEST corpus is harder to rank than the UEDINDEV corpus, since the ranking performance is slightly worse for the former one. The experiments with the confusion matrix variants CMWITHDIAG and CMWITHOUTDIAG show for most of the normalizations that the latter one is easier to rank (i.e. better performance of the model). The same argument holds here, that it contains more confused pronouns and is therefore easier to distinguish from the original document. Compared to the best-performing coarse-grained model, the fine-grained model still achieves a boost in accuracy of 25.33% up to 96.00%. This shows that the fine-grained model is more informative and better captures the bilingual coherence of the documents involved. In the remainder of the experiments, we will therefore only report results with the fine-grained model.

5.3.4 Experiment II: N-ary Ranking

The experiments in Section 5.3.3 show that there is a difference that can be captured between documents from various confused corpora and the original corpus. It also

	levelNorm	noNorm	srcNorm
Accuracy (pairwise)	83.33	82.89	84.22
Accuracy (three-way)	61.33	63.33	64.00

Table 5.13: Results in percent of three-way ranking experiment on the same train/test split of 1442/150 documents.

suggests that we can distinguish between the confused versions themselves. This is an important piece of evidence making these models useful for ranking different actual SMT system translations. To provide further evidence we run an experiment, where the data consists of triples of feature vectors. The first one represents the original document pair, the second one the TURKUNLPTEST-confused document pair, and the third one the UEDINDEV-confused document pair (*CmWithDiag, both KmExtensions*-version). The assumed gold ranking labels are given in that order as (3,2,1) for each triple, where a higher rank represents a higher coherence.

The results of this experiment are given in Table 5.13 for the single train/test split. In addition to pairwise accuracy as before (i.e. how often is any pair ranked correctly), we report the three-way accuracy, since now we have data triples. This is the number of ranking triplets where all rankings in each triplet are ranked correctly divided by the total number of ranking triplets. This metric more harshly penalizes the model, since each ranking triplet has to be perfectly predicted to count as correct. Performance of both metrics is very good, especially the second one, considering that any error in ranking one of the three data points is penalized. This is strong evidence, that our bilingual coherence model is capable of distinguishing different versions of confusions, thus being able to detect and judge entity-based coherence across two languages. This will also be important when it is applied to scoring different translation outputs of SMT systems for the same document.

5.3.5 Experiment III: Transfer of Learnt Ranking Function

In the final experiment we test if the learnt binary ranking function from experiment I based on one confused corpus A vs. the original corpus can be applied to rank a different confused corpus B against the original corpus. This tests whether the model can be applied to rank data confused by a different error distribution and if this is the case then we have evidence that we can apply the learnt ranking function to rank

	levelNorm	noNorm	srcNorm
train on: UedinDev, CmWithDiag test on: TurkunlpTest, CmWithDiag model type: both KmExtensions	89.33	87.33	88.00
train on: TurkunlpTest, CmWithDiag test on: UedinDev, CmWithDiag model type: both KmExtensions	92.67	89.33	90.00

Table 5.14: Accuracy in percent when applying a ranking function trained on a corpus with one degree of confusion vs. the original one to a corpus with another degree of confusion vs. the original one. Based on the same train/test split of 1442/150 documents.

different translation hypotheses of the same document.

Results are given in Table 5.14. They show a high accuracy between 92% and 89%, which confirms that the learnt ranking function is general enough and can successfully be applied to rank document pairs from a differently confused corpus against original ones.

5.4 Discussion

We first analysed a small set of documents to verify our hypothesis that source- and target-side coherence have a strong systematic connection. The analysis showed that there is a strong correspondence of source- and target-side coherence as measured by the EGraph formulation of the EGM both based on near-gold standard entity annotations and on automatically resolved annotations. This enables us to exploit entity-based coherence information with respect to the source document. In common translation scenarios the entire source document is usually available completely and is produced as natural language text by humans.

Based on the findings that there is a strong correspondence of coherence, we developed a bilingual model of coherence inspired by the monolingual EGM. We test its performance on three experiments. These experiments are designed to capture three different aspects of the model: (1) Can it distinguish original documents from documents with translation errors (i.e. confusions) with respect to the source documents?

(2) Can it distinguish also among different levels of confusions? (3) Can we learn a ranking function on one particular confused data set and apply it to a different one and still be able to distinguish the original from the confused documents?

The first experiment aimed at providing evidence whether our model performs well at capturing the coherence and the systematic connections of coherence between source and target side. We assume that the coherence of the original target-side document is closer to the source document than the confused target-side document, hence providing automatic gold ranking labels. The learnt ranking function is then expected to rank the original document pair higher than the confused one, which is confirmed by the results.

The second experiment aimed at providing evidence that the model can also be used to provide more than a binary ranking (i.e. original vs. confused document pair). So we tested whether we can rank document pairs of different degrees of confusion. This tests the general ability of the model to make decisions about whether a document pair is slightly better than another one, rather than comparing a perfect translation with an inferior one. This is working towards evaluating or ranking different translation hypotheses or different SMT system outputs. We obtain the different versions of confused data sets by using two different confusion matrices as basis for the confusion. We assume that the confusion from TURKUNLPTEST should always rank higher than the one from UEDINDEV. This assumption comes from the fact that the first system generally has a higher accuracy (i.e. more probability mass is on the diagonal of the confusion matrix distribution) and therefore fewer pronouns get changed (i.e. more pronouns get replaced with themselves). Investigation of both confusion matrices also showed that the first CLPP system generally performed very well on the non-referential pronoun *man*, therefore leaving more of these pronouns untouched and therefore introducing fewer referential pronouns that could potentially be resolved by the coreference resolution system. Furthermore, our assumption as before that the original document pair should be ranked highest still holds. The experimental results confirm what we wanted to show.

In the first two experiments we created the supervised ranking labels automatically based on well-founded assumptions, that our target-side confusions result in lower coherence than the original translations. However, in an SMT evaluation setting, we do not know the true rankings of different translation hypotheses and therefore have no training data for learning the ranking function. The final experiment is therefore designed to test whether it is possible to take a learnt ranking function from one particular confused data set, and see whether it generalizes and can be applied to rank another

version of the confused data set vs. the original data set. Results in the experiment confirm that this is possible in general.

All of the experiments are performed on data where errors of pronoun translation were introduced artificially. We take these confused corpora as a proxy of actual SMT output exhibiting one particular error commonly found in SMT system outputs. While we are aware that this does not reflect the noisier and lower quality of automatic translation, we still think that this is a useful proxy of such data. The advantage of this approach is the fact that we can work with almost natural translations, and can therefore rely on existing tools for parsing and coreference resolution.

The errors are introduced by sampling from confusions a CLPP system makes and are therefore not completely arbitrary confusions. However, the decision what pronoun to replace with what other pronoun is made independently of the surrounding context and therefore does not necessarily reflect the actual prediction of the CLPP system in that context. However, we cannot use the predictions of the CLPP systems directly, since the corpus we use in our experiments was part of the systems' training data. They would therefore have to be retrained and code or parameter settings are not always available. The approach of sampling from confusion matrices is therefore a useful compromise, since all the confusion matrices are published by the shared task.

Coherence has many aspects and with the EGM we only capture one of it. The conclusions of our experiments make no statement about other parts of coherence, such as discourse relations.

5.5 Conclusions

We compared the output of the EGM, a monolingual coherence model, for English and German and showed that there is a correlation between source- and target-side coherence. We then defined a coarse- and fine-grained bilingual coherence model. The former is based on the AOD of a merged source-target EGraph and the latter is defined as a feature vector of cross-lingual patterns on entity and entity mention level. To test different aspects of our model, we ran three ranking experiments, where target-side documents with a coherence more related to the source coherence are expected to be ranked higher than less coherent ones. We work with human translations and different degrees of confusion of these translations by substituting pronouns with other pronouns sampled from a confusion matrix which represents typical errors a SMT system could make. The experiments test whether the model can distinguish confused docu-

ment pairs from the original document pair, whether it can distinguish among different degrees of confusion, and whether we can learn a ranking function based on one degree of confusion and use it to rank another degree of confusion against the original translations. Results show that the coarse-grained model is not expressive enough resulting in poor performance. However, the fine-grained model has a high performance in all experiments. All experiments work towards using bilingual coherence for SMT evaluation and provide encouraging results.

Chapter 6

A discourse-aware Evaluation Metric for SMT

Recent work on document-level and discourse-aware SMT reported only minor improvements in performance according to BLEU when explicitly handling a particular discourse phenomenon (Le Nagard and Koehn, 2010; Rios Gonzales and Tuggener, 2017; Hardmeier, 2014, Chapter 9). However, BLEU can only capture strictly sentence-internal performance, is restricted to an n-gram context within sentences (conventionally n is 4) and gives equal importance to every word (cf Section 2.3). Furthermore, manual evaluation revealed that the covered phenomena are indeed preferred by the human judge (Guillou, 2012; Luong and Popescu-Belis, 2016). However, manual evaluation is not scalable. We therefore need an automatic discourse-aware SMT evaluation metric to enable and further the development of discourse-aware SMT approaches. Similarly to BLEU as tuning objective, this automatic metric could then also enable tuning SMT systems not only towards BLEU, but also towards a better handling of discourse-phenomena.

In the previous chapter we defined and experimented with our bilingual model of coherence, i.e. our bilingual EGM. We tested the model's performance on a task where translations should be ranked according to their coherence. This task resembles the SMT evaluation setting and in fact was set up with SMT evaluation in mind. We created artificial corpora, that were constructed in such a way, that we could automatically obtain gold ranks with a high certainty that they are reliable. The artificial corpora are based on parallel documents from TED talks with both source- and target-side created by human authors. We then confused the target-side by changing pronouns, where the change was informed by typical errors made by CLPP systems.

In this chapter, we move away from this artificial setup and experiment with our bilingual EGM to see if it can be used as a full discourse-aware SMT evaluation metric. We therefore need to show that it can be used when applied to real SMT system output. This also requires that we test the evaluation metric for correlation with human judgements. In the previous chapter, we could use automatically obtained gold labels of coherence from the artificially constructed corpora. This is no longer available in the setting for this chapter. Obtaining human coherence judgements is an expensive and not well-defined task. Instead, we employ our annotated corpus from Chapter 4 for this purpose. We use the pronoun translations chosen by human annotators, which are contributing to the entity-based coherence of a document, as a proxy for human coherence judgements. Based on these judgements, we compute a gold ranking of entity-based coherence between two system outputs, by counting which system agrees more often with the human pronoun choice.

If we can show that there is a correlation of our SMT evaluation metric with these human judgements, that shows that our evaluation metric adequately captures this particular concept of coherence. Instead of correlation, the equivalent view is if we can show that we can predict the gold rankings with a high accuracy. This is the approach we take in this chapter.

We first present the two SMT systems that we use to create translations as basis for testing our SMT evaluation metric and describe the test data set including how we obtained system rankings from our annotated corpus (Section 6.1). We then present how we use our bilingual EGM as evaluation metric and run two sets of experiments (Section 6.2). One is based on reusing pre-trained weights from our earlier chapter on artificial data, the other one uses a cross-validation setup. Both experiments are designed to find out if there is a high correlation with human judgements of our discourse-aware SMT evaluation metric. We then present a simpler evaluation metric based on information extracted from our bilingual EGM, but without the additional tuning procedure (Section 6.3). Finally, we discuss the overall results from this chapter (Section 6.4).

6.1 SMT Systems and Data

6.1.1 SMT Systems

In order to show the full potential of our discourse-aware SMT evaluation metric, we cannot rely on simply comparing two or more standard sentence-level SMT systems.

These systems could only introduce changes affecting the coherence of a translation at the sentence level. Furthermore, even those changes would have been performed by chance, since standard SMT systems do not explicitly model discourse-phenomena (neither at the sentence-level, nor at the document-level).

For testing our evaluation metric, we therefore work with output from two different SMT systems one of which is a discourse-aware system. The first system (henceforth: NMTBASELINEPLAIN) is the state-of-the-art NMT system Nematus (Sennrich et al., 2016a), that we used for our corpus creation (cf. Chapter 4). This system is not specifically designed to handle any discourse phenomena. We therefore take it as our discourse-unaware SMT system.

The second system is an automatic post-editing system that attempts to tackle pronoun translation. It is based on pronoun predictions made by our CLPP system. It takes the output of the above NMT baseline, and replaces pronouns on the target side, if the prediction is different from the baseline output. We experiment with three variants of this system based on different settings of our CLPP system. The settings of these variants are described in detail in Section 4.5.2. The first two (henceforth: NMTCLPPPLAIN and NMTCLPPNONEFEAT) differ, in that in the latter system the NONE feature is enabled, attempting to obtain better predictions for the OTHER class. However, none of these two systems have NONE predictions, i.e. that the source pronoun should be translated with the empty word, in their final output. On the other hand, the third variant (henceforth: NMTCLPPNONEFEAT&PREDICT) has NONE predictions in its final output.

Compared to the baseline NMT system, these post-editing systems attempt to tackle a discourse phenomenon that can be both sentence-internal, but more importantly can also cross sentence boundaries. They therefore make changes to the SMT output that can no longer be adequately captured with sentence-level evaluation metrics such as BLEU. This is confirmed by the respective BLEU scores in Table 6.1, which indeed do not change considerably. As such, our post-editing systems provide us with data as a basis for evaluating whether our SMT evaluation metric performs well. Table 6.1 furthermore shows the inadequacy of BLEU in that the translation output resulting from human pronoun annotation on top of the NMTBASELINEPLAIN output (HUMANC2, cf. Section 6.1.3) is in the same performance range as the BLEU scores for system outputs, with the NMTBASELINEPLAIN ranking higher with a small margin of 0.04 BLEU points.

In order to find out how different the output with respect to pronouns is between

	BLEU (uncased)
NMTCLPPPLAIN	34.54
NMTCLPPNONEFEAT	34.58
NMTCLPPNONEFEAT&PREDICT	34.72
NMTBASELINEPLAIN	34.82
HUMANC2	34.78

Table 6.1: Uncased BLEU scores on the WMT16 news test set.

	Mac-R	Acc
NMTBASELINEPLAIN vs. CLPPPLAIN	71.06	74.29
NMTBASELINEPLAIN vs. CLPPNONEFEAT	62.00	76.68
NMTBASELINEPLAIN vs. CLPPNONEFEAT&PREDICT	41.85	68.31

Table 6.2: Macro-averaged recall and accuracy on the WMT16 news test set (with automatically translated target-side documents) when comparing pronouns as translated by the NMTBASELINEPLAIN system against pronouns as predicted by each of our CLPP systems.

the NMTBASELINEPLAIN system and our post-editing systems, we compute a confusion matrix, where we consider the rows as the post-editing system predictions, and the columns as the baseline predictions. Based on this confusion matrix, we compute macro-averaged recall and accuracy as done in Section 3.2.2. These scores tell us how different the two sets of document translation hypotheses are. If the accuracy is very high, this means that there are not many differences. If it is low, then they propose many different pronoun translations. The results are shown in Table 6.2. The accuracy shows that between 68.31% and 74.29% of the pronouns do not differ. Macro-averaged recall is lower, i.e. between 41.85% and 71.06%. The reason for the lowest score is that the NMTBASELINEPLAIN system never predicts NONE, whereas the NMTCLPP-NONEFEAT&PREDICT system does, leading to a recall of zero for that class.

6.1.2 Data

We run our experiments on the WMT16 news test set, which also formed the basis of our annotated corpus. It consists of 155 parallel documents. For the automatic translation and for the individual pronoun predictions, we reuse the tokenized data created in Section 4.5.1. Note that for experiments in this section, we reuse the data *before* it has been lemmatized and POS-tagged. The translation of each document from the baseline system can then be taken directly from our corpus preparation. The post-edited translations of each document from the CLPP-enriched translation systems are obtained by replacing all the pronoun instances on the target side in the translation from the baseline with the pronoun predictions made by each CLPP system.

In case a CLPP system predicts NONE, we remove the target-side pronoun. With respect to OTHER-class predictions, we do not know what the actual word would be that the CLPP system prefers. The possibilities are too large since the OTHER class groups many different words together (i.e. other pronouns, nouns, or even words from other POS-classes). In our experiments, we therefore replace the baseline predictions with the token *OTHER* in these cases. Furthermore, we made sure that each *OTHER* token is unique by appending an identifier, so that the coreference resolution system does not accidentally resolve them into one coreference chain.

For our bilingual EGM we then pre-process the parallel documents with coreference resolution systems both on the source and on the target side. For the English source side, we use the Stanford DCoref system, and for the German target side, we use CorZu. These are the same tools we used in earlier experiments (Section 5.2.1).

6.1.3 Human Judgements for Gold Coherence Rankings

Eliciting human judgements of coherence in a document is a task that would be both expensive and is not well defined. If the term is left underspecified and human annotators were asked to either provide a judgement on a scale how coherent a text is, or to provide a ranking between two documents whether one is more coherent than the other, then this would require a lot of annotation time and effort. Annotators would have to read at least one (or possibly two) long documents and then give one final score for the entire document.

In our translation setting, this is further complicated by imperfect translations making it harder to understand the entire document and to judge only coherence and not be influenced in this judgement by grammatical and other errors. An additional layer of

complexity would be added if we required the annotator not only to judge the coherence of the target-side document, but to get a coherence assessment with respect to the source-side (e.g. verifying that all source-side entities are translated properly). Furthermore, coherence has many aspects (e.g. based on topics, discourse relations, event structure, entities, etc.) and if the term is left underspecified in a human assessment, then it would be unclear which concept of coherence would influence the annotator's decision. Finally, the WMT translation shared tasks elicit human judgements on a sentence by sentence basis, which simplifies the situation. However, even there they do not ask for judgements of each sentence of a document, but only sample sentences to be judged, as otherwise the task would be too time-consuming.

In their summary evaluation experiment, Barzilay and Lapata (2008) elicit coherence judgements via crowd-sourcing. However, in their setup, many of the mentioned problems do not occur. First of all, it is in a monolingual setting, not requiring the annotators to read documents in two languages. Second, the summaries they evaluate on are inherently short (i.e. around five sentences). Such short texts are faster to judge, even when asking for a ranking between two such short texts. Therefore, this specific monolingual setting is not comparable to our bilingual setting.

We therefore turn to our manually annotated corpus from Chapter 4. This parallel corpus consists of documents, where the translations were obtained with the NMT baseline system (NMTBASELINEPLAIN) we again use in the experiments in this chapter. The target-side pronouns have then been removed and human annotators were asked to provide a pronoun translation from a fixed set of choices. Hence, these annotations focus on one specific aspect of coherence, i.e. entity-based coherence. We take these pronoun annotations as a proxy for direct coherence judgements. This is based on the simplifying assumption that entity-based coherence is only influenced by pronoun choice. This is only an approximation, since in entity-based coherence coreferring nouns also play a role. Nevertheless, it provides a feasible data collection scheme with clearly defined annotation guidelines. Furthermore, when annotating the pronouns, the human annotators do not make a judgement about the grammaticality or fluency of the rest of the translation and therefore only focus on entity-based coherence.

We use the pronoun annotations in the corpus to create coherence rankings between two system outputs. In order to convert the pronoun annotations into ranked coherence judgements, we proceed as follows: For system A and B and for each document, we count how often pronouns from the system are equal to the human annotation. If system A has more such counts, then it gets a higher rank than system B, and vice

system A vs. system B	both right	both wrong	only A right	only B right	# of proInst
NMTBASELINEPLAIN vs. NMTCLPPPLAIN	272	67	100	22	461
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT	261	52	80	23	416
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT	279	49	104	37	469

(a) Counts from documents, where one system is favoured over the other system (i.e. documents **without ranking ties**).

system A vs. system B	both right	both wrong	only A right	only B right	# of proInst
NMTBASELINEPLAIN vs. NMTCLPPPLAIN	153	35	10	10	208
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT	181	46	13	13	253
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT	144	40	8	8	200

(b) Counts from documents, where both systems have the same count in one document (i.e. documents **with ranking ties**).

Table 6.3: The tables give counts that represent how often system A and B agree or disagree with human annotations for a particular pronoun instance (proInst). Four cases are possible, i.e. both system A and B agree with the human annotations, both systems disagree, only system A agrees, but not system B, and vice versa.

versa. If both system A and B have the same number of counts, then they get the same rank (i.e. ranking tie). In Tables 6.3a and 6.3b, we list the statistics for these counts separated for documents, which do not result in ranking ties and documents that result in ranking ties, respectively.

We take a consolidated version of the annotations from our corpus as basis for the above statistics and all of our experiments below (henceforth: HUMANC2). Whenever we have multiple annotations for a document (i.e. by two annotators), we take the one that is correct (if there are annotation errors) or the one that is most plausible (if they are all correct).

6.2 Bilingual EGM as SMT Evaluation Metric

In this section we want to answer the general question whether our bilingual EGM can serve as a discourse-aware SMT evaluation metric and whether it correlates with human judgements of coherence. If this is the case, then it can be considered a good and valid evaluation metric for assessing coherence in SMT output. To answer this question we design experiments in a ranking task setup. We conduct a series of experiments that are designed to answer different aspects of this general question.

6.2.1 Experimental Setup

In two experiments below (i.e. experiment I and II) we report results on two types of setup. In the first setup, we let our evaluation metric rank a human reference translation against a system output (i.e. *human vs. system*). In the second one, we let our evaluation metric rank two systems against each other (i.e. *system A vs. system B*).

The first type of setup (i.e. *human vs. system*) wants to answer the question, whether we can distinguish a perfect translation (with respect to pronouns) from an automatic translation that we assume is inferior. In particular this means that we want to find out if we can distinguish human gold output from system output. To generate system outputs, we use the SMT systems as described in Section 6.1.1. In this setting, we assign the human gold output the higher coherence rank, and the system output the lower coherence rank. This is based on the assumption that the automatic output exhibits a lower coherence than the output with human pronoun annotations. In quite a few cases, there is no difference with respect to pronouns between the human output document and the automatic translation (either because there are no pronouns in that particular document, or the automatic system predicted all pronouns in accordance with humans). In that case, we assign equal coherence ranks.

The second type of setup (i.e. *system A vs. system B*) wants to answer the question, whether we can rank two outputs from different systems according to the coherence of the produced documents. This is the more realistic setting, where either two completely different SMT systems can be ranked and the better one identified, or similarly the same system with different parameters. In the first type of setup, we assumed gold labels for coherence rankings. However in the *system A vs. system B* setup, we cannot do that. Here we use the human judgements converted into gold rankings of coherence as to which system produced a more coherent output from Section 6.1.3.

Both setups are binary ranking tasks, since in each setup two parallel input docu-

ments are compared. This binary ranking task has the following three possible rankings: (2,1), (1,2) and (1,1), where the first part refers to either human or system A output, and the second part to the system or system B output, respectively for the two types of setups. A higher number indicates a higher rank, i.e. 2 is more coherent than 1. If the rank is the same, then the pair (1,1) is used. In all experiments, our metric is expected to rank each parallel document pair in a ranking order that is most similar to the one given by the gold labels (either obtained by assumption, or by human judgements). Results are given in accuracy (i.e. how many rankings are ranked in the correct order). Furthermore, we report results only for the bilingual EGM which uses the LEVELNORM normalization setting (cf. Section 5.2.5), since the other normalization variants showed similar results in Chapter 5 and we do not expect them to differ here.

6.2.2 Experiment I: Reusing Weights from Trained Ranker

For this experiment we reuse a trained ranker from earlier experiments on the artificially confused corpus (cf. Section 5.3.3). If this works, then this would show that a ranker can be trained once, and can be reused on different data sets and different setups. The ranker for the earlier experiment was trained against data points that come from parallel documents where the target side is the human reference translation vs. parallel documents where pronouns in the human reference are randomly confused. In that experiment, we tried out two different types of confusions, which were referred to as *UedinDev, CmWithDiag* and *TurkunlpTest, CmWithDiag*. We now refer to them as *UedinDevDiag* and *TurkunlpTestDiag*, respectively.

6.2.2.1 Human vs. System

Results on the entire test set are shown in Table 6.4a. The highest accuracy is 66.45% providing good initial results. The pre-trained model parameters trained on the *TurkunlpTestDiag* corpus always perform better than the ones pre-trained on *UedinDevDiag*. Further inspection shows that there are quite a few parallel input document pairs that have an identical surface form for both human and the system (i.e. between 65 and 79 out of 155 documents are *input data ties*). These input data ties are trivially easy to rank, since they have an equal gold rank, i.e. (1,1), and the resulting bilingual coherence model scores are also equal due to the identical input data.

We therefore also look at the performance on input data pairs, which are strictly

different with respect to pronouns. Results for these are given in Table 6.4b. They are considerably lower, with the highest accuracy at 31.58%. Note, that the values in each row of the table without input data ties are not directly comparable. Each row reports results on a slightly different test set size (between 76 and 90 documents), since each *human vs. system* combination produces a different number of ties.

We more closely inspect the ranking errors by computing the confusion matrices of predicted rankings vs. gold rankings. The confusion matrices for experimental results reusing the *UedinDevDiag* model are shown in Table 6.5. Columns show the predicted rankings and rows show the gold rankings. First of all, we can see that the huge drop of performance between the full test data set (with input data ties) and the reduced test data set (without input data ties) can be mostly attributed to the fact that there are quite a lot of these input data ties. Additionally, these input data ties are trivially easy to rank (performance is perfect for this class) as noted before. The major source of error is that many input data pairs are ranked as ties (1,1), when they should have been ranked as untied (2,1). In other words, despite the fact that the input document pair is different, our metric is not capable of detecting the difference and predicts the same ranking (1,1). Note, we leave out the same type of table for the *TurkunlpTestDiag*-trained ranker.

6.2.2.2 System A vs. System B

In this set of experiments, we run the pre-trained ranker in the same way as in the previous experiment. Only now we rank two systems against each other, using gold ranking labels assigned by humans. Results are given in Tables 6.6a and 6.6b for the entire test set, and for the test set without input data ties, respectively. We omit results for the *TurkunlpTestDiag*-trained ranker. The number of input data ties is slightly smaller than in the *human vs. system* setting with a maximum of 71 input data ties compared to a maximum of 79 for the previous setting. The results are in the same range in *system A vs. system B* setting compared to the *human vs. system* setting from before, with a better maximum accuracy in the reduced test set without input data ties in the current setting (i.e. 35.11% compared to 31.58%).

We provide confusion matrices for the *system A vs. system B* experiments with the *UedinDevDiag* ranker in Table 6.7. Note that here again the numbers to the right side of the slash in the confusion matrices represent the experiments where we removed input data ties. On the reduced data set, in the *human vs. system* experiments we never have the tied ranking (1,1) as gold label, since in this setup only input data ties can

	accuracy	# of doc pairs	# of input data ties
used trained ranking model: UedinDevDiag			
HUMANC2 vs. NMTBASELINEPLAIN	65.81	155	79
HUMANC2 vs. NMTCLPPPLAIN	59.35	155	65
HUMANC2 vs. NMTCLPPNONEFEAT	60.00	155	70
HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT	58.06	155	67
used trained ranking model: TurkunlpTestDiag			
HUMANC2 vs. NMTBASELINEPLAIN	66.45	155	79
HUMANC2 vs. NMTCLPPPLAIN	60.00	155	65
HUMANC2 vs. NMTCLPPNONEFEAT	60.65	155	70
HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT	58.71	155	67

(a) Test data includes all 155 documents (i.e. **with input data ties**).

	accuracy	# of doc pairs	# of input data ties
used trained ranking model: UedinDevDiag			
HUMANC2 vs. NMTBASELINEPLAIN	30.26	76	0
HUMANC2 vs. NMTCLPPPLAIN	30.00	90	0
HUMANC2 vs. NMTCLPPNONEFEAT	27.06	85	0
HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT	26.14	88	0
used trained ranking model: TurkunlpTestDiag			
HUMANC2 vs. NMTBASELINEPLAIN	31.58	76	0
HUMANC2 vs. NMTCLPPPLAIN	31.11	90	0
HUMANC2 vs. NMTCLPPNONEFEAT	28.24	85	0
HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT	27.27	88	0

(b) Test data only includes documents **without input data ties**.

Table 6.4: Results of experiment I: human vs. system. Results are given in accuracy in percent.

	2,1	1,2	1,1	Total
2,1	23	19	34	76
1,2	0	0	0	0
1,1	0	0	79/0	79/0
Total	23	19	113/34	155/76

(a) HUMANC2 vs. NMTBASELINEPLAIN

	2,1	1,2	1,1	Total
2,1	27	31	32	90
1,2	0	0	0	0
1,1	0	0	65/0	65/0
Total	27	31	97/32	155/90

(b) HUMANC2 vs. NMTCLPPPLAIN

	2,1	1,2	1,1	Total
2,1	23	30	32	85
1,2	0	0	0	0
1,1	0	0	70/0	70/0
Total	23	30	102/32	155/85

(c) HUMANC2 vs. NMTCLPPNONEFEAT

	2,1	1,2	1,1	Total
2,1	23	30	35	88
1,2	0	0	0	0
1,1	0	0	67/0	67/0
Total	23	30	102/35	155/88

(d) HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT

Table 6.5: Confusion matrices of experiment I: human vs. system. These results are obtained using the trained ranking model *UedinDevDiag*. Each column represents the predicted ranking order, and each row represents the gold ranking order. The numbers to the left of the slash are from the experiments *with input data ties*, the ones to the right are from the experiments *without input data ties*.

	accuracy	# of doc pairs	# of input data ties
used trained ranking model: UedinDevDiag			
NMTBASELINEPLAIN vs. NMTCLPPPLAIN	61.94	155	71
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT	63.87	155	75
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT	60.65	155	61

(a) Test data includes all 155 documents (i.e. **with input data ties**).

	accuracy	# of doc pairs	# of input data ties
used trained ranking model: UedinDevDiag			
NMTBASELINEPLAIN vs. NMTCLPPPLAIN	29.76	84	0
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT	30.00	80	0
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT	35.11	94	0

(b) Test data only includes documents **without input data ties**.

Table 6.6: Results of experiment I: system A vs. system B. Results are given in terms of accuracy in percent.

result in such a ranking. However, in the *system A vs. system B* experiments, there is another way how gold rankings of (1,1) can occur. If system A makes an error in e.g. sentence 5, and system B makes an error in sentence 10, and nowhere else, they both receive the same rank according to the human annotation, since they both make exactly one error and predict every other pronoun correctly. In other words, document pairs that have a different surface form, can still result in (1,1) gold rankings. This explains, why the gold ranking row for (1,1) in the confusion matrices is not necessarily zero in the right-hand side confusion matrices. However, one can still see, that the number of these tied gold rankings (1,1) in the right tables drops with respect to the left-hand side of the table.

Compared to Table 6.5, in the confusion matrices in the current setup the confusions are more spread out across the whole range of the matrix. This seems to suggest that it is harder to rank two system outputs against each other than a system output against a human annotation.

6.2.3 Experiment II: Cross-validation on Realistic Corpus

The results in the above two experiments have good accuracy on the full test set, but a poor performance on the reduced set. One possible explanation for this is that the trained weights from the rankers that we reused from earlier experiments on different data (i.e. on corpora with artificially introduced coherence errors), cannot be applied to this data set and need to be retrained. The amount of data with gold labels we have at hand is quite small (i.e. 155 documents, and even smaller if input data ties are removed), so a fixed training-test split might not provide sufficient training or test data for reliable results. We therefore perform a 10-fold cross validation to get an average performance measure across the 10 folds. For these experiments, we use rankSVM integrated into LIBSVM (Kuo et al., 2014) with the default radial basis function kernel.

6.2.3.1 Human vs. System

Results for the *human vs. system* experiments are given in Tables 6.8a and 6.8b for the full test set and for the reduced test set without input data ties. With the full test data set, we can see that all *human vs. system* experiment pairs perform in the cross-validation experiment at least as good as, and in most cases 0.61% to 6.31% (absolute) better than their counterparts in the experiment using a pre-trained ranker (Section 6.2.2). With respect to the setting without input data ties, two *human vs. system* pairs are better in

	2,1	1,2	1,1	Total
2,1	17	17	22	56
1,2	1	4	5	10
1,1	3	11	75/4	89/18
Total	21	32	102/31	155/84

(a) NMTBASELINEPLAIN vs. NMTCLPPPLAIN

	2,1	1,2	1,1	Total
2,1	12	17	20	49
1,2	1	5	6	12
1,1	4	8	82/7	94/19
Total	17	30	108/33	55/80

(b) NMTBASELINEPLAIN vs. NMTCLPPNONE-FEAT

	2,1	1,2	1,1	Total
2,1	16	17	20	53
1,2	2	6	8	16
1,1	5	9	72/11	86/25
Total	23	32	100/39	155/94

(c) NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT

Table 6.7: Confusion matrices of experiment I: system A vs. system B. These results are obtained using the trained ranking model *UedinDevDiag*. Each column represents the predicted ranking order, and each row represents the gold ranking order. The numbers to the left of the slash are from the experiments *with input data ties*, the ones to the right are from the experiments *without input data ties*.

	accuracy
HUMANC2 vs. NMTBASELINEPLAIN	72.12±14.99
HUMANC2 vs. NMTCLPPPLAIN	60.00±14.49
HUMANC2 vs. NMTCLPPNONEFEAT	60.61±9.62
HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT	60.61±9.62

(a) Experiment data for the 10 folds includes all 155 documents (i.e. **with input data ties**).

	accuracy
HUMANC2 vs. NMTBASELINEPLAIN	40.69±13.05
HUMANC2 vs. NMTCLPPPLAIN	26.66±12.37
HUMANC2 vs. NMTCLPPNONEFEAT	23.41±13.58
HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT	29.54±16.23

(b) Experiment data for the 10 folds only includes documents **without input data ties**.

Table 6.8: Results of experiment II: human vs. system. Performance is given in terms of accuracy in percent, which in turn is averaged over the 10 folds. Standard deviation is given after ±.

experiment I (between 3.34% and 4.83%), and the other two are better in experiment II (between 2.27% and 9.57%). At least in some cases, learning new model parameters seem to help, however, in other cases, performance goes down showing that reusing the learnt model parameters is sometimes better.

The standard deviation of the averaged performance scores is relatively high (between 9.62 and 14.99) on the full data set, and even higher (between 12.37 and 16.23) on the restricted data set. This high variance might point towards the fairly small size of the test set folds, hence leading to overfitting to the training data.

As done in Table 6.5 for the pre-trained *human vs. system* experiment, we also show the confusion matrices from the *human vs. system* cross-validation experiments in Table 6.9. Each of these confusion matrices is obtained by adding up the confusion matrices from each of the 10 folds. Comparing the confusion matrices in the two tables from experiment I and II one can see that in the latter the strict ranking (2,1) is less frequently confused for HUMANC2 vs. NMTBASELINEPLAIN (but not for the other combinations), thus showing a partial benefit of learning new weights. The latter table also shows that a lot of times our metric cannot make a distinction between the data pairs, and as are result of that predicts the (1,1) ranking.

The minor variations between confusion matrices on the left and right (i.e. with and without input data ties) in Table 6.9 come from the fact that the training data in each fold is slightly different, since input data ties are missing in the latter setup.

One of the reasons, why there is a high number of equal rankings produced by our evaluation metric (especially in the restricted data set without input data ties) might be that the differences in coreference chains are not captured by this. Since the bilingual EGM only captures differences on mention references across sentence boundaries, it might be the case that in the parallel document pairs resulting in (1,1) rankings, there are fewer cross-sentential coreference chains. To test this hypothesis, we collect coreference chain statistics, separated by parallel document pairs resulting in (1,1) against all other ranking outcomes. In Table 6.10, we distinguish between pronoun instances that are part of a sentence-internal coreference chain, ones that are part of cross-sentential coreference chains, and ones that are not part of a coreference chain. The statistics confirm our hypothesis. The parallel document pairs resulting in a (1,1) ranking exhibit a much lower count of cross-sentential coreference chains. The counts also correlate with the performance of the different system pairings, i.e. a higher performance has a higher count of cross-sentential coreference chains.

6.2.3.2 System A vs. System B

Results for the *system A vs. system B* experiments are given in Tables 6.11a and 6.11b. In the cross-validation experiments for the full data set, there is an increase of performance of 2.80-2.91% (absolute) in two combinations and a slight decrease of 0.65% (absolute) in one combination compared to *system A vs. system B* results from experiment I with pre-trained models (Section 6.2.2). In the same comparison on the reduced data set, performance goes down by 0.21-3.80% (absolute) in two combinations, and up by 1.25% (absolute) in one combination. The standard deviation of the performance scores from the 10-fold cross-validation experiments is again quite large (between 6.96 and 21.10).

We also provide confusion matrices for the cross-validation experiment testing the *system A vs. system B* setup in Table 6.12. Compared to *system A vs. system B* confusion matrices from experiment I (cf. Table 6.7), one can observe that strict rankings are less frequently confused (i.e. fewer (2,1) rankings are predicted as (1,2)). This provides additional indication that relearning model parameters rather than reusing them is beneficial for ranking accuracy.

When comparing the cross-validation results from both setups in experiment II, one

	2,1	1,2	1,1	Total
2,1	30	11	35	76
1,2	0	0	0	0
1,1	0	0	79	79
Total	30	11	114	155

(a) HUMANC2 vs. NMTBASELINEPLAIN

	2,1	1,2	1,1	Total
2,1	31	10	35	76
1,2	0	0	0	0
1,1	0	0	0	0
Total	31	10	35	76

(b) HUMANC2 vs. NMTBASELINEPLAIN

	2,1	1,2	1,1	Total
2,1	24	33	33	90
1,2	0	0	0	0
1,1	0	0	65	65
Total	24	33	98	155

(c) HUMANC2 vs. NMTCLPPPLAIN

	2,1	1,2	1,1	Total
2,1	24	33	33	90
1,2	0	0	0	0
1,1	0	0	0	0
Total	24	33	33	90

(d) HUMANC2 vs. NMTCLPPPLAIN

	2,1	1,2	1,1	Total
2,1	22	30	33	85
1,2	0	0	0	0
1,1	0	0	70	70
Total	22	30	103	155

(e) HUMANC2 vs. NMTCLPPNONEFEAT

	2,1	1,2	1,1	Total
2,1	20	32	33	85
1,2	0	0	0	0
1,1	0	0	0	0
Total	20	32	33	85

(f) HUMANC2 vs. NMTCLPPNONEFEAT

	2,1	1,2	1,1	Total
2,1	25	27	36	88
1,2	0	0	0	0
1,1	0	0	67	67
Total	25	27	103	155

(g) HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT

	2,1	1,2	1,1	Total
2,1	26	26	36	88
1,2	0	0	0	0
1,1	0	0	0	0
Total	26	26	36	88

(h) HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT

Table 6.9: Confusion matrices of experiment II: human vs. system. Each column represents the predicted ranking order, and each row represents the gold ranking order. The left set of matrices are from the experiments *with input data ties*, the right set of matrices are from the experiments *without input data ties*.

	rank (1,1)		rank (2,1) or (1,2)	
	H	S	H	S
sentence-internal coreference chain	7.12	9.39	9.22	9.31
cross-sentential coreference chain	21.15	20.40	29.66	33.01
not part of coreference chain	71.73	70.22	61.12	57.68

(a) HUMANC2 vs. NMTBASELINEPLAIN

	rank (1,1)		rank (2,1) or (1,2)	
	H	S	H	S
sentence-internal coreference chain	4.79	5.80	9.45	5.76
cross-sentential coreference chain	18.78	18.77	27.94	30.97
not part of coreference chain	76.44	75.43	62.61	63.27

(b) HUMANC2 vs. NMTCLPPPLAIN

	rank (1,1)		rank (2,1) or (1,2)	
	H	S	H	S
sentence-internal coreference chain	3.65	6.18	9.66	4.57
cross-sentential coreference chain	14.67	15.18	29.17	31.44
not part of coreference chain	81.68	78.64	61.18	63.99

(c) HUMANC2 vs. NMTCLPPNONEFEAT

	rank (1,1)		rank (2,1) or (1,2)	
	H	S	H	S
sentence-internal coreference chain	1.87	3.25	10.3	3.93
cross-sentential coreference chain	14.64	14.43	28.98	29.71
not part of coreference chain	83.49	82.32	60.73	66.37

(d) HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT

Table 6.10: Experiment II: human vs. system. Statistics about parallel document pairs that have a predicted rank of (1,1) vs. all other rank predictions. H stands for the statistics of the human-authored documents, S stands for statistics of the system output. Counts are based on the target-side of the documents only. The figures are in percent and represent the respective values averaged over the entire test corpus.

	accuracy
NMTBASELINEPLAIN vs. NMTCLPPPLAIN	64.85±11.75
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT	66.67±10.64
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT	60.00±6.96

(a) Experiment data for the 10 folds includes all 155 documents (i.e. **with input data ties**).

	accuracy
NMTBASELINEPLAIN vs. NMTCLPPPLAIN	29.55±12.23
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT	31.25±21.10
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT	31.31±17.61

(b) Experiment data for the 10 folds only includes documents **without input data ties**.

Table 6.11: Results of experiment II: system A vs. system B. Performance is given in terms of accuracy in percent, which in turn is averaged over the 10 folds. Standard deviation is given after ±.

can see that it depends on the exact combination of system outputs involved whether the *human vs. system* or *system A vs. system B* combination performs better. This might seem counter-intuitive at first, since one would expect that the *human vs. system* combinations are easier to handle, since the human output should be considerably more different to a system than one system's output compared to another one. However, it has to be noted that the human output is not the full reference translation, but only contains the human annotation choices embedded in the NMT baseline output, similarly to the other post-editing systems.

We again seek to test our hypothesis that there is a high number of equal rankings (1,1) because parallel document pairs with these predicted rankings have a lower number of cross-sentential coreference chains. The statistics in Table 6.13 confirm this (only one system combination is shown). Compared to the statistics in the *human vs. system* experiment setup (cf. Table 6.10), we have a slightly higher count of cross-sentential coreference chains, which is also directly reflected in the fact that in the second setup there are fewer equal rankings predicted when they should not be equally ranked.

	2,1	1,2	1,1	Total
2,1	24	9	23	56
1,2	5	0	5	10
1,1	7	7	75	89
Total	36	16	103	155

(a) NMTBASELINEPLAIN vs. NMTCLPPPLAIN

	2,1	1,2	1,1	Total
2,1	21	12	23	56
1,2	5	0	5	10
1,1	7	7	4	18
Total	33	19	32	84

(b) NMTBASELINEPLAIN vs. NMTCLPPPLAIN

	2,1	1,2	1,1	Total
2,1	21	7	21	49
1,2	5	1	6	12
1,1	5	7	82	94
Total	31	15	109	155

(c) NMTBASELINEPLAIN vs. NMTCLPPNONE-FEAT

	2,1	1,2	1,1	Total
2,1	17	11	21	49
1,2	5	1	6	12
1,1	6	6	7	19
Total	28	18	34	80

(d) NMTBASELINEPLAIN vs. NMTCLPPNONE-FEAT

	2,1	1,2	1,1	Total
2,1	18	15	20	53
1,2	6	2	8	16
1,1	10	3	73	86
Total	34	20	101	155

(e) NMTBASELINEPLAIN vs. NMTCLPPNONE-FEAT&PREDICT

	2,1	1,2	1,1	Total
2,1	16	17	20	53
1,2	5	3	8	16
1,1	11	2	12	25
Total	32	22	40	94

(f) NMTBASELINEPLAIN vs. NMTCLPPNONE-FEAT&PREDICT

Table 6.12: Confusion matrices of experiment II: system A vs. system B. Each column represents the predicted ranking order, and each row represents the gold ranking order. The left set of matrices are from the experiments *with input data ties*, the right set of matrices are from the experiments *without input data ties*.

	rank (1,1)		rank (2,1) or (1,2)	
	A	B	A	B
sentence-internal coreference chain	5.37	4.54	10.33	4.21
cross-sentential coreference chain	13.85	16.28	33.89	30.59
not part of coreference chain	80.78	79.18	55.78	65.20

(a) NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT

Table 6.13: Experiment II: system A vs. system B. Statistics about parallel document pairs that have a predicted rank of (1,1) vs. all other rank predictions. A stands for statistics of the system A output, B stands for statistics of the system B output. Counts are based on the target-side of the documents only. The figures are in percent and represent the respective values averaged over the entire test corpus.

6.2.3.3 Accuracy With Prediction Ties

Accuracy is generally on the lower side if all three possible rankings are considered as equally important (which was done in the above experiments). A slightly different perspective on the coherence ranking problem is to give different weights to the severity of a ranking error. Predicting the opposite ranking between two parallel documents (e.g. predicting (2,1) for a (1,2) document pair) is definitely a serious error, however, predicting a ranking tie (e.g. predicting (1,1) for a (1,2) document pair) can be interpreted as not such a severe error. Sim Smith et al. (2016) take this view and provide results for both standard accuracy, and for $\text{accuracy}_{with-prediction-ties}$, where prediction ties are considered to be true positives no matter what the true gold label is.

In their experiments they rank human translations against automatically created ones for coherence, similarly to our *human vs. system* setup. In this setup their interpretation of $\text{accuracy}_{with-prediction-ties}$ is that it provides a score as to how often a human translation is ranked no worse than any of the automatic translations. A more general interpretation, which is then also relevant for our *system A vs. system B* setup, is that the $\text{accuracy}_{with-prediction-ties}$ gives a score as to how well the model performs at ranking parallel document pairs no worse than their true ranking (i.e. without reversing their ranking).

Therefore, we also compute the $\text{accuracy}_{with-prediction-ties}$ for both *human vs. system* and *system A vs. system B* settings and compare it to the standard accuracy scores reported in all of our previous experiments. Results are given in Table 6.14. These results show a much higher score for $\text{accuracy}_{with-prediction-ties}$ with a difference of

	accuracy		accuracy$_{with-prediction-ties}$	
	full	reduced	full	reduced
HUMANC2 vs. NMTBASELINEPLAIN	72.12±14.99	40.69±13.06	93.33±06.36	86.58±10.32
HUMANC2 vs. NMTCLPPPLAIN	60.00±14.49	26.67±12.37	80.00±11.72	63.33±18.63
HUMANC2 vs. NMTCLPPNONEFEAT	60.61±09.62	23.41±13.58	80.61±06.64	61.59±16.93
HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT	60.61±09.62	29.55±16.23	82.42±08.66	70.45±13.35
NMTBASELINEPLAIN vs. NMTCLPPPLAIN	64.85±11.75	29.55±12.24	83.03±10.00	63.64±18.04
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT	66.67±10.64	31.25±21.10	84.24±06.53	65.00±16.58
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT	60.00±06.96	31.31±17.61	78.18±07.57	62.12±15.50

Table 6.14: Standard accuracy compared to accuracy counting prediction ties as true positives for both *human vs. system* and *system A vs. system B* setups, on both full and reduced (i.e. without input data ties) test data sets in percent.

between 17.57% and 21.81% (absolute) compared to standard accuracy in both setups. These scores provide the improvement potentials of our evaluation metric if the high number of (1,1) predictions is tackled. Considering these results together with the positive results from the artificial experiments in Chapter 5, we can conclude that there is potential in our evaluation metric, if further development specifically focusses on tackling the (1,1) predictions (e.g. by making the metric more sensitive to entities mentioned only within the same sentence).

6.2.3.4 Random Baseline Performance

We perform an analysis of the results with regards to a random baseline performance. As random baseline we consider the expected accuracy of our model, defined as follows:

$$p_e = \frac{1}{N^2} \sum_{c \in C} (\mathrm{tp}_c + \mathrm{fn}_c) \times (\mathrm{tp}_c + \mathrm{fp}_c)$$

where $\mathrm{tp}_c, \mathrm{fn}_c$ and fp_c are true positives, false negatives and false positives of a given class (as obtained from a confusion matrix) and N is the number of parallel document pairs in the test set. The expected accuracy provides an intuition of the proportion of rankings our model could achieve correctly by chance. It is also used in the definition of Cohen's Kappa coefficient κ (Cohen, 1960):

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed accuracy (i.e. as reported in the experiments above). For each fold in the cross-validation we compute expected accuracy and Cohen's Kappa,

	expected accuracy		Cohen's Kappa	
	full	reduced	full	reduced
HUMANC2 vs. NMTBASELINEPLAIN	50.14±11.66	40.69±13.06	47.41±23.90	0.00±0.00
HUMANC2 vs. NMTCLPPPLAIN	37.37±07.65	26.67±12.37	37.78±21.05	0.00±0.00
HUMANC2 vs. NMTCLPPNONEFEAT	39.39±05.13	23.41±13.58	35.14±14.95	0.00±0.00
HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT	39.15±05.54	29.55±16.23	35.62±13.69	0.00±0.00
NMTBASELINEPLAIN vs. NMTCLPPPLAIN	48.36±05.24	36.22±06.69	32.53±20.50	-10.13±12.97
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT	50.10±03.55	33.91±10.34	32.38±22.80	-3.17±23.78
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT	44.85±05.48	32.02±07.70	26.55±15.91	-1.05±24.56

Table 6.15: Expected accuracy and Cohen's Kappa coefficients in percent for both *human vs. system* and *system A vs. system B* setups, on both full and reduced (i.e. without input data ties) test data sets.

and provide the average and standard deviation of each in Table 6.15. The observed accuracy used for the calculation of Cohen's Kappa is taken from Table 6.14.

First, when considering the *human vs. system* setup, one can observe that for the reduced test sets the observed and expected accuracies are the same. This is due to the fact that in this setup there are no (1,1) or (1,2) gold rankings, and therefore in those cases the expected accuracy is always equal to the observed accuracy. This also means that for those cases Cohen's Kappa values are not informative. This issue shows that it is difficult to get meaningful results from a random baseline in the case where only one class is actually observed in the test set. For the full test sets the (1,1) gold ranking is possible in addition, and therefore the expected accuracy can be different from the observed one. For the *system A vs. system B* setup, the expected accuracy and Cohen's Kappa can be interpreted on both full and reduced test sets, since all three possible rankings exist in both sets.

Comparing the observed accuracy (cf. Table 6.14) with the expected accuracy (cf. Table 6.15) one can see that for the full test data sets, the observed accuracy is always much higher (between 11.98% and 22.63% absolute) than the expected accuracy. This is also reflected in Cohen's Kappa scores, ranging between 26.55% and 47.41%. For the reduced test sets (in the *system A vs. system B* setup), however, the expected accuracies are in fact higher than the observed ones (by between 0.71% and 6.67% absolute). This is also reflected in negative Kappa scores, ranging between -1.05% and -10.13%. These results lead to the conclusion that the model in its current form performs worse than what it is expected to perform by chance.

6.2.4 Discussion

On the full data sets, all experiments showed a good accuracy score around 60%. This initially seems to suggest, that our SMT evaluation metric can successfully rank translations in accordance with both assumed (i.e. *human vs. system*) and human (i.e. *system A vs. system B*) rankings. However, a closer inspection revealed that a large portion of this high accuracy score is due to parallel document pairs that are trivially easy to rank, since they are identical with respect to pronoun translation. They are either identical because they do not contain pronoun instances to begin with, or the involved systems predicted the same pronouns for each pronoun instance in a document. When we removed these data pairs from the corpus, accuracy dropped considerably to around 31% on average.

Our SMT evaluation metric makes a large number of (1,1) predictions, i.e. ranks the document pairs equally. We investigated why this is the case and found out that in these document pairs, the number of pronoun instances that are part of cross-sentential coreference chains is much lower compared to document pairs, where strict rankings (2,1) or (1,2) are predicted. In fact, the former document pairs exhibit fewer coreference chains in general. Our SMT evaluation metric only captures cross-sentential coreference chains, which explains why it cannot distinguish well in those cases, yielding a high number of (1,1) predictions.

In all cross-validation experiments we observed a high variance of accuracy (with the standard deviation ranging between 6.96 and 21.10 percentage points). This is an indicator that our data sets are too small, leading to overfitted parameters of our model. An additional pattern can be observed in that the variance increases in five out of seven experimental settings between the full test data set and the reduced data set. This further shows that the even smaller reduced data sets lead to an even higher variance. The results therefore have to be interpreted with care. Testing our evaluation metric on a larger and more diverse data set (e.g. where the pronouns are not just predicted by the same CLPP model with different parameters, but actually different models) might lead to more reliable results (either rejecting or confirming our model as a good SMT evaluation metric).

Contrary to our expectations, reusing a trained ranker from the artificially confused corpora from Chapter 5 worked as well or better than retraining our SMT evaluation metric on the gold ranking labels obtained by human annotations in a 10-fold cross-validation. One explanation for this is, that the training folds are very small, hence,

the learnt weights are overfitted and prediction on the test fold performs poorly. This is confirmed by the high variance of accuracy.

In summary, the main issues that we identified leading to this overall negative result in this section are (1) the small size of the training and test sets, (2) the lack of diversity between the different automatically translated documents, and (3) the insensitivity of our bilingual coherence model to entities that are mentioned only strictly within one sentence. In the following section, we direct our attention to the first issue.

6.3 A Score-based SMT Evaluation Metric

The results in the previous section using our bilingual model of coherence with either pre-trained model parameters or within a cross-validation setup did not show a satisfactory performance. This is partly attributed to training on data that is from a different distribution (i.e. the corpus with artificially introduced coherence errors) or is very small (i.e. our corpus with human judgements), leading to a worse generalization or overfitting.

In this section, we therefore experiment with reusing features extracted from our bilingual coherence model without the additional learning of weights. This simpler formulation is based on a combination of scores more similar to traditional untuned evaluation metrics. By removing the requirement of tuning from our evaluation metric, it can also be more easily applied to other datasets and languages without requiring a tuning step (and therefore manually annotated data).

We first define the score-based discourse-aware SMT evaluation metric (Section 6.3.1). We then test the metric in experiments using the corpus where we introduced coherence errors artificially (Section 6.3.2). This experiment is again in a more controlled setting, where the target-side is based on human translations, thus showing if the metric generally has the potential to be used for SMT evaluation, or if it is less expressive due to the missing weight tuning. We then apply the metric also to the more realistic test set used in the previous experiments, with automatically translated target-sides and post-edited pronouns (Section 6.3.3). This experiment tests whether results on the controlled setting generalize to a more realistic SMT setting. Both experiments tell us whether tuned weights are productive or counterproductive when using the bilingual model of coherence for SMT evaluation.

group	no.	feature
1	1	AOD of merged EGM: P_U
	2	AOD of merged EGM: P_W
	3	AOD of merged EGM: P_{Acc}
2	4	number of mention insertions
	5	number of mention deletions
	6	number of mention preservations
3	7	number of aligned entities
	8	number of unaligned source entities
	9	number of unaligned target entities
4	10	number of entities in source document
	11	number of entities in target document

Table 6.16: Set of features of the fine-grained model of bilingual coherence.

6.3.1 Metric Definition

As a reminder, we reproduce the features involved in the bilingual model of coherence in Table 6.16 from Section 5.2.5. We base our definition on a combination of a subset of these features. The features consist of several groups, which represent different aspects of bilingual coherence. The first group contains three variants of the AOD (i.e. the output of the three variants of our coarse-grained bilingual coherence model). The problem with the AOD is that it is not normalized, and the values depend on the size of the involved source- and target-side entity graphs from which it was computed. We therefore do not use this group of features here and integration is left for future work.

The second group of features is based on the entities that could successfully be aligned, and considers source-target mention pairs. To make use of the information captured with the features in this group, we define the following mention-level scoring function:

$$h_m(d) = \frac{m_d^{pre}}{m_d^{pre} + m_d^{ins} + m_d^{del}}$$

where d is a document, m are entity mention pairs from successfully aligned entities, and e^{pre} stands for the count of preserved entity mention pairs, e^{ins} stands for the count of mention pairs, where the target-side mention does not have a source-side counterpart, and e^{del} stands for the count of mention pairs, where the target-side

mention is missing. This score is one, if we only have preserved mention pairs, and decreases to zero if we only have inserted or deleted mention pairs. The intuition behind this is that a more coherent target-side document exhibits a larger amount of entity mentions that have corresponding mentions on the source-side (i.e. a high m_d^{pre} count). If the target-side is not a truthful translation of the source-document (i.e. a less coherent document), the other two counts will be higher, resulting in a lower metric score.

The third group of features looks at aligned entities and the number of unaligned source or target entities. We define the following two variants a and b as entity-level scoring function:

$$h_e^a(d) = \frac{e_d^{al}}{min(e_d^s, e_d^t)}$$

$$h_e^b(d) = \frac{e_d^{al}}{min(e_d^{us}, e_d^{ut})}$$

where e are entities and e^{us} stands for the number of unaligned source entities, e^{ut} for the number of unaligned target entities and e^{al} for the number of aligned entities. Furthermore, e^s stands for the number of source entities and e^t for the number of target entities. Variant a is one, if all possible source and target entities are aligned, and zero if none of the source and target entities are aligned. Variant b provides the ratio between aligned entities and those entities that could not be aligned. The main difference between the two denominators is that the quantities in the former contain counts for singleton entities and entities that are mentioned more than once within one sentence. The quantities in the latter are derived only from counts of entities that are mentioned more than once across sentences. Furthermore, the former has a slightly better chance at capturing sentence-internal entities, since they are reflected in the count, and therefore the expectation is that it has fewer (1,1) predictions. The general intuition behind this component is that a coherent target-side document should contain a higher number of entities that have a matching counterpart in the source-side.

Finally, we combine the mention- and entity-level scoring functions as follows:

$$h(d) = \frac{1}{2} \cdot (h_m(d) + h_e(d))$$

where $h_e(d)$ is either variant a or b. This gives us two variants of the final definition of our score-based SMT evaluation metric, variant h_m, h_e^a, and h_m, h_e^b.

type of confusion	accuracy	
	h_m, h_e^a	h_m, h_e^b
UedinDevDiag	75.33	74.67
TurkunlpTestDiag	70.00	64.67

Table 6.17: Results of different metric definitions in accuracy in percent, as tested on 150 document pairs from experiments with the artificial corpora.

6.3.2 Experiments on Artificial Corpora

We experiment with the above two definitions of the score-based metric. As test corpus, we consider the setup we had before with our corpora where we artificially introduced coherence errors via confusion matrices from CLPP systems as described in Section 5.3.3. In this experiment we use the same two test corpora, one based on pronoun confusions from our *UedinDevDiag* confusion matrix, and the other based on confusions from the *TurkunlpDiag* confusion matrix. Each of the test corpora contains 150 document pairs from the same single train/test split as in the artificial corpus experiments (albeit here we ignore the training data, since we do not need it).

Results are shown in Table 6.17 and performance is given in accuracy in percent. Both metric definitions perform well (between 64.67% and 75.33% accuracy), especially considering that the metric is untuned. One can observe again that the performance is generally lower on the *TurkunlpTestDiag* data set, since it contains fewer pronoun confusions and is therefore closer to the original document of each document pair. It is harder to make accurate predictions if the documents in a pair are more similar. We made the same observation in earlier experiments with the tuned metric (cf. Section 5.3).

We also show the predictions of each metric definition on both artificial corpora in Table 6.18. In these corpora we do not have any tied gold rankings (nor (1,2) rankings), which is why we do not show full confusion matrices. The metric only predicts very few ties (1,1).

6.3.3 Experiments on the Realistic Corpus

In this experiment we test our score-based evaluation metric on SMT system outputs as done before with the tuned SMT evaluation metric (cf. Section 6.2). Results for the *hu-*

type of confusion	metric definition	2,1	1,2	1,1
UedinDevDiag	h_m, h_e^a	113	35	2
UedinDevDiag	h_m, h_e^b	112	35	3
TurkunlpDiag	h_m, h_e^a	105	42	3
TurkunlpDiag	h_m, h_e^b	97	48	5

Table 6.18: Counts of ranking predictions made by different definitions of the score-based SMT evaluation metric on the two differently confused artificial corpora. The gold predictions are all (2,1).

man vs. system setup are given in Table 6.19. On the full test data set, the performance decreases by between 0.61% and 11.47% (absolute) compared to the results in the 10-fold cross-validation (cf. Table 6.8a). On the reduced data set without input data ties performance decreases for two *human vs. system* combinations by between 1.13% and 20.95% (absolute), and increases for the two other combinations by between 1.12% and 3.65% (absolute), when comparing to the corresponding Table 6.8b.

The counts of different predictions for metric definition h_m, h_e^b are shown in Table 6.20. These counts show promising results in that consistently over all *human vs. system* combinations, there is a reduced number of wrong strict rankings compared to the cross-validation setup.

Results for the *system A vs. system B* setup are given in Table 6.21. On the full test data set, the results are comparable with the trained metric (cf. Table 6.11a). On the reduced data set without input data ties, however, the score-based metric performs better, with the h_m, h_e^b metric definition having an increase between 4.97 and 10.17 percentage points in accuracy (cf. Table 6.11b). On both full and reduced data sets, the h_m, h_e^b metric definition performs better than the other one. This is the opposite behaviour compared to results on the artificial corpora.

The confusion matrices for metric definition h_m, h_e^b are shown in Table 6.22. We compare them to the ones from the tuned evaluation metric in Table 6.12. Two major observations can be made. First, the score-based metric predicts an even higher number of equal ranks (1,1). It has the same issues with a lack of representation for sentence-internal entities as the tuned metric, since it is based on the same bilingual coherence representation (i.e. this is not a problem we tried to solve with the score-based metric). More importantly, it makes fewer wrong (1,2) predictions, especially in those cases

	accuracy	
	h_m, h_e^a	h_m, h_e^b
HUMANC2 vs. NMTBASELINEPLAIN	61.94±12.51	60.65±13.66
HUMANC2 vs. NMTCLPPPLAIN	58.06±8.08	58.06±8.57
HUMANC2 vs. NMTCLPPNONEFEAT	58.71±10.57	60.00±12.39
HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT	58.06±7.56	59.35±10.81

(a) Results on full test set (i.e. with input data ties).

	accuracy	
	h_m, h_e^a	h_m, h_e^b
HUMANC2 vs. NMTBASELINEPLAIN	22.37±9.21	19.74±12.52
HUMANC2 vs. NMTCLPPPLAIN	27.78±13.38	27.78±17.39
HUMANC2 vs. NMTCLPPNONEFEAT	24.71±16.05	27.06±19.92
HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT	26.14±16.39	28.41±22.03

(b) Results on reduced test set (i.e. without input data ties).

Table 6.19: Results of the *human vs. system* experiment setup for the score-based SMT evaluation metric. Performance is given in terms of accuracy in percent, which in turn is averaged over the same 10 folds from experiment I (Section 6.2.3). Standard deviation is given after ±.

where they should have been ranked as (2,1), and generally also predicts more (1,2) predictions correctly. The confusion matrices (not shown) for the metric definition h_m, h_e^a show a similar picture, although not as clearly.

6.3.4 Discussion

Experiments with the score-based SMT evaluation metric showed promising results in general being comparable or better than the tuned metric in the *system A vs. system B* setup. This is encouraging, since it points towards easier reuse of the metric, if it does not require to be tuned, while still performing well. Nevertheless, the results have to be interpreted relative to their overall low performance. Experiments across both setups showed that the score-based metric performed better at predicting cases, where the weaker post-editing system outputs were preferred by human judges (i.e. (1,2)

	2,1	1,2	1,1
HUMANC2 vs. NMTBASELINEPLAIN	15	11	50
HUMANC2 vs. NMTCLPPPLAIN	25	15	50
HUMANC2 vs. NMTCLPPNONEFEAT	23	12	50
HUMANC2 vs. NMTCLPPNONEFEAT&PREDICT	25	11	52

Table 6.20: Counts of ranking predictions made by the h_m, h_e^b metric definition of the score-based SMT evaluation metric for the *human vs. system* experimental setup on the reduced data set (i.e. without input data ties). The gold predictions are all (2,1).

	accuracy	
	h_m, h_e^a	h_m, h_e^b
NMTBASELINEPLAIN vs. NMTCLPPPLAIN	61.94±10.67	64.51±9.54
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT	63.87±12.18	67.74±12.53
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT	60.64±8.74	64.52±10.35

(a) Results on full test set (i.e. with input data ties).

	accuracy	
	h_m, h_e^a	h_m, h_e^b
NMTBASELINEPLAIN vs. NMTCLPPPLAIN	29.76±20.00	34.52±19.42
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT	30.00±23.18	37.50±26.81
NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT	35.11±19.53	41.48±23.86

(b) Results on reduced test set (i.e. without input data ties).

Table 6.21: Results of the *system A vs. system B* experiment setup for the score-based SMT evaluation metric. Performance is given in terms of accuracy in percent, which in turn is averaged over the same 10 folds from experiment II (Section 6.2.3). Standard deviation is given after ±.

	2,1	1,2	1,1	Total
2,1	15	7	34	56
1,2	1	4	5	10
1,1	5	3	81/10	89/18
Total	21	14	120/49	155/84

(a) NMTBASELINEPLAIN vs. NMTCLPPPLAIN

	2,1	1,2	1,1	Total
2,1	14	4	31	49
1,2	1	4	7	12
1,1	5	2	87/12	94/19
Total	20	10	125/50	155/80

(b) NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT

	2,1	1,2	1,1	Total
2,1	18	6	29	53
1,2	2	4	10	16
1,1	4	4	78/17	86/25
Total	24	14	117/56	155/94

(c) NMTBASELINEPLAIN vs. NMTCLPPNONEFEAT&PREDICT

Table 6.22: Confusion matrices of the h_m, h_e^b metric definition for the *system A vs. system B* experimental setup. Each column represents the predicted ranking order, and each row represents the gold ranking order. The numbers to the left of the slash are from the experiments *with input data ties*, the ones to the right are from the experiments *without input data ties*.

rankings).

6.4 General Discussion

The accuracy scores in all of the above metrics are to be understood as to how well the rankings produced by the evaluation metric correlate with human judgements. A high accuracy score means that the evaluation metric ranked a high number of system outputs according to how humans ranked these system outputs. At first sight, the accuracy results in both experimental setups (i.e. *human vs. system* and *system A vs. system B*) seem to be quite good on the full test data set (i.e. on average 64%). However, a closer inspection revealed that this result is misleading, since the test data set contains a number of document pairs, where the SMT systems produce non-distinguishable output (i.e. input data ties). This is either due to the fact that there are no pronoun instances in a particular parallel document, and thus our post-editing systems cannot make any changes to the baseline output. Or it is due to the fact that our CLPP system variants predicted exactly the same pronouns as the baseline. However, these data points in our experiments are trivially easy to rank as having an equal rank (1,1). With input data ties the evaluation metric operates on document pairs with identical target sides from two system outputs, hence producing the same metric score. We therefore also included experiments, where we filtered out these input data ties from the corpus, with the attempt to reveal the more realistic performance of the evaluation metric. Performance on these reduced data sets is considerably lower than on the full data sets (i.e. on average 31%). However, this might be due to the fact that the resulting data sets are up to 50% smaller than the full data sets. This is partially confirmed by the high variance of accuracy scores in the 10-fold cross-validation experiments. Due to the small fold-size, the ranker overfits the parameters to the training fold, thus making more errors on the test fold.

We reused the annotated corpus from Chapter 4 as a proxy for human coherence rankings. By counting the number of times a pronoun from a system output agreed with the human pronoun annotation, we converted these human annotations into human judgements of coherence. These coherence judgements were then directly converted into gold rankings between two system outputs (or also between human output taken directly from these annotation and a system output in the *human vs. system* setup) according to these counts. This has the underlying assumption that entity-based coherence of a document can be captured by just looking at how pronouns are translated.

This is a simplifying assumption, since full noun phrases are also part of coreference chains, and thus contribute to the entity-based coherence of a document. Furthermore, our annotations only consider a small subset of pronouns in a document (i.e. the ones that are also used in the CLPP shared tasks) thus only capturing a part of entity-based coherence. However, our annotations do not just evaluate pronoun translation. The annotations were obtained by showing the entire parallel documents, so they are sensitive to the surrounding context, e.g. the automatically translated noun antecedent.

To test for correlation with human judgements of our discourse-aware SMT evaluation metric, there is the one-time overhead of obtaining these human annotations. Once a high correlation has been shown, the evaluation metric can be applied to new data sets without testing for such a correlation under the assumption that human judgements are consistent across different domains and genres. Therefore, the human annotations have to be collected only once. However, since our evaluation metric requires model parameters to determine, it also requires human judgements for training these parameters. In the experiments on the artificially confused corpora (cf. Chapter 5) it has been shown that the learnt weights generalize well from one confused corpus to another. This remains to be shown for the experiments on realistic data from this chapter.

Earlier work on testing SMT evaluation metrics often relied on existing SMT system output and human ratings obtained from sample translations (sentence by sentence). With our approach of comparing a discourse-aware system with a baseline system we can test whether the handled discourse phenomenon is actually beneficial for the translation, instead of simply testing discourse-related translations that are correct by chance (and not due to explicit modelling). In the DiscoMT15 shared translation task (Hardmeier et al., 2015), there are a few SMT system outputs available from systems that explicitly attempt to model discourse phenomena. However, the data sets used there are between English-French, so they are not applicable to our experiments. Nonetheless, they could be considered for future work when testing our SMT evaluation metric on this language pair. Furthermore, they provide more varied system outputs, rather than our post-editing systems that focus their changes on pronouns only.

Our evaluation metric predicts a high number of equal ranks (1,1) when the document pairs should have been ranked as (2,1) or (1,2). This means it has troubles distinguishing these document pairs from each other. Analysis showed that in those document pairs with predicted equal ranks (1,1), there are on average fewer cross-sentential coreference chains. This explains why our metric tends to perform worse on

these document pairs, since it only captures cross-sentential chains well. This prompts for integrating these sentence-internal coreference chains into the EGM representations as well. This is left for future work. A simpler, but also more error-prone approach would be to combine our evaluation metric with BLEU, following the assumption that if coreference chain mentions occur within the 4-gram window BLEU covers, differences in coherence could be captured. However, this comes with the same problems noted as deficiencies that BLEU has when it comes to evaluating pronoun translation (cf. Section 2.2). For example, it requires a matching of pronouns in reference and system output regardless of how the antecedent is translated.

Previous work on discourse-aware SMT relied on pre-existing system output. These systems did not attempt to model discourse by design. A better translation of pronouns for example would happen by pure chance. Testing the evaluation metric on such output only tests if the metric can capture these changes that are done by chance. We on the other hand want to test our evaluation metric on system outputs from systems that specifically model a discourse phenomenon. We therefore created the setup of a state-of-the-art NMT system as baseline output, and a set of post-editing systems that operated on this output. These post-editing systems consist of our CLPP systems applied to that output. They make more informed decisions about how pronouns should be translated, taking the entire document context into account. Furthermore, it should be noted that in this setup we do not have the requirement that the systems that model a discourse-phenomenon must also produce better translations than the baseline. This would be desirable, but for testing our discourse-aware evaluation metric, this not crucial.

Our evaluation metric does not take reference translations into account. This is mainly motivated by our initial hypothesis that there is a strong relation between entities in the source document with those in the target document. And that in incoherent translations there is a systematic difference for these source-target entity pairs with respect to a coherent translation that we can exploit with our evaluation metric. The general approach of relying on the source-side for SMT evaluation is also taken by Guzmán et al. (2014) and Joty et al. (2014) in their SMT evaluation metric which compares the discourse structure between a source sentence and target sentence.

Our evaluation metric is unlexicalized, i.e. it does not represent the actual words that are used in the source- or target-side document. This is in line with the original monolingual formulation of the EGM (Barzilay and Lapata, 2008). This unlexicalized representation is one way of circumventing the problem mentioned earlier that BLEU

has with expected pronoun translations in the reference that are not checked against the system translation of the antecedent. In other words, if a pronoun is translated differently to the reference translation, but agrees with the antecedent in the translation, then this would still create a coreference chain, which is the basis for our evaluation metric. Like this an actually correct pronoun translation that appears superficially wrong with respect to the reference contributes positively to the metric. At the same time though this means that our evaluation metric is not designed to be used on its own, since it does not make any judgements about actual translations of individual words. It will therefore have to be combined with evaluation metrics that do take words into account, e.g. BLEU or METEOR. In our experiments this was not a problem, since the system outputs were only distinguished by words (i.e. pronouns) that influence the structure of the translation and are therefore captured by our evaluation metric. Words other than pronouns were not changed from the baseline in the post-editing systems.

In the WMT metrics shared task (Bojar et al., 2016) SMT evaluation metrics are tested for correlation with human judgements on two levels. The first level, i.e. the *system level*, considers the final SMT system rankings produced by the human assessment of SMT systems that participated in the WMT news translation shared task. A metric is then tested how well it correlates with these system rankings. Since we only have four systems which are very similar to each other, we do not test for such a system-level correlation. Typically there is a much larger set of systems compared in this setup (i.e. 15 systems in WMT16), but even then many of these systems do not show a distinction that is statistically significant resulting in the same ranking cluster (i.e. producing 7 clusters of systems in the final ranking in WMT16). The second level, i.e. *segment level*, is what is most similar to our setup. The major difference is that in the WMT metrics shared task, a segment is a sentence, whereas in our metric, a segment is a document. Results in Bojar et al. (2016) show that correlation with human judgements on the sentence-level correlation test is generally much lower compared to the system-level correlation test. With this context in mind, the fairly low accuracy scores (i.e. low correlation with human judgements) of 31% on average on the reduced data sets are relativized to a certain degree, since it is a hard task in general.

The experiments with the score-based SMT metric showed promising results in that it could achieve an increase of 5-10% (absolute) in accuracy on the reduced data sets. This is promising, since it means that the score-based metric has the potential to be used directly as an SMT metric without training model parameters. It further means that it has the potential to be more straight-forwardly applied to other data sets

and possibly new language pairs without the requirement of human annotations for training parameters. The fact that the score-based evaluation metric performs better than the tuned SMT metric provides additional evidence that the latter one is overfitting parameters towards the small training sets. In addition to that, the metric less frequently confused the coherence of the document pairs, i.e. it less frequently ranked documents in the wrong order. This provides a good starting point for overall good performance, once the expressivity of the metric with sentence-internal coreference chains is improved to handle the many (1,1) predictions.

A further issue is the frequency of changes with respect to pronoun translations that occur within each document pair, often leading to little or no change at all. If we have longer documents (e.g. TED talks as used in the ParCor corpus or in the CLPP shared task test sets) or a higher frequency of pronouns, we expect the differences with respect to entity-based coherence to increase. This in turn will increase the chance that our SMT metric will provide more distinct predictions for document pairs, hence reducing the number of (1,1) predictions. This would be a complementary experiment to increasing the granularity of the underlying bilingual model of entity-based coherence.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The main focus of this thesis has been two-fold: (1) modelling pronoun translation as a phenomenon of entity-based coherence; and (2) discourse-aware SMT evaluation with respect to entity-based coherence. In the first part, we looked at modelling pronoun translation to better understand the requirements of how this should be modelled such that it provides pronoun translations with high accuracy in a realistic setting. In the second part, we focused on evaluation of SMT with a discourse-aware metric, which we considered a necessity to enable advances in discourse-aware and document-level SMT. The recurrent theme throughout our experiments is taking advantage of dividing the main problem of pronoun translation and entity-based discourse-aware SMT evaluation into smaller steps, which enabled us to more deeply investigate the problems at hand. These subdivisions abstract away from the full SMT pipeline, thus introducing artificial settings. In both modelling and evaluation parts of this thesis we worked on ways of lifting these artificial restrictions moving towards verifying and integrating the approaches back into the final goal of end-to-end SMT with discourse-aware evaluation of these systems. Bringing back the issues into the more realistic setting exhibited weaknesses and drawbacks of proposed approaches and setups. There is still a big gap to fill between the current understanding of modelling pronoun translation and a fully discourse-aware evaluation metric. The results in this thesis should provide good starting points to move the field further.

In Chapter 3 we created CLPP systems for English-German and English-French that produce state-of-the-art results on the test sets of the CLPP shared tasks from 2015 and 2016. With the help of these systems we could show in a feature ablation

study that having access to the antecedent of a pronoun and the grammatical features (gender and number) of this antecedent is beneficial for increasing the prediction performance. Similarly, predicting NONE is not only beneficial for higher performance, but also reduces the amount of OTHER predictions. This makes the predictions more useful in downstream tasks, since NONE has a direct mapping to a surface form (i.e. the empty string), whereas OTHER is an artificial string that cannot be substituted directly into natural text. Experiments on both language pairs showed that a similar setup can be used. This showed that the same approach generalizes well to other languages that have similar agreement constraints on pronouns and their antecedents. Furthermore, we could confirm linguistic knowledge for the pronouns and language pairs under discussion by showing the usefulness of having access to the antecedent and its grammatical features in particular and by providing evidence that target-side features are more informative than source-side ones in general.

We then pointed out in Chapter 4 that the two CLPP shared tasks work within an artificially restricted setting by working on target-side data that has been translated from the source by human translators (and lemmatized in the CLPP16 shared task). This deliberate restriction was good for understanding what is involved in pronoun translation without the noise added by full SMT systems. Furthermore, this restriction enabled the automatic creation of large quantities of training data for applying machine learning techniques, since the gold pronoun labels could be directly extracted via word alignments from the parallel documents. We worked towards removing some of the restrictions aiming at a more realistic setting by creating a test corpus for English-German with automatic translations from a state-of-the-art NMT system rather than human translations. In this setting, labels (i.e. target-side pronouns) can no longer be extracted automatically. We therefore devise an annotation scheme with guidelines and an annotation tool, with which we collect gold labels within the context of the full parallel document with the automatically translated target-side. Inter-annotator agreement showed that this annotation is a feasible task producing reliable pronoun annotations. The annotation also showed that the set of pronouns from the CLPP shared task only show little unresolvable ambiguity for humans. We then applied our trained CLPP systems on this more realistic test corpus. Results did not show a clear tendency and depending on both the system and evaluation metric a decrease or increase of performance could be observed. More importantly, however, experiments showed that the state-of-the-art NMT system had a better performance in pronoun prediction than the specialized models of our CLPP systems. This corpus and the experiments therefore

made it possible to exhibit this weakness in the CLPP system that were undiscovered with the abstracted setting of the first two shared task instances. At the same time it enables future work on more realistic CLPP modelling, such as considering pronoun translations of an underlying baseline SMT system.

The lack of appropriate evaluation metrics for SMT to handle discourse phenomena provided us with the motivation to investigate in Chapter 5 how monolingual models of entity-based coherence can be extended to the bilingual case. Focussing on entity-based coherence, we worked with the hypothesis that an entity in the source-side shows direct correspondences with entities on the target-side, and that an incoherent translation shows systematic differences in these correspondences compared to a coherent counterpart. We showed that there is such a correspondence via correlation of source- and target-side coherence model scores obtained from the monolingual EGMs. We then devised an automatic entity alignment procedure. The aligned entities are then basis for a bilingual model of coherence inspired by the monolingual EGM. The model comes in two formulations, a coarse-grained one inspired by the EGraph representation of the EGM (i.e. by merging the source- and target-side EGraphs) and a fine-grained one (i.e. by capturing cross-lingual entity transition patterns inspired by the EGrid representation). The models were tested on corpora where in the human-authored target side the coherence of the document was disturbed by artificially confusing pronouns based on typical errors CLPP systems make. This method provided us with automatic gold labels for entity-based coherence. In a learning-to-rank framework, the models were tested, such that more coherent documents should be ranked higher than less coherent ones. The coarse-grained model did not perform well, as it only provides one score for the entire parallel document. The fine-grained model performed much better (by 25 to 44 percentage points) being able to accurately rank document pairs according to their gold ranking coherence. The bilingual model of coherence therefore was shown to be useful in judging the bilingual coherence of parallel documents, with the restriction that the target-side documents contained human translations with artificially confused pronouns.

Finally, in Chapter 6 we tested how well the above bilingual model of entity-based coherence is suited as a discourse-aware SMT evaluation metric, i.e. how well it performs on non-artificial data that comes from actual SMT system outputs. For this purpose, several outputs of SMT systems are necessary. Furthermore, we can no longer obtain gold rankings of coherence automatically. We therefore need human judgements of coherence. The major goal was to find out if the ranking order of document pairs

translated by different SMT systems as established with our model correlates with the ranking by humans. Instead of relying on existing SMT system outputs, which is a common approach in SMT metric evaluation, we tested against output from a state-of-the-art NMT system, and systems which explicitly handle a discourse phenomenon, i.e. pronoun translation. To obtain the latter, we used our CLPP systems as post-editing components operating on the NMT output. For obtaining the human judgements of coherence, we reused our annotated corpus from Chapter 4 by converting the pronoun annotations with respect to SMT system outputs into ranking labels. This was done by comparing the translated pronouns of a pair of system outputs with the pronouns as translated by humans and counting the matching translations. The system in each pairwise system comparison that yielded a higher number of matching pronouns was assigned the higher rank of coherence. In experiments we discovered that with our setup on the news data set, almost half of the document pairs are indistinguishable from each other, since there are either no pronoun instances, or the two involved systems made the same prediction. Furthermore, we realized that in this setup, the evaluation metric predicts a lot of ranking ties. We found out that this happens more frequently if in the involved parallel document pairs there are fewer cross-sentential coreference chains and fewer coreference chains in general. Since our evaluation metric is predominantly sensitive to cross-sentential coreference chains, this means that in those parallel documents there is less information that can be extracted with our metric. This is a problem that future work on entity-based models of coherence for discourse-aware SMT evaluation has to focus on.

One of the recurring approaches in this thesis is starting to investigate a problem from an abstraction of the final goal, followed by exploring ways to remove some of these abstractions. Our experimental results show that this approach comes both with advantages and disadvantages. On the one hand, it enables understanding of pronoun translation and testing our CLPP systems in a controlled setting and with automatic means. This helps identifying relevant features (e.g. antecedent of a pronoun, modelling prediction of the empty word) for modelling cross-lingual pronoun behaviour. On the other hand, when applied to a more realistic test set, it was shown that a full NMT system performed better at predicting the right pronoun than the dedicated CLPP systems. Similarly, experiments with our bilingual model of entity-based coherence showed very promising results on the artificially confused data set, providing us with evidence that the model works in principle. However, when applied as discourse-aware SMT metric on a realistic test set it pointed towards limitations of the abstraction ap-

proach, since it performed much worse on the realistic test set. Despite the drawbacks, we still believe that these abstracted settings are valuable approaches, since they verify the models at least on a restricted setting and more importantly allow for more extensive and faster experimentation than with manual annotation, since gold labels can be obtained automatically. However, it remains important to keep the original goal in mind, so that these abstractions do not make the problem at hand too simple or unrealistic, such that once applied to the original setting the previously working models then fail.

In our thesis, we aimed at providing annotations based on a generally used and unbiased test set (i.e. WMT news test set). Complementary to this approach are the CLPP shared tasks, which explicitly compiled test sets with documents where they ensured a good coverage of rare pronouns. We opted against filtering out documents with a low frequency of pronouns, because we wanted to verify both our CLPP systems and our SMT evaluation metric on an unbiased and generally applicable data set. However, this poses additional difficulties to the already difficult problem of discourse-aware SMT in making the sparsity issue even stronger. In our approach with using CLPP systems for post-editing this resulted in many document pairs with few to no changes or differences. We conclude from this that if such sparse, but important problems are attempted to be tackled, then more targeted evaluation sets seem unavoidable. Nevertheless, any experiment with a targeted test set, needs to be complemented with experiments on unbiased test sets to be able to compare how proposed models and metrics would fare in a realistic setting.

The unexpectedly high results of NMT systems in the CLPP task compared to dedicated CLPP systems prompt for a reevaluation to what degree pronoun translation is still a problem in NMT systems. When analysing this it needs to be identified what pronouns NMT systems are good at, and which ones they still cannot handle. Pronouns with sentence-internal antecedents beyond the context of n-gram based LMs might be less problematic in NMT systems, whereas pronouns with cross-sentential antecedents could still require specialised handling.

We only considered one aspect of coherence (i.e. entity-based coherence). This makes the already sparse discourse-level problem even sparser. Rather than considering individual aspects of coherence separately one should address this problem as one discourse-aware SMT evaluation metric with different components that take different aspects into account. This makes it harder to pinpoint what element contributed to the final metric score, but might make the metric overall less sparse and more success-

ful, since it has a higher chance of revealing coherent or incoherent components in a translation.

7.2 Future Work

CLPP

In Chapter 4 we tested our CLPP systems on realistic data where the target-side has been automatically translated. In these experiments, we reused our pre-trained models that were trained on the CLPP16 shared-task data sets with a human-authored and lemmatized target-side. In this setup, potentially useful information is removed from the training data by lemmatization. Further experiments should be conducted over what kind of training data is more helpful to learn better CLPP models, whether it is training data where the target-side contains full form human reference translations, or whether it is the training data with the lemmatized target-side, or combination of the two.

The CLPP shared tasks come with a few restrictions. One of these we lifted by creating a data set with an automatically translated target-side and manually annotated target-side pronouns (cf. Chapter 4), bringing it closer to a more realistic setting. Another restriction is the set of pronouns the CLPP shared task focuses on. On the source-side only subject-position 3rd person pronouns are considered. For the tasks from English into German and French, these are *it* and *they*. This considerably restricts coverage of pronouns to which a CLPP system can be applied. Some pronouns, such as *I* might have a trivial one-to-one mapping in these language pairs, however, there are other pronouns, such as *you*, that are also ambiguous in the source language and can be translated into different pronouns on the target-side. Opening up the small set of pronouns on the target-side, will reduce the large number of OTHER class labels, therefore, making a potentially large portion of predictions more useful for downstream tasks. Rather than predicting OTHER, which we for example cannot use in our post-editing setup (cf. Chapter 6), we could then predict an actual pronoun. Difficulties with this are the potentially low frequency of pronouns currently grouped into the OTHER class. Furthermore, increasing the number of possible classes might make learning an accurate classifier harder. This might require increasing the size of the training data.

We experimented with predicting pronouns in a sequence with the hypothesis that if we jointly predict all pronouns belonging to the same coreference chain this should

be beneficial for the overall performance. Rather than individual predictions for each pronoun instance, we therefore grouped all those pronouns into one sequence that occur in the same coreference chain. In experiments with a linear CRF however we could not show that such an approach increased the prediction performance. Nevertheless we believe there is valuable information in the fact that pronoun instances belong to the same coreference chain that should be exploited. Addressing the issue of possibly diverging grammatical genders of nouns in a coreference chain, e.g. by an additional Boolean feature that records whether all nouns in a coreference chain have the same gender or not, might be worth an experiment.

Bilingual EGM

The current formulation of our bilingual model of coherence does not explicitly model entities with multiple mentions that remain strictly within one sentence. These are only indirectly captured by including them in the count of source-side and target-side entities. Excluding sentence-internal non-singleton entities is in line with the monolingual EGraph representation of the EGM, which only captures entity transitions across sentences. Our experiments however showed that our bilingual model did not adequately capture bilingual coherence in documents with a larger number of sentence-internal non-singleton mentions. This suggests that even though they do not seem as relevant in the monolingual case, they seem more important in the bilingual case. Experiments should be conduced to find out whether including these entities improves the correlation with human judgements.

All EGM formulations (monolingual and bilingual) in this thesis only record the highest ranking entity mention within a sentence if there are multiple mentions. This is a coarse-grained perspective on the document. These other syntactically lower ranking mentions also contribute to the coherence of the document, and it should be experimented whether they can be incorporated into the model as well. On the other hand, in the translation setting removing this abstraction layer might increase the number of one-to-zero or zero-to-one mappings between entity mentions in the source- and target-side document.

The fine-grained formulation of our bilingual model of coherence currently considers coherence patterns across two languages for example recording how often entity mentions in aligned source-target entity pairs are preserved, deleted and inserted. This horizontal view (across language, but within one sentence) does not consider tran-

sitions of bilingual entity mention pairs of successive sentences. This vertical view (across languages and across sentences) would take a more direct inspiration from the EGrid formulation of the EGM in that it captures the entity mention transitions across sentences. Similarly to that, we could capture how sequences of entity mention transitions within the source language correspond to such transitions in the target language.

SMT Evaluation Metric

In experiments using the NMT baseline as starting point for creating different systems via post-editing by substituting target-side pronouns with CLPP predictions, these systems are only different with respect to the target-side pronouns, i.e. all other translations are the same. Future experiments will have to look into how well the SMT evaluation metric performs on SMT outputs that are different in other aspects in order to close the gap between the current setting and a realistic scenario of ranking arbitrary SMT systems. In our experimental setup this would require more manual annotations, again asking human annotators to provide pronoun choices for a given automatically translated target side. This annotation would have to be collected for each tested SMT system.

Related work on discourse-level SMT evaluation often incorporates sentence-level metrics, such as BLEU and METEOR, in addition to a proposed discourse-aware component. In our experiments, we do not follow this approach, since BLEU and METEOR are not focussed specifically on particular phenomena, i.e. they treat each word or n-gram overlap with equal importance. They would therefore capture other changes in the SMT output, which would make it harder to pinpoint the contribution of the discourse-aware evaluation component. However, this approach makes no claim about the usefulness of the combination of evaluation metrics in general to achieve a global SMT evaluation metric that takes all aspects of translation quality into account. Therefore, once shown that a particular discourse-phenomenon is well-captured by a discourse-aware evaluation metric, the next step would be to experiment with these combinations.

The exact role and influence of the coreference resolution systems should be investigated in detail. They provide the basis for our bilingual model of coherence. If for example a coreference resolution system is particularly bad at resolving a specific pronoun leading to wrong or missed coreference chains, this could provide our evaluation metric with false data. This issue is to be distinguished from the potentially noisy data

the target-side coreference resolution system has to work with, i.e. since it consists of automatically translated text. Coreference resolution systems are not developed with noisy data in mind. However, this indirectly provides us with a coarse-grained assessment of the overall translation quality. If the translated text is of low quality, then the coreference resolution system is expected to make a higher number of errors, whereas if we have a state-of-the-art translation, it has higher quality text to work with, thus reducing the amount of potential errors caused by noisy translations. We did not investigate this relationship in our thesis, but the exact interaction should be studied. This could for example take the form of exposing our metric to parallel document pairs translated by two discourse-unaware SMT systems, one state-of-the-art system and an inferior baseline system, where the expected coherence rank would be (2,1). If the metric still can distinguish the two systems this would be evidence in favour of the above hypothesis.

On the other hand the exact performance of the coreference resolution systems can be studied by manually annotating parallel documents for coreference, where the target-side is generated by different SMT systems. This oracle experiment would reveal the true potential of our evaluation metric if the underlying coreference resolution systems were providing perfect coreference chains. This requires a substantial amount of manual annotation. As a starting point the ParCor corpus (Guillou et al., 2014) could be considered. It is a parallel corpus containing TED talks in English and French and has been annotated with partial coreference chains by linking all pronouns to their closest noun antecedent. These annotations could be completed on the source side to provide full coreference chains. Annotations on the target side would have to be performed completely from scratch, since these would be on automatic translations.

Our SMT evaluation metric only focuses on one specific aspect of coherence. We only consider entities in a text and how they are mentioned and referenced. In understanding whether coherence can serve as a way to measure SMT quality it is important to look at single concepts rather than a whole range right away. Nevertheless, findings should be combined with other approaches on measuring coherence for SMT. Guzmán et al. (2014) and Joty et al. (2014) for example consider discourse structure as one aspect contributing to the overall coherence of a translation. Extending their work from the sentence level to the document level and combining it with our complementary view of entity-based coherence metric seems like a promising direction to go. In addition to that, with respect to event-based coherence, entities are participants in events and work has been done on monolingual modelling of how events are connected with

each other in a document (e.g. via event schemata) and how entities are connected to these events (e.g. via narrative schemata or scripts). It seems like a natural extension of our hypothesis that entities are preserved in translation, in that event schemata are also preserved in translation and that it is possible to capture systematic differences if we have coherent vs. incoherent document pairs.

Bibliography

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *Proceedings of the LREC Workshop on Linguistic Coreference*, pages 563–566.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Barzilay, R. and Lapata, M. (2005). Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1).

Barzilay, R. and Lee, L. (2004). Catching the drift: Probabilistic content models, with applications to generation and summarization. In Susan Dumais, D. M. and Roukos, S., editors, *HLT-NAACL 2004: Main Proceedings*, pages 113–120, Boston, Massachusetts, USA. Association for Computational Linguistics.

Bergsma, S. and Yarowsky, D. (2011). NADA: A robust system for non-referential pronoun detection. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, pages 12–23, Faro, Portugal.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. 3:993–1022.

Bojar, O., Graham, Y., Kamran, A., and Stanojević, M. (2016). Results of the wmt16 metrics shared task. In *Proceedings of the First Conference on Machine Translation*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Bos, J. (2008). Wide-Coverage Semantic Analysis with Boxer. In Bos, J. and Delmonte, R., editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 277–286. College Publications.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Chambers, N. and Jurafsky, D. (2008). Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Chambers, N. and Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Cheung, J. C. K. and Penn, G. (2010). Entity-based local coherence modelling using topological fields. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 186–195, Uppsala, Sweden. Association for Computational Linguistics.

Chrupała, G., Dinu, G., and van Genabith, J. (2008). Learning morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Clark, K. and Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Collins, M. and Duffy, N. (2001). Convolution kernels for natural language. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 625–632. MIT Press.

Comelles, E., Gimenez, J., Marquez, L., Castellon, I., and Arranz, V. (2010). Document-level automatic mt evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 333–338, Uppsala, Sweden. Association for Computational Linguistics.

Curran, J., Clark, S., and Bos, J. (2007). Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic. Association for Computational Linguistics.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Elsner, M. and Charniak, E. (2011). Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129, Portland, Oregon, USA. Association for Computational Linguistics.

Fellbaum, C. (2006). Wordnet(s). In Brown, K., editor, *Encyclopedia of Language and Linguistics*, pages 665–670. Elsevier, Oxford, 2nd edition.

Filippova, K. and Strube, M. (2007). Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 139–142, Saarbrücken, Germany. DFKI GmbH. Document D-07-01.

Giménez, J. and Màrquez, L. (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Gimenez, J., Marquez, L., Comelles, E., Castellon, I., and Arranz, V. (2010). Document-level automatic mt evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 333–338, Uppsala, Sweden. Association for Computational Linguistics.

Gong, Z., Zhang, M., and Zhou, G. (2015). Document-level machine translation evaluation with gist consistency and text cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40. Association for Computational Linguistics.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2).

Guillou, L. (2012). Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, France. Association for Computational Linguistics.

Guillou, L. (2016). *Incorporating Pronoun Function into Statistical Machine Translation*. PhD thesis, University of Edinburgh, School of Informatics.

Guillou, L. and Hardmeier, C. (2016). Protest: A test suite for evaluating pronouns in machine translation. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Guillou, L., Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., Webber, B., and Popescu-Belis, A. (2016). Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation (WMT16)*, Berlin, Germany. Association for Computational Linguistics.

Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B. (2014). Parcor 1.0: A parallel pronoun-coreference corpus to support statistical mt. In *Proceedings of the Ninth International Conference on Language Resources and Eval-*

uation (LREC'14), Reykjavik, Iceland. European Language Resources Association (ELRA).

Guinaudeau, C. and Strube, M. (2013). Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria. Association for Computational Linguistics.

Guzmán, F., Joty, S., Màrquez, L., and Nakov, P. (2014). Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland. Association for Computational Linguistics.

Hajlaoui, N. and Popescu-Belis, A. (2013). Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, CICLing'13, pages 236–247, Berlin, Heidelberg. Springer-Verlag.

Hardmeier, C. (2014). *Discourse in Statistical Machine Translation*. PhD thesis, Uppsala University, Disciplinary Domain of Humanities and Social Sciences, Faculty of Languages, Department of Linguistics and Philology.

Hardmeier, C. and Federico, M. (2010). Modelling Pronominal Anaphora in Statistical Machine Translation. In Federico, M., Lane, I., Paul, M., and Yvon, F., editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289.

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015). Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, Lisbon, Portugal. `http://www.idiap.ch/workshop/DiscoMT/shared-task`.

Hardmeier, C., Stymne, S., Tiedemann, J., and Nivre, J. (2013a). Docent: A Document-Level Decoder for Phrase-Based Statistical Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 193–198, Sofia, Bulgaria. Association for Computational Linguistics.

Hardmeier, C., Stymne, S., Tiedemann, J., Smith, A., and Nivre, J. (2014). Anaphora models and reordering for phrase-based smt. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 122–129, Baltimore, Maryland, USA. Association for Computational Linguistics.

Hardmeier, C., Tiedemann, J., and Nivre, J. (2013b). Latent anaphora resolution for cross-lingual pronoun prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 380–391, Seattle, Washington, USA. Association for Computational Linguistics.

Heafield, K. (2011). KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, 44(4):311–338.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142.

Joachims, T. (2006). Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.

Joty, S., Guzmán, F., Màrquez, L., and Nakov, P. (2014). Discotk: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 402–408, Baltimore, Maryland, USA. Association for Computational Linguistics.

Kamp, H. (1981). A theory of truth and semantic representation. In Groenendijk, Janssen, S., editor, *Formal Methods in the Study of Language*. Mathematisch Centrum.

Kehler, A. and Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1-2):1–37.

Klenner, M. and Tuggener, D. (2011). An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility. In *Proceedings of the*

International Conference Recent Advances in Natural Language Processing 2011, pages 178–185, Hissar, Bulgaria. RANLP 2011 Organising Committee.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, Phuket, Thailand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Kuo, T.-M., Lee, C.-P., and Lin, C.-J. (2014). Large-scale kernel ranksvm. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 812–820.

Lappin, S. and Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4).

Le Nagard, R. and Koehn, P. (2010). Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden. Association for Computational Linguistics.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Louis, A. and Nenkova, A. (2012). A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1157–1168, Jeju Island, Korea. Association for Computational Linguistics.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Luong, N. Q. and Popescu-Belis, A. (2016). Improving pronoun translation by modeling coreference uncertainty. In *Proceedings of the First Conference on Machine Translation*, pages 12–20, Berlin, Germany. Association for Computational Linguistics.

Luong, Q. N., Popescu-Belis, A., Rios Gonzales, A., and Tuggener, D. (2017). Machine translation of spanish personal and possessive pronouns using anaphora probabilities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 631–636. Association for Computational Linguistics.

Luotolahti, J., Kanerva, J., and Ginter, F. (2016). Cross-lingual pronoun prediction with deep recurrent neural networks. In *Proceedings of the First Conference on Machine Translation*, pages 596–601, Berlin, Germany. Association for Computational Linguistics.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Miculicich Werlen, L. and Popescu-Belis, A. (2017). Using coreference links to improve spanish-to-english machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Misra, H., Cappe, O., and Yvon, F. (2008). Using lda to detect semantically incoherent documents. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 41–48, Manchester, England. Coling 2008 Organizing Committee.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., and Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Rios Gonzales, A. and Tuggener, D. (2017). Co-reference resolution of elided subjects and possessive pronouns in spanish-english statistical machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 657–662. Association for Computational Linguistics.

Ruiz, N. and Federico, M. (2014). Complexity of spoken versus written language for machine translation. In *Proceedings of 17th Annual conference of the European Association for Machine Translation*, pages 173–180.

Schank, R. C. and Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. L. Erlbaum, Hillsdale, NJ.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, United Kingdom.

Schwarm, S. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R. and Kunz, B. (2014). Zmorge: A german morphological lexicon extracted from wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sim Smith, K., Aziz, W., and Specia, L. (2015). A proposal for a coherence corpus in machine translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 52–58, Lisbon, Portugal. Association for Computational Linguistics.

Sim Smith, K., Aziz, W., and Specia, L. (2016). The trouble with machine translation coherence. *Baltic Journal of Modern Computing – Special Issue: Proceedings of EAMT 2016*, 4(2):178–189.

Stymne, S. (2016). Feature exploration for cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, pages 609–615, Berlin, Germany. Association for Computational Linguistics.

Tiedemann, J. (2010). Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden. Association for Computational Linguistics.

Tien Nguyen, D. and Joty, S. (2017). A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330, Vancouver, Canada. Association for Computational Linguistics.

Tuggener, D. (2016). *Incremental Coreference Resolution for German*. PhD thesis, University of Zurich.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference*, pages 45–52.

Weiner, J. (2014). Pronominal anaphora in machine translation. Master's thesis, Karlsruhe Institute of Technology.

Wetzel, D. (2016). Cross-lingual pronoun prediction for english, french and german with maximum entropy classification. In *Proceedings of the First Conference on Machine Translation*, pages 620–626, Berlin, Germany. Association for Computational Linguistics.

Wetzel, D., Lopez, A., and Webber, B. (2015). A maximum entropy classifier for cross-lingual pronoun prediction. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 115–121, Lisbon, Portugal. Association for Computational Linguistics.

Wong, B. T., Pun, C. F. K., Kit, C., and Webster, J. J. (2011). Lexical cohesion for evaluation of machine translation at document level. In *7th International Conference on Natural Language Processing and Knowledge Engineering, NLPKE 2011, Tokushima, Japan, November 27-29, 2011*, pages 238–242.

Wong, B. T. M. and Kit, C. (2012). Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Xiong, D., Ding, Y., Zhang, M., and Tan, C. L. (2013). Lexical chain based cohesion models for document-level statistical machine translation. In *Proceedings of the*

2013 Conference on Empirical Methods in Natural Language Processing, pages 1563–1573, Seattle, Washington, USA. Association for Computational Linguistics.