2015-11-01

Data Selection using Topic Adaptation for Statistical Machine Translation

Hitokazu Matsushita
Brigham Young University - Provo

Data Selection Using Topic Adaptation for Statistical Machine Translation

Hitokazu Matsushita

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Eric K. Ringger, Chair
Kevin D. Seppi
Ryan M. Farrell

Department of Computer Science

Brigham Young University

November 2015

ABSTRACT

Data Selection Using Topic Adaptation for Statistical Machine Translation

Hitokazu Matsushita
Department of Computer Science, BYU
Master of Science

Statistical machine translation (SMT) requires large quantities of bitexts (i.e., bilingual parallel corpora) as training data to yield good quality translations. While obtaining a large amount of training data is critical, the similarity between training and test data also has a significant impact on SMT performance. Many SMT studies define data similarity in terms of domain-overlap, and domains are defined to be synonymous with data sources. Consequently, the SMT community has focused on domain adaptation techniques that augment small (in-domain) datasets with large datasets from other sources (hence, out-of-domain, per the definition). However, many training datasets consist of topically diverse data, and not all data contained in a single dataset are useful for translations of a specific target task.

In this study, we propose a new perspective on data quality and topical similarity to enhance SMT performance. Using our data adaptation approach called *topic adaptation*, we select topically suitable training data corresponding to test data in order to produce better translations. We propose three topic adaptation approaches for the SMT process and investigate the effectiveness in both idealized and realistic settings using large parallel corpora. We measure performance of SMT systems trained on topically similar data and their effectiveness based on BLEU, the widely-used objective SMT performance metric. We show that topic adaptation approaches outperform baseline systems (0.3 – 3 BLEU points) when data selection parameters are carefully determined.

Keywords: topic adaptation, data selection, statistical machine translation

ACKNOWLEDGMENTS

I am deeply indebted to Dr. Ringger, my advisor and mentor. I would not have been able to pursue this research without his consistent support and guidance to this work. I would also like to thank Dr. Seppi and Dr. Farrel, my other committee members, for providing their valuable comments and feedback on this work.

I also give thanks to Steve Richardson, my supervisor in Translation Systems of The Church of Jesus Christ of Latter-Day Saints. Since I started as an intern in his team, he has helped me develop skills required as an SMT specialist and a computational linguist and provided me with various opportunities to pursue this research in these years. Especially, I thank him for having made the datasets used at the Church available for the experiment conducted in this study. I would also like to thank Ryan Lee for having helped me obtain the dataset from the TM server and pre-process them for the experiments.

I am deeply thankful for Lynne Hansen, my sponsor. I would not have been able to reach this point without her considerable support for my schooling here at BYU.

I would like to thank my family for their support and love throughout the years at BYU. Especially, I would like to extend my deepest gratitude to Megumi and our precious children, for their selfless sacrifice, patience, and love.

Table of Contents

List of Figures vii

List of Tables ix

1 Introduction 1

1.1 Data for Statistical Machine Translation 1

1.2 Critical Properties of Training Data 2

1.2.1 Quantity of Training Data 3

1.2.2 Similarity between Training and Target Data 4

1.3 Domain Adaptation 5

1.4 LDS Dataset 7

1.5 Contributions of This Thesis 8

2 Related Work 10

2.1 Overview 10

2.2 Data Weighting for Domain Adaptation 11

2.3 Data Weighting for Topic Adaptation 12

2.4 Data Selection for Domain Adaptation 13

2.5 Data Selection for Topic Adaptation 15

3 Topic Adaptation for Data Selection 16

3.1 Types of Datasets 16

3.2 Data Selection using Topic Models 17

3.3 Three Approaches to Topic Adaptation 20
3.3.1 User-Dev Approach 22
3.3.2 LDA-Dev Approach 22
3.3.3 PLTM-Dev Approach 23
3.3.4 From Data Selection to SMT Training 23
3.4 Thesis Statement 25

4 Preliminary Experiments **27**
4.1 Overview 27
4.2 Experiment 1: Optimal Segment Length 27
4.3 Experiment 2: Comparison of Multilingual and Monolingual Topic Models 29
4.4 Experiment 3: Topic Adaptation in the Idealized Scenario 31
4.4.1 SMT System Configuration 31
4.4.2 Topic Adaptation in the Simplified Data Selection Approach 32
4.4.3 Topic Counts in the Simplified Data Selection Approach 35
4.4.4 Topic Counts with Pre-Selected Test Set 36
4.4.5 Learning Curve 38
4.5 Discussion 38

5 Realistic Scenario **40**
5.1 Overview 40
5.2 Surrogate Selection 1 40
5.3 Surrogate Selection 2 45
5.4 Top-*N* Per-Document (TNPD) Features 47
5.5 Comparison with Cross-Entropy Approach 51
5.5.1 Cross-Entropy Approach 51
5.5.2 Experiment Setup 52
5.5.3 Results 53

5.6 LDS Data Experiment 55
5.6.1 Experiment Setup 56
5.6.2 Results 57
5.7 Discussion 59

6 Conclusions **60**
6.1 Contributions of This Work 60
6.2 Limitations of This Work 62
6.3 Future Work 62

A Derivation of the PLTM Complete Conditional Distribution for Gibbs Sampling **64**

References **72**

List of Figures

1.1 Learning Curve with the Europarl EN-FR Dataset 3

2.1 Taxonomy of Related Work 11

3.1 Monolingual and Bilingual Data Selection Processes 20

3.2 Three Approaches of Topic Adaptation for Data Selection 21

3.3 Graphical Models of LDA and PLTM 24

4.1 Sentence Counts and JS Divergence 28

4.2 Comparison of LDA and PLTM using English-Synthetic Bitexts 30

4.3 Simplified Topic Adaptation 32

4.4 Dendrogram of Clustering Process 33

4.5 An Example Topic Discovered by PLTM 34

4.6 SMT Performance Results 34

4.7 Topical data selected with various topic counts and BLEU scores 36

4.8 Topical data selected with various topic counts and BLEU scores 37

4.9 TU count increase and BLEU scores 39

5.1 SMT Results Evaluated on the Surrogate Sets 41

5.2 SMT Results Evaluated on the Surrogate Sets (Cont.) 42

5.3 Three BLEU Scores on the Blind Dataset with the Topic Counts of the Best Three BLEU Scores on the Surrogate Dataset 43

5.4 PLTM-Dev Modified for Topic Count Sweeping with a Surrogate Set 44

5.5 Three BLEU Scores on the Blind Dataset with the Topic Counts of the Best Three BLEU Scores on the Surrogate Dataset (Cont.) 45
5.6 SMT Results Evaluated on the surrogate Data 46
5.7 SMT Results Evaluated on the surrogate Data (Cont.) 47
5.8 Three BLEU scores on Blind Set at the Topic Counts of the Best Three BLEU Scores on the Surrogate Set 48
5.9 Weighted Feature Combination and BLEU scores (Surrogate Set) 49
5.10 Weighted Feature Combination and BLEU scores (Blind Set) 49
5.11 Topic Count Sweeping with `test2006` 54
5.12 LDS Dataset Results 58

List of Tables

1.1 Hansards-EMEA Experiment Results 4

1.2 Example English-German (EN-DE) TUs in LDS Dataset 7

5.1 Averaged BLEU Scores on the Europarl Translation Task with `test2007` 55

5.2 Summary of LDS Dataset 57

Chapter 1

Introduction

Machine translation produces translations automatically from one language into another with computers [56]. Statistical machine translation (SMT) is an approach for producing translations based on machine learning techniques [68]. SMT has made a rapid progress and become a mainstream approach to machine translation problems in the last two decades [100].

This thesis investigates a novel data selection method for the enhancement of statistical machine translation (SMT) performance. In this introduction, we provide an overview of the relationship between the characteristics of training data and SMT performance, discuss the types of datasets used for SMT, explain critical factors for training data which affect SMT performance significantly, discuss domain adaptation for SMT and identify a weakness in the existing perspective, and illustrate the problem based on characteristics of a specific dataset.

1.1 Data for Statistical Machine Translation

SMT requires a substantial amount of parallel training data to yield desired translation results. Parallel training data are called *bitexts*, collections of documents paired with their corresponding translations, and they are the main source of training instances for SMT [93]. For SMT training, bitexts need to be segmented into sentences in both source and target languages, and these segmented sentences should be aligned to form translation units (TUs), consisting of minimal corresponding sets of consecutive sentences in both source and target languages. Various approaches for bitext segmentation and alignment have been investigated in the past two decades ([12, 13, 23, 34, 75, 101],

inter alia). Furthermore, a wide variety of bitexts have recently become available for SMT training. In general, the following kinds of datasets are used as the source of training TUs:

1. Parallel corpora, such as proceedings of the European Parliament (Europarl) [55], the Canadian Parliament (Canadian Hansards) [85], and the United Nations (UN) [29], are collected from aligned bilingual documents translated by people working in multilingual organizations. These corpora are used widely for research purposes in the SMT community (e.g., [10]).

2. Translation memories (TMs) are datasets collected from human-translated texts, which have mainly been used in the translation industry to facilitate re-use of previously translated materials. Many TMs created by various organizations have recently been shared and utilized as another source of training data for SMT.[1]

3. Comparable corpora are collections of documents from different sources that presumably discuss the same topics. Unlike 1 and 2 above, this type of data is not directly usable for SMT because it consists of *semi-aligned* or *unaligned* texts. Various approaches have been investigated to extract usable TUs from web contents [82, 90] and social media [27, 65] for SMT training. Also, methods for utilizing word or phrase pairs extracted from comparable corpora to augment translation models have also been investigated ([49, 53, 58, 77, 84, 89], *inter alia*).

1.2 Critical Properties of Training Data

For successful SMT performance, collecting usable training TUs from these data sources is crucial. In this section, we discuss two critical properties of training TUs which determine the quality of SMT performance, namely the quantity of training data and the similarity between training and target data.

[1] See `https://www.tausdata.org/` as an example.

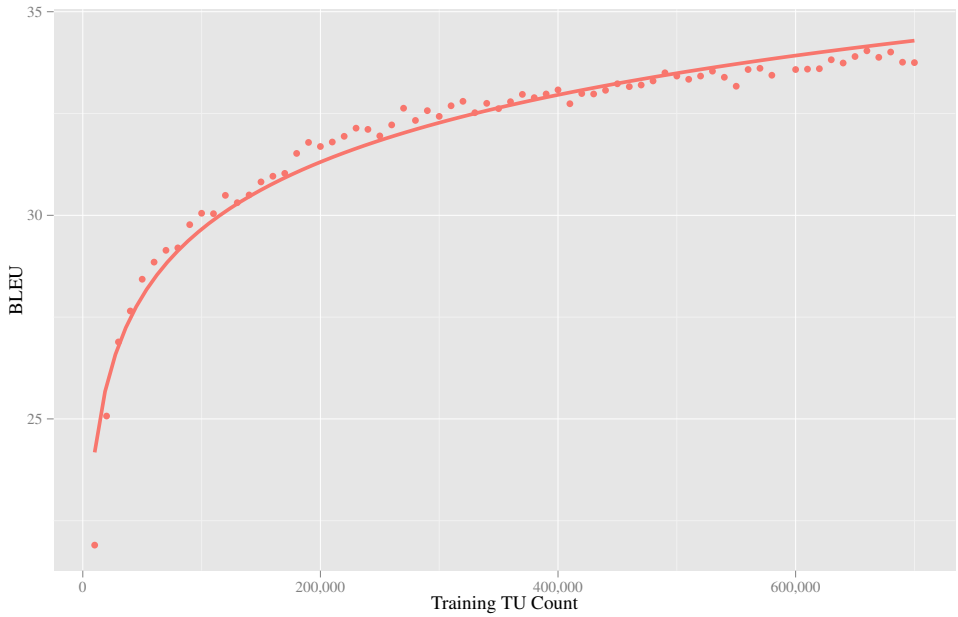


Figure 1.1: Learning Curve with the Europarl EN-FR Dataset

1.2.1 Quantity of Training Data

The first factor for desirable SMT performance is the quantity of training TUs. As with other machine learning problems, SMT performance generally improves when models are trained on large amounts of data, due to the statistical or data-driven approach employed by SMT systems. Irvine [48] describes why larger quantities of data help SMT performance:

- Fewer Out-Of-Vocabulary (OOV) words: more data provide more words and phrases for a translation model to learn, leaving fewer gaps in translation output.
- Better parameter estimation: more data provide more accurate word alignments and phrase pair extraction, which leads to better translation probability and reordering probability estimation.

Although it is important to provide a substantial amount of training TUs, data size increase is not always proportional to SMT performance. Figure 1.1 depicts the relationship between training data size and SMT performance. We randomly selected TUs from the English-French (EN-FR) portion of Europarl,[2] a widely-used parallel corpus, and trained an SMT system with the selected TUs. Then we translated a validation test set with the trained system. We increased the training TU

[2] `http://www.statmt.org/europarl/`

count from 10,000 to 700,000 in increments of 10,000 and computed BLEU [81], a common SMT performance metric, on translations of the validation set for each training data size. As shown in the figure, the linear growth of training data size does not cause linear growth in SMT performance. Based on the trend shown in the figure, it is easily anticipated that the performance improvement slows with further data. Therefore, there are quickly diminishing returns from simply increasing the amount of training TUs. Moreover, adding more data can potentially hurt SMT performance if the data are greatly dissimilar to the target task (i.e., documents to be translated), which leads to the next critical property.

1.2.2 Similarity between Training and Target Data

If training and target data are similar, then SMT is more likely to yield desirable results. To illustrate, Carpuat et al. [15] conducted an SMT experiment using two heterogeneous corpora, Hansards and EMEA [92]. Both copora are collections of EN-FR bilingual documents, but the former consists of Canadian parliament proceedings, whereas the latter is collected from medical-related bitexts. The Hansards dataset is about twenty-five times larger than EMEA in terms of the number of word tokens. In this study, the following three SMT systems were trained:

- An SMT system trained only on Hansards (System 1)
- An SMT system trained only on EMEA (System 2)
- An SMT system trained on both Hansards and EMEA (System 3)

Using a test set created from EMEA, they compared the performance of the three systems. The results are shown in Table 1.1:

System	BLEU
System 1	27.72
System 2	34.83
System 3	34.76

Table 1.1: Hansards-EMEA Experiment Results

The results clearly indicate the importance of the similarity between training and target data. There is a difference of about 7 BLEU points between System 1 and 2.[3] Moreover, the performance of System 3 is slightly lower than System 2, which indicates that the Hansards data did not contribute to the performance.[4] Furthermore, Carpuat et al. [15] mention that the Hansards data would hurt the performance more significantly if the dataset were much larger (see also [38, 66]).

In this experiment setting, the distinction between similar and dissimilar data is made simply by the datasets (EMEA and Hansards), and data in a specific dataset are essentially treated as homogeneous training instances. Based on this simple distinction, the TUs in EMEA contributed positively to the performance of System 2, whereas those in Hansards exerted no or negative impact on the performance of System 3 due to the dissimilarity to the target task. As in this experiment, categorizing training TUs based on data sources is a common approach in the SMT community, and many previous studies have examined methods to effectively utilize TUs which do not belong to target corpora based on the source-based data distinction. This problem is called *domain adaptation*, and various domain adaptation approaches for SMT have been investigated in the last decade ([4, 50, 61, 70], *inter alia*).

1.3 Domain Adaptation

In this section, we briefly discuss the main idea of domain adaptation and issues that arise in current domain adaptation approaches for SMT.

Intuitively, the concept of a textual "domain" refers to a set of shared characteristics, including vocabulary and phrases. Textual domains are related to genres, registers, sources, and topics. Data collected from the same domain should exhibit similarities. In many SMT studies, data sources are typically regarded as the primary way to specify a data domain. A domain is usually

[3] Statistical significance of performance comparison between two systems is usually attained at p-value of 0.01 when the difference is more than 1 BLEU point [54].

[4] Due to the stochastic nature of the SMT training and tuning processes, the small decrease in the BLEU scores beween System 2 and 3 in Table 1.1 can be within a margin of error, and it is difficult to conclude that the Hansards hurt the performance.

defined based on the genre of contained bitexts (e.g., legal, medical, or news [50]) or simply the name of a dataset (e.g., Hansards or EMEA). Thus, "domain" is often used as a proxy for "source".

Domain adaptation is a machine-learning problem which considers how to utilize available data in order to perform well on a given target task, potentially in a domain that is not similar to the available data [80]. The typical domain adaptation scenario discussed in the SMT literature explores how to augment a small amount of in-domain training data (i.e., data from the same source as the target set) with a large amount of out-of-domain data (i.e., data from other sources) (e.g., [61]). With regard to the Hansards-EMEA example above, it was not effective to simply use all of the (out-of-domain) Hansards data alone or to outright combine all of the Hansards and EMEA data together in the training set. What is needed for better SMT performance is to properly mitigate the differences between the EMEA and Hansards data. One approach to doing so is to augment the small in-domain EMEA training data with only the useful portions of the Hansards data.

There is a spectrum of domain adaptation approaches ranging from simply using all available training data adapting nothing at all — on the one hand — to selecting data sources that are (potentially) relevant to the source of a given test set. Adaptation involves augmenting a small amount of in-domain data with a suitable amount of out-of-domain data. For adaptation to be possible, the target domain must be known *a priori* by its source, so that suitable data can be included in the training set. It is also worth noting that most of the domain adaptation experiments reported in prior studies are designed to deal with a single "in-domain" test set. This is not always the case in real-world situations, wherein users translate documents on various subjects in a single translation process. In such situations, it is highly likely that a particular target domain is not identifiable before translation time because the dataset consists of data from various domains that can only be recognized at translation time.

In addition, the notion of making data sources synonymous with domains can be misleading and is not necessarily applicable to all cases. The following section describes one real-world situation where the notion is not applicable.

Type	TU	
Technology	EN	The database that receives and processes all survey responses may collect the Internet Protocol (IP) addresses through which responses are transmitted to us.
	DE	In der Datenbank, in der alle Umfrageergebnisse eingehen und verarbeitet werden, werden unter Umständen die IP-Adressen gespeichert, über die die Antworten an uns übermittelt werden.
Medical	EN	Doctors and local health providers who know simple infant resuscitation techniques can be the difference between joy and heartache.
	DE	Ob die Ärzte und das medizinische Fachpersonal vor Ort einfache Techniken zur Wiederbelebung von Säuglingen kennen oder nicht, kann den Ausschlag geben fÃijr Glück oder Kummer.
Facilities Management	EN	The facilities manager plans and manages the deep cleaning and maintenance procedures and performs all other tasks necessary to prepare the building for use.
	DE	Der Regionalleiter BI plant die Grundreinigungs- und Instandhaltungsverfahren, veranlasst deren Ausführung und erledigt alle weiteren Aufgaben, die vor Nutzung des Gebäudes noch anstehen.
Religious	EN	Therefore, all things whatsoever ye would that men should do to you, do ye even so to them, for this is the law and the prophets.
	DE	Alles nun, was auch immer ihr wollt, daßÃ§ euch die Menschen tun sollen, das tut ihnen auch, denn das ist das Gesetz und die Propheten.
Colloquial	EN	Hey, how are you doing, man?
	DE	Wie läuft's, Alter?

Table 1.2: Example English-German (EN-DE) TUs in LDS Dataset

1.4 LDS Dataset

TUs in a single "domain" can be very diverse in terms of vocabulary and style although they are collected from a specific data source, such as a set of bitexts created by a particular organization. According to the conventional notion of domain adaptation, the TUs in such a dataset are regarded as data generated in a specific d omain. H owever, i t i s questionable t o a lways c onsider a ll the TUs in the dataset as instances which contribute to SMT performance. An example of a dataset consisting of diverse TUs is seen in the TMs used at The Church of Jesus Christ of Latter-Day Saints (henceforth, the LDS Church). The LDS Church publishes various materials in English and translates them into more than 100 languages to support communication among its 15 million members in various countries. Along with a substantial amount of religious texts [90], the LDS

Church produces documents on a wide variety of subject domains including education, legal, medical, finance, technology and so forth. Also, the formality of language use varies according to the intended use (e.g., instruction manuals vs. casual conversations in a video clip). Some example TUs from the English-German (EN-DE) TM are shown in Table 1.2. According to the notion of domain adaptation, this dataset can be labeled as a "religious" or "LDS" domain. However, it is unreasonable to categorize this dataset with such a single domain label and to disregard the diversity exhibited in the contained TUs. In order to effectively utilize the TUs in a dataset such as this LDS dataset for SMT performance, it is highly important to identify TUs which exhibit similarities to the target task regardless of the ostensible characteristics such as arbitrary domains.

In this thesis we investigate a data selection method using *topic adaptation* for MT training. Topic adaptation has been actively investigated recently as an approach to enhance internal components of MT systems by utilizing latent features or topics found in words and phrases. Such models can be found using text mining methods called topic models (c.f. [6]). The previous topic adaptation studies are based mainly on the idea that topic models capture various thematic contexts indicated by topics and help MT systems identify suitable word and phrase translations according to the contexts [42]. Although we adopt the idea of using topic models for the improvement of MT performance discussed in the previous studies, the focus of this study is using topic models for *data selection* rather than the enhancement of internal MT components, which enables us to use our approach regardless of the chosen MT systems or translation models.

1.5 Contributions of This Thesis

The contributions of this thesis are listed below:

1. Data selection approaches that are applicable to any SMT methods
2. Topical training data selection using topic models and clustering algorithms
3. Using a surrogate set, which discussed in Chatper 5, for regirous experimentation
4. An investigation of the impact of topical training data on SMT performance

We discuss these items based on the findings of our experiments in the remaining chapters. We organize this thesis as follows. In Chapter 2, we present various domain and topic adaptation approaches investigated in the previous studies. In Chapter 3, we describe the three topic adaptation approaches examined in this thesis. In Chapter 4, we conduct several preliminary experiments and report the results. In Chapter 5, we expand our experiments in realistic settings and examine the effectiveness of the proposed approaches. In Chapter 6, we review our conclusions, contributions, and future work.

Chapter 2

Related Work

2.1 Overview

In this section, we classify, summarize, and discuss previous work on domain and topic adaptation for SMT, closely related to the proposed study. Domain adaptation has been one of the active research areas in SMT, and topic adaptation has attracted attention of the SMT community in recent years mainly because of the issue discussed in Section 1.4. Although they are closely related problems, the notions of these two problems are different. In a typical domain adaptation scenario, TUs within a training corpus or domain are not distinguished from one another and are handled in the same way. Also, their contribution to the SMT performance primarily depends on whether the target task is generated from the same domain. Based on this assumption, many domain adaptation studies focus on the approaches of adjusting SMT systems trained on a particular corpus to apply to target tasks from the other corpus [3]. On the other hand, the proponents of topic adaptation believe that TUs can be thematically diverse even within a corpus and such diversity is closely related to multiple latent topics. Thus, we separate prior studies based on these different perspectives regarding data sources.

Also, domain and topic adaptation approaches discussed in prior studies can be categorized in two groups. One focuses on weighting corpora, TUs, phrases or words extracted from training TUs according to the closeness to the in-domain data and target tasks. The other focuses on selecting TUs from a large pool of data based on similarity criteria. Therefore, we call the former "Data Weighting" and the latter "Data Selection" as in [38] and classify prior studies based on these two types of the approaches.

Figure 2.1 shows the taxonomy of prior work based on the types of problems (i.e., domain adaptation and topic adaptaion) and approaches (data weighting and data selection). The cell (4) in the figure is highlighted because it is the focus of this thesis. To the best of our knowledge, using data selection for topic adaptation has not been reported in the previous studies. In the following subsections, we summarize the prior studies categorized by (1) through (3) in the taxonomy.

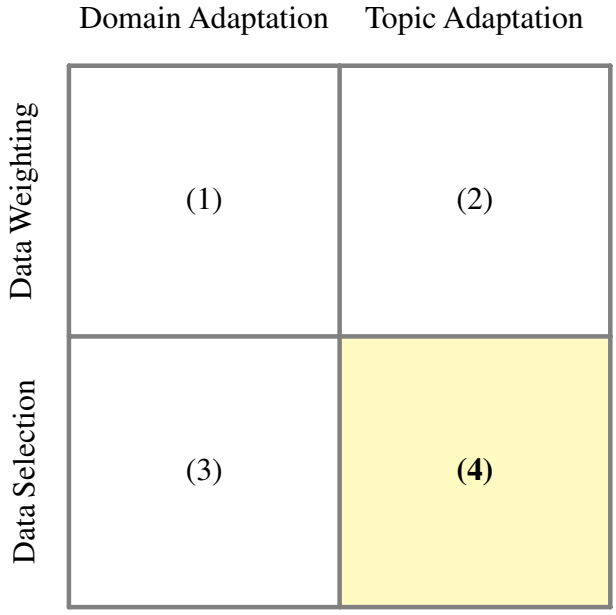


Figure 2.1: Taxonomy of Related Work

2.2 Data Weighting for Domain Adaptation

Data weighting for domain adaptation is one of the most widely-used approaches for various SMT systems. The typical phrase-based SMT system combines multiple statistical models, such as translation models, language models, reordering models, and so forth [56] to yield translations of given source texts. Using the independence assumptions of these statistical models, posterior translation probabilities are computed using log-linear combination. Researchers have trained multiple translation and/or language models with in- and out-of-domain training corpora separately and learned mixing weights in the tuning phase using discrimative learning algorithms such as minimum error rate training (MERT [79]). This domain adaptation approach is used in a wide

variety of SMT engines (e.g., [60], [97]) because of its simplicity and efficiency. Several studies examine the effectiveness of a log-linear model to combine multiple language and translation models at the corpus level [31, 61] or the phrase level [4, 38] and observe moderate gains with this discriminative combination approach. Furthermore, Matsoukas et al. [70] and Foster et al. [33] extract sentence-level features from each training TU and train a perceptron and an SVM to map features. Niehues and Waibel [78] used phrase-pair features and incorporated them into factored translation models [57]. Daumé III and Jagarlamudi [22] identified that Out-Of-Vocabulary (OOV) is the primary problem in translating documents in a divergent domain. Using a techinique called Canonical Correlation Analysis, they mine unseen words from out-of-domain corpora to integrate the identified OOV items into the translation model. Irvine et al. [51] extend their approach and examine a method to extract OOV words from comparable corpora, and Irvine et al. [50] identify that sense errors (i.e., mapping errors between known source-language words and unknown target-language words) frequently occur along with OOV errors. Carpuat et al. [16] propose a classification method to weight seen and unseen target phrases at decoding time using various textual features. Carpuat et al. [15] propose a domain adaptation method called phrase sense disambiguation (PSD) [14] to address cross-domain translation problems using classification techniques.

The primary problem with the mixing weight approach is that it downweights the entire out-of-domain dataset uniformly without considering the similarities between training and test [15]. Also, this approach works only when the provided development and blind test data are similar. Since a development test set is usually a subset of training data, it is difficult to guarantee this condition. Also, the data weighting requires modifications of existing SMT systems in order for computed weights to be incorporated into final phrase tables. Also, devising effective methods for training weights of training data can be very challenging [38].

2.3 Data Weighting for Topic Adaptation

Data weighting for topic adaptation has been actively investigated in recent years. The difference from data weighting for domain adaptation is that this type of approach dynamically weights word

and phrase pairs, which are treated as atomic units in the SMT models, based on latent topic distributions of training and test data obtained with topic modeling, rather than on the domain or data source features. For example, Xiong and Zhang [99] show that topic information discovered by topic models can be used as features for a word sense disambiguation classifier integrated in the MT decoder [99]. Also, Hasler et al. [41] used topics in order to be incorporated in log-linear models as features and optimized in the MT tuning process using MIRA [21, 40]. The other previous studies on topic adaptation focus on utilizing topic information for (re-)weighting word and phrase pairs during the SMT training and tuning processes. Eidelman et al. [28] use Latent Dirichlet Allocation (LDA; Blei et al. [9]) to identify word topics using the source side of training TUs and compute lexical weighting probabilities [20, 59] augmented by topic distributions. Hu et al. [47] extend the approach by Eidelman et al. [28] using bilingual topic models called polylingual tree-based topic models to capture topic information through not only the source side but also the target side of a training parallel corpus. They discover word-pair topics with this model and incorporate them into lexical weighting probability estimation. Hasler et al. [43] focus on topical phrase pair identification for the phrase-based translation models. In their approach, source-side phrases formed in the phrase table are treated as documents. The topic distributions over phrases identified by the model are used for computing cosine distance of training and test phrases at training time and dynamically re-weight the phrase table based on the similarity values at translation time.

As in the case of data weighting for domain adaptation, the challenge inherent to these approaches is devising methods to assign reasonable weights to each data unit. Also, some approaches such as Hasler et al. [43] can be computationally expensive because the phrases in the translation table are re-weighted for every test instance (see also [39]). Therefore, such approaches are not necessarily suitable for SMT training with multiple large-scale training datasets.

2.4 Data Selection for Domain Adaptation

Data selection for MT has also been actively investigated by several researchers. The basic approach of data selection is to identify TUs similar to in-domain training or target instances from a large

pool of TUs, typically called the mixed-domain dataset, based on similarity criteria. One of the most popular approaches used for SMT utilizes information-theoretic metrics such as similarity measures. For example, Yasuda et al. [103] use perplexities computed with language models trained on the source side of in-domain data. Mansour et al. [69] and Axelrod et al. [2] use cross entropy computed by language models trained on in-domain and mixed-domain datasets to measure the similarities between the TUs. The approach of Axelrod et al. [2] is based on a language model enhancement method used by Moore and Lewis [76], but they apply the cross-entropy approach to translation models by utilizing language models trained on both source-side and target-side TUs. As an extension of the approaches using *n*-gram language models, Duh et al. [26] and Mediani et al. [71] use neural language models [72] to deal with word contexts in the continuous vector space so that the data selection system can capture TUs in the mixed-domain sets that are similar but contain unknown words more effectively. Other popular approaches are based on information-retrieval (IR) techniques. For example, Duh et al. [25] report that they select TUs similar to the target data using a TF/IDF-based technique developed by Hildebrand et al. [45]. Mirkin and Besacier [74] incorporate the IR-based approach called Vocabulary Saturation Filter [64] to form a subset which is similar to the in-domain set but contains a wide variety of *n*-grams to avoid overfitting, which tends to occur when the in-domain set is significantly small. Kirchhoff and Bilmes [52] propose another data selection method based on an optimization approach involving submodular functions [62]. The motivation for this approach is based on the fact that training data size growth causes sublinear growth in SMT performance, which coincides with the idea of diminishing returns, one of the fundamental properties of submodular functions. They devise several feature-based submodular functions which act as surrogates for the BLEU metric [81] using *n*-grams and TF-IDF. They report that the SMT performance with the selected TUs is about the same as or exceeds that of the baseline systems.

The main criticism of the data filtering approach is that it is often difficult to obtain a reasonable amount of "good" data from out-of-domain corpora if the discrepancies between datasets are large [86]. Also, this approach can be effective only when the TUs in the in-domain dataset are

guaranteed to be homogeneous. Therefore, if the in-domain set consists of various types of TUs, such as the one described in Section 1.4, these approaches may not necessarily be effective.

2.5 Data Selection for Topic Adaptation

Our approach is an application of topic adaptation to data selection. As far as we know, topic adaptation has not been applied to data selection for MT. Although a similar idea has been investigated in a previous study [1], ours is different from it because the previous study focuses on using topic models constrained by the in- and out-of-domain distinction, which is the main concept of domain adaptation. On the other hand, because documents in a specific corpus can be dissimilar to one another, we treat documents contained in training and target datasets as topically diverse entities, and our main focus is to identify the topically similar portions in these datasets using topic models for the enhancement of MT performance.

Chapter 3

Topic Adaptation for Data Selection

Our goal is to select training TUs which are topically similar to the target task using topic models. In this chapter we describe the data selection method that we have chosen to accomplish the goal. First, we describe the types of datasets used in the SMT training and evaluation processes. Second, we explain the three approaches of topic adaptation. Lastly, we declare the thesis statement.

3.1 Types of Datasets

The three datasets used in the SMT system training and evaluation are the training set, the development set and the test set respectively. They are shown and highlighted in red ("Training Dataset(s)", "Dev", and "Test") in Figure **??**. In this figure, we depict training and development sets as a pair of two documents to indicate that they are collections of bilingual texts or bitexts. The test set is shown in the figure as a collection of documents, which indicates that the set consists of monolingual documents to be translated. We use these two types of pictures to distinguish between parallel and non-parallel datasets hereafter. The purposes of these datasets are summarized as follows:

- The training dataset consists of a large number of TUs (hundreds of thousands to millions) and is used to generate various statistical models, as indicated in Figure **??**. They are comprised of translation models, language models, reordering models, lexical weighting models, and so forth. These models are trained independently of one another, and the contribution of each model to translation quality is not known at the time of their creation. These statistical models are called sub-models and are sometimes refered to as features because they are components of the entire SMT system and used jointly to produce translations [56].

- To produce translations with the sub-models, the SMT system is tuned with the parallel data in the development test set ("dev"), which usually consist of a few thousand TUs. In this tuning process, the sub-models are combined in log-linear fashion, and the weights on sub-models are estimated based on the translation quality measured on the development test set by automatic translation quality metrics such as BLEU [56]. The weights are iteratively adjusted using parameter estimation algorithms such as MERT [79], MIRA [19], and Rampion [36], until the weight values converge.

- The test set contains documents to be translated for system evaluation. The documents are trainslated with an SMT system trained and optimized with the training and development sets, and the translation quality is measured with quality metrics in the same manner as for the tuning process. Only the source-language segments in the test set are made available to the translation system. Target language segments are available for the evaluation metrics.

As described above, training and development sets have separate impacts on the SMT performance. However, one simple but important aspect is that if both training and development sets are similar to the test set, the final SMT system is likely to yield target-like translations because such similar training data provides language features, such as vocabulary items and phrases, which are likely to be found in the test set, and a development set similar to the test set leads the log-linear model to be tuned favorably to the test set. Based on these foundations, we explain our data selection approach using topic models in the following section.

3.2 Data Selection using Topic Models

Algorithm 3.1 shows the process of data selection using monolingual and bilingual topic models. This algorithm takes three inputs: training datasets *Trn*, which are the SMT training datasets in Figure **??**, a target dataset *Trg*, and the size of a dataset selected from the training dataset based on the similarities to the target dataset, which is *S* in Algorithm 3.1. The output is the collection

Algorithm 3.1: Data Selection

input : training datasets *Trn*, target dataset *Trg*, size of selected training dataset *S*
output : selected training dataset *ST*

1 **if** *Trg is monolingual* **then**
2 remove target-language data from *Trn*
3 pre-process *Trn* and *Trg*:
4 1. create segments
5 2. remove stopwords
6 3. collect vocabulary items
7 4. transform segments to a document-term matrix for topic modeling
8 **if** *Trg is monolingual* **then**
9 train a monolingual topic model
10 **else**
11 train a bilingual topic model
12 extract a vector representation of each segment
13 **initialize**: create singleton clusters with vector representations
14 **initialize**: $ST \leftarrow \varnothing$
15 **repeat**
16 compute distances between clusters
17 merge two clusters with the minimum distance
18 $ST \leftarrow ST \cup$ training segments in clusters with segments from *Trg*
19 **until** *TU count in* $ST \geqslant S$
20 **return** *ST*

of training data which are topically similar to the target set. In this algorithm, we consider the following two scenarios based on the type of the target set:

- The target set is a monolingual dataset. In this case, this is a test set of the SMT System in Figure **??**.
- The target set is a bilingual dataset. This set is guaranteed to be similar to the target set and can be generated by expert knowledge or by a automatic process. We will discuss this type of datasets later in Section 3.3.3.

The type of the provided target set is examined in lines 1 and 2 of the algorithm. If the given target set is monolingual, then the source-language data will be extracted from *Trn* because a monolingual topic model will be chosen and the target-side data in the training set cannot be used

for the subsequent data selection processes. If the given target set is bilingual, both source- and target-side data in *Trn* will be used.

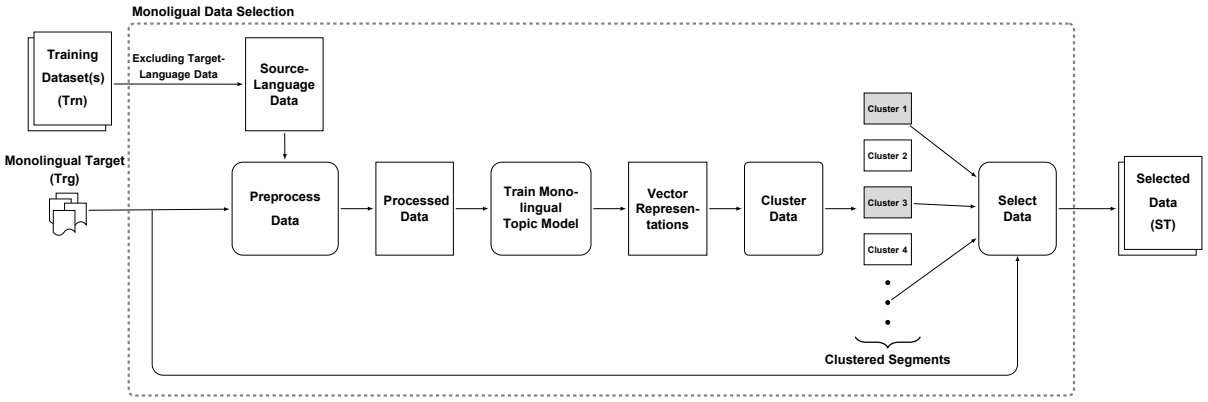
Lines 3 through 7 in the algorithm show the necessary processes to be conducted prior to the topic model training. First, the training and target sets need to be segmented to form a collection of documents because topic models need documents as its inputs. We describe our approach to generate segments for topic modeling in Section 4.2. Removing stopwords and transforming segments into a document-term matrix are typical pre-processes for topic modeling (see [5]). These processes are also conducted in this step.

Lines 8 through 11 show the process of choosing a topic model to be trained for data selection based on the type of the target set *Trg*. If *Trg* is monolingual, then a monolingual topic model will be trained on the generated segments; otherwise, a bilingual topic model will be used.

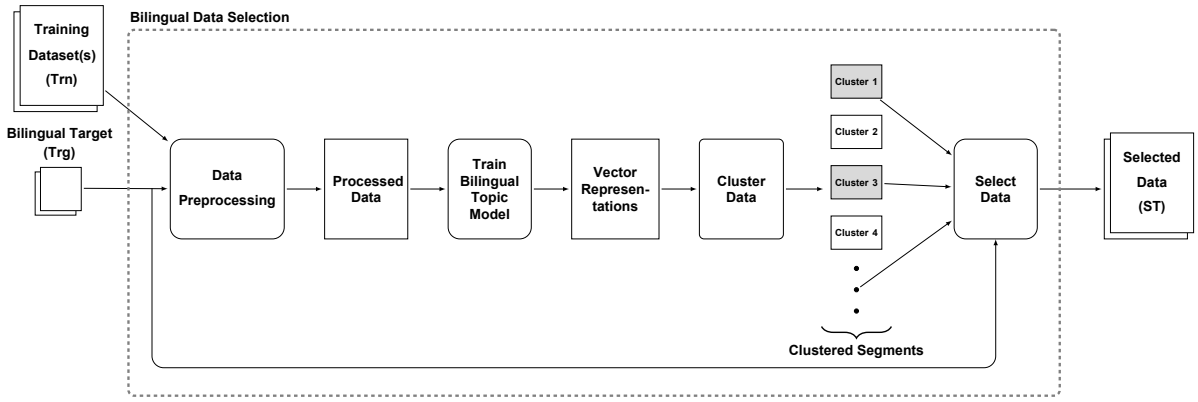
Lines 12 through 19 show the data selection process using a hierarchical agglomerate clustering (HAC) algorithm. Line 12 extracts vector representations of the segments generated from *Trn* and *Trg*. These vector representations are obtained with the chosen topic model. We explain the vector representations in Section 3.3.4 below. As the initialization processes, singleton clusters are formed with the extracted vector representations, and an empty placeholder for selected training data *ST* is created.

Lines 15 through 19 show the clustering process. In this loop, the distances between all the pairs of two clusters are computed, and the pair with minimum distance is merged to form a single cluster. Following the merge process, the algorithm goes through each of the clusters containing segments obtained from *Trg* to examine whether training segments are also stored in those clusters. If training segments are found in the clusters, then those training segments are included as elements of *ST*. Once the number of training TUs in *ST* reaches the requested data size *S*, the clustering process is terminated, and *ST* will be returned.

Figure 3.1 depicts monolingual and bilingual data selection cases dealt with in Algorithm 3.1. As shown, the main difference in these two scenarios are (1) whether the target side of training bitexts is used or not and (2) monolingual and bilingual topic models are used for data selection,



(a) Monolingual Data Selection



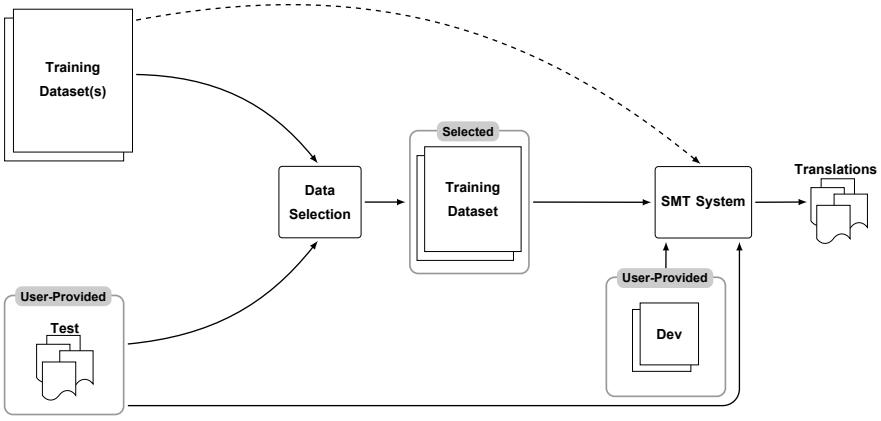
(b) Bilingual Data Selection

Figure 3.1: Monolingual and Bilingual Data Selection Processes

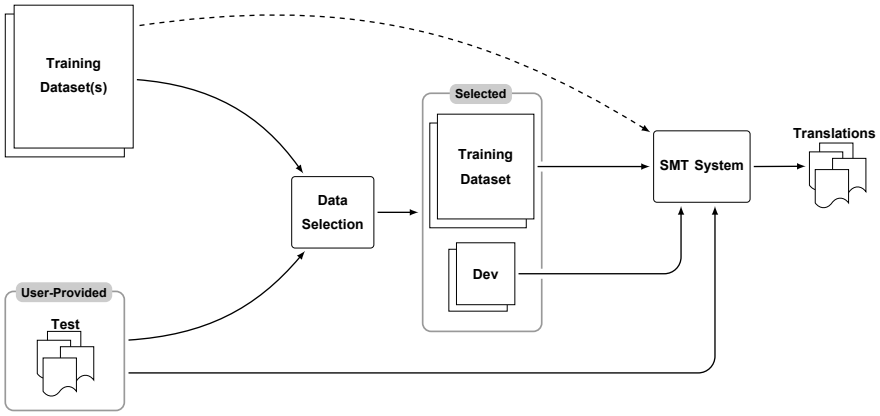
and the difference is distinguished simply by the type of target set (i.e., monolingual or bilingual). As shown in Algorithm 3.1, the choice of monolingual and bilingual data selection processes can be handled automatically. Henceforth, we use a single term "Data Selection" to indicate both cases in the subsequent sections of this chapter rather than using two separate terms for the sake of simplicity. With this data selection algorithm, we propose three data selection approaches of topic adaptation and describe them in the following section.

3.3 Three Approaches to Topic Adaptation

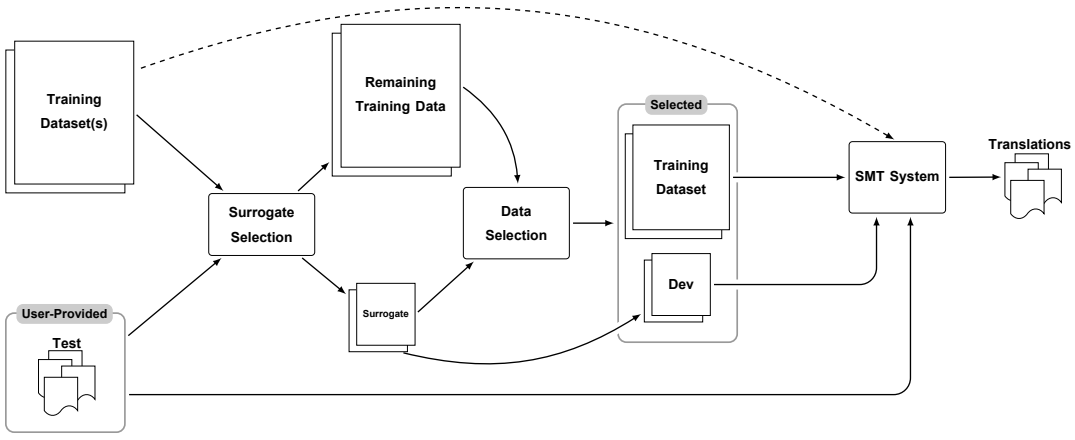
Figure 3.2 shows our three approaches to topic adaptation for data selection. The three approaches are depicted according to the types of user-provided SMT development set and according to the



(a) User-Dev



(b) LDA-Dev



(c) PLTM-Dev

Figure 3.2: Three Approaches of Topic Adaptation for Data Selection. "Training Dataset(s)", "Dev", and "Test" in this figure correspond to those in Figure **??**. "Data Selection" and "SMT System" correspond to the parts surrounded by dashed-line rectangles in Figure 3.1 and Figure **??**, respectively.

topic models used for data selection, which we described in the previous section. We call these three approaches User-Dev, LDA-Dev, and PLTM-Dev, respectively.

3.3.1 User-Dev Approach

The User-Dev is applicable when all three types of datasets are provided by the user and is illustrated in Figure 3.2a. "Training Dataset(s)" in the figure indicates all of the training TUs from parallel corpora and/or translation memories available to the user. "Dev" in the "User-Provided" box in the figure indicates the development set used in the SMT tuning process. "Test" in the "User-Provided" box in all parts of Figure 3.2 indicates a collection of documents to be translated and used to evaluate SMT performance with a chosen translation quality metric. The User-Dev approach assumes that the TUs in the development set are selected based on the user's preference or domain knowledge, and this approach uses the development set in the SMT tuning process without any modification. "Data selection" indicates the data selection process using the test documents and/or the development set, and all the available training datasets, as described in the previous section. In this approach, data selection is conducted with a monolingual topic model by using test documents as the target set in the process. It is also possible that a bilingual topic model is applied if the development set, which provides both source-language sentences and corresponding translations, assuming that the development set is a true representation of the test documents. In the following chapters, however, we do not consider using a bilingual model with this approach.

3.3.2 LDA-Dev Approach

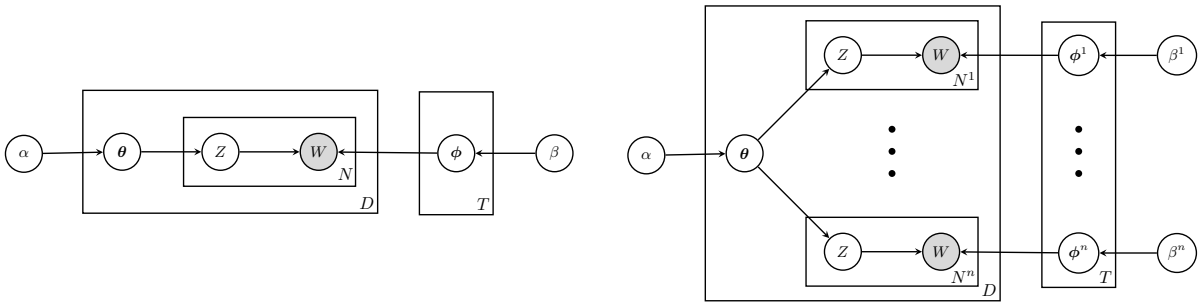
The LDA-Dev approach is used when no development set is provided by the user and a monolingual topic model such as LDA is used for data selection using source-language only test segments as the target set. As shown in Figure 3.2b, the training and test sets are involved in the data selection. When data selection is complete, both a training set and a development set are generated. Both of these sets are topically similar to the test segments and used in the SMT training process described in Figure **??**.

3.3.3 PLTM-Dev Approach

This approach is similar to the LDA-Dev approach above in that no user-provided development set is available in the beginning of the process. However, this approach utilizes a bilingual topic model by creating a "Surrogate" set in "Surrogate Selection" before the data selection process, as shown in Figure 3.2c. "Surrogate" in the figure indicates a parallel dataset created from the training datasets which is similar to the provided source-language only test documents based on the chosen similarity measure. It is called a surrogate set because it is a substitute for the actual test segments in the data selection process enabling a bilingual topic model to be used for data selection. Thus, unlike the test documents, the surrogate set contains TUs, where target-language segments are available. Therefore, the bilingual data selection described in Section 3.2 above is applicable in this approach. As illustrated in Figure 3.2c, the surrogate set is generated from the provided training datasets. After the surrogate selection process, the remaining training data is used as a training set for the subsequent data selection process. The SMT development set can be selected in the same manner as in the LDA-Dev approach above, or the surrogate set can be directly used as the topically similar development set as indicated in Figure 3.2c. In our experiments in the next two chapters, We choose the latter approach and use the surrogate set as the SMT development set.

3.3.4 From Data Selection to SMT Training

Upon the completion of data selection, "SMT System" in Figure 3.2 is trained on the selected and user-provided datasets. "SMT System" in all parts of Figure 3.2 corresponds to the portion surrounded by the dotted-line rectangle in Figure **??**. The SMT system training process is common among the three approaches, and it is described in Section 3.1. The SMT system is tuned by either a user-provided development test set or the one selected in the data selection process or the surrogate set, depending on the chosen approach described above. Then the trained SMT system yields the final translations with the given test documents in an ordinary manner. One significant aspect of these data selection approaches is that the user can choose any SMT system and translation model approach (e.g., phrase-based and syntax-based) in this process because the data selection is



(a) LDA (b) PLTM

Figure 3.3: Graphical Models of LDA and PLTM. Nodes denote random variables; edges denote dependence between random variables. Shaded nodes denote observed random variables; unshaded nodes denote hidden random variables. The rectangular boxes are "plate notation" which denote replication [7].

completely independent of this SMT system training. Also, the TUs in the training datasets which are not selected in the data selection process can be utilized in this process using existing domain or topic adaptation approaches, which is indicated by the dotted line in these figures. We do not consider this option in this study and leave it as a possible direction for future work.

The topic models we chose for this study are LDA as a monolingual topic model and the polylingual topic model (PLTM; [73, 95]) as a biligual topic model. For the surrogate set creation, only LDA is used because the source-language only test documents are involved in this process. The graphical models of LDA and PLTM are shown in 3.3. Each model represents the interaction of between observed words (shaded nodes in the graphs) in the provided documents and latent topics in the probabilistic generative process [7]. For LDA, it is assumed that a word is drawn from the multinomial distribution parameterized by $\phi_{z_{d,n}}$, where ϕ_t denotes the distributions of words over topics drawn from the Dirichlet distribution Dir(β), and $z_{d,n}$ denotes a topic drawn from the distributions over the documents ($\boldsymbol{\theta}$ in the figure), which are drawn from the Dirichlet distribution Dir(α). PLTM is the multilingual version of LDA. The main difference from LDA is that the latent topics of the words in the document pairs are drawn from the mutual distributions of latent topics over the documents (indicated as $\boldsymbol{\theta}$) based on the assumption that the latent topics are shared by the words in the document pairs. In the surrogate creation and data selection processes, the

chosen topic model is trained with all the datasets. Training a topic model provides two important components: the distributions of latent topics over the documents and the distributions of words over topics (henceforth, per-document topic distributions and per-topic word distributions, respectively) are discovered. Both types of distributions are probability vectors and used as the numerical representations, and they are depicted as θ and ϕ in Figure 3.3. Particularly, per-document topic distributions are suitable for the data selection process because each of the θ vectors is assigned to each document as a vector representation as illustrated in Figure 3.1. Using the vectors obtained by topic models as numerical representations of documents, the similarity of each pair of training and test (or surrogate) documents are computed by a distance metric such as cosine distance and Jensen-Shannon (JS) divergence.

For the surrogate data selection, the similarity measure is computed for each combination of the test and training documents, and the training document which is the closest to each test document is chosen to form the surrogate set.

For data selection, we use the hierarchical agglomerative clustering (HAC) algorithm with complete linkage. In the initial stage, singleton clusters are formed with individual documents from the datasets. The clusters are merged based on the similarity measure in each iteration. In each merging process, the clusters with test or surrogate documents are examined to count the training TUs contained in those clusters. Once the desired number of training TUs is collected in the clusters with the test or surrogate documents, the iteration is terminated, and the selected TUs form a training dataset. With this process, we can utilize identified topic distributions to cluster segments based on similarities.

Based on this simple approach, Chapter 4 presents several experiments examining the effectiveness of our data selection methods.

3.4 Thesis Statement

Translation units (TUs) selected for training in such a way that they match the topical content of texts to be translated improve statistical machine translation performance on those texts over

approaches that select training TUs only from a matching source when data selection parameters are carefully determined.

Chapter 4

Preliminary Experiments

4.1 Overview

In this chapter, we conduct several preliminary experiments and discuss the effectiveness of topic adaptation for SMT. These experiments are designed to select document segments to facilitate using topic models and to examine the effectiveness of types of topic models using a synthetic dataset. We also report SMT experiments in an idealized setting, which will be described in the following sections. In the preliminary experiments discussed in this chapter, we used Europarl [55], a diverse parallel corpus used widely in SMT research.

4.2 Experiment 1: Optimal Segment Length

Most of the available parallel corpora are obtained from texts such as speeches, the web, news, and so on, which are relatively unstructured compared with formally written documents. These parallel corpora often do not or cannot provide internal document segment information. On the other hand, segment boundary information is significantly important for topic modeling because document scope is an important factor for topic identification processes [11]. Therefore, it is often the case that studies on topic models use collections of academic journal papers, where document or segment boundaries are easily identified due to their unified length and format [7]. In such a case, topics are reliably discovered in the process by treating each document equally. If such length uniformity is not guaranteed, it is advised that segments contained in corpora should be processed properly in order for them to have similar lengths [11]. While parallel corpora can be separated systematically based on some identifiable boundaries such as paragraphs and speakers, the amounts of text in

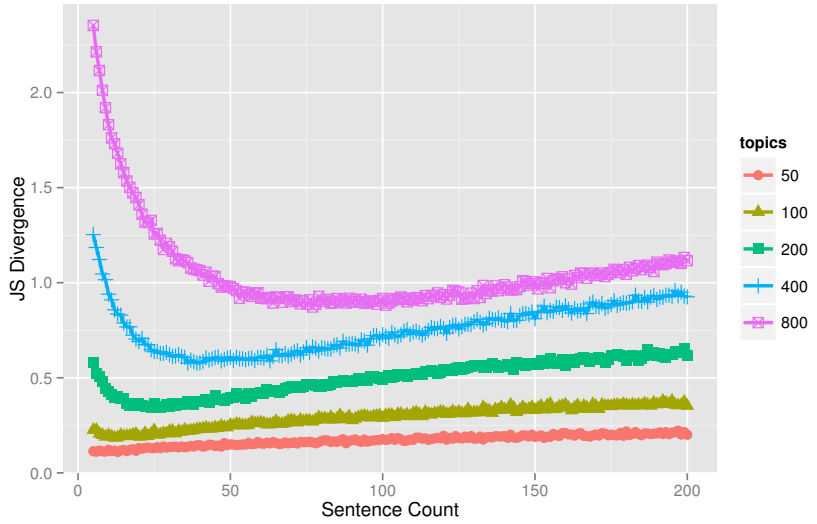


Figure 4.1: Sentence Counts and JS Divergence

these segments are often radically different, which is not desirable for topic modeling [73]. Several previous studies investigate various approaches to segment documents based on topical coherence (e.g., Blei and Moreno [8], Du et al. [24], Eisenstein and Barzilay [30], Hearst [44], Shafiei and Milios [88]). These approaches are effective for corpora with formally written documents but might not be applicable to parallel corpora due to the aformentioned reasons. In order for topic adaptation to be possible, creating reasonable segments from parallel corpora is crucial.

To investigate a solution to this non-trivial issue, we conducted an experiment to identify a reasonable working segment length by uniformly changing TU count in each segment, regardless of speaker or paragraph boundaries indicated in Europarl. To identify optimal segment length, we conduct the following experiment. First, we created a synthetic dataset using the English portion of Europarl by substituting all English words with corresponding random strings. Second, we created a dictinary between English and the syntheic words. Therefore, this synthetic dataset is essentially the same as the original English set but with different word representations. We ran LDA on both original English and synthetic datasets with various topic and TU counts. Then we unmasked synthetic words using the dictionary in the topics to compute JS divergence of all combinations of topics identified for both datasets. Lastly, we averaged the values to examine what TU count yields the lowest JS divergence throughout various topic counts. Figure 4.1 illustrates

the results. As shown, extremely small TU counts cause high JS divergence, which indicates that topic word assignments are significantly different between these datasets especially when the topic count is large. Also, the JS divergence gradually increases after exceeding particular TU counts, depending on the topic count. Although these datasets are basically the same, TU count in a segment affects topic discovery processes significantly. We chose 50 TUs in a segment for the rest of our experiments because JS divergence is relatively low throughout all the topic counts (i.e., across all trend lines) as shown in Figure 4.1.

4.3 Experiment 2: Comparison of Multilingual and Monolingual Topic Models

One question that arises is whether multilingual topic models are more suitable than monolingual counterparts such as LDA for the purpose of obtaining topically similar data for SMT. To investigate this question, we conducted the following experiment:

1. Create synthetic target language data derived from the English source side of data in the Europarl English-German (EN-DE) pair, as described in Section 4.2.

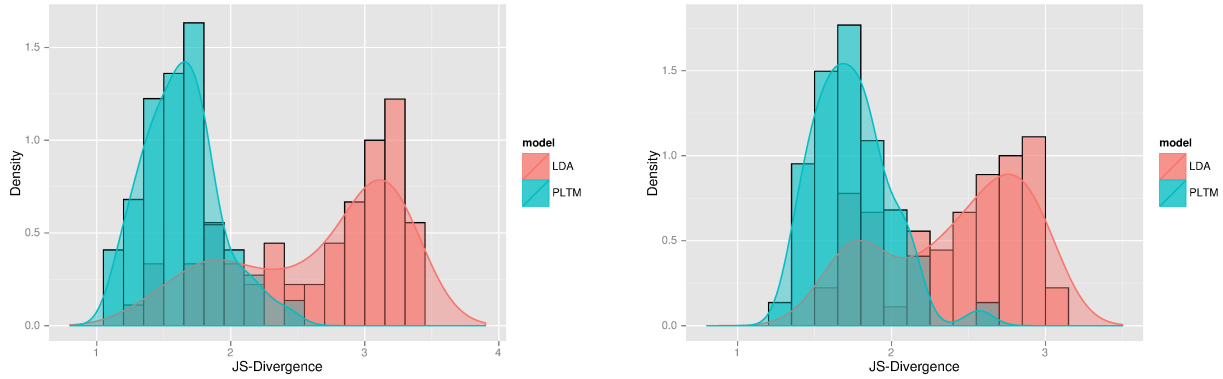
2. Generate two versions of the simulated target language instances from the synthetic data above: One version involves splitting long English words (eight characters or more) in two and assigning separate synthetic words to the respective parts (one-to-many), and the other is merging short words (three characters or less) to the subsequent, adjacent long words (five characters or more) and assigning a single synthetic word to the combined words (many-to-one), in order to simulate real-world target language morphology.

3. Run LDA and PLTM with these two English-Synthetic parallel datasets.

4. Unmask the discovered synthetic topic words and evaluate the similarity of those unmasked topic words with the English topic counterparts.

We chose to use the synthetic data approach (as in the experiment in Section 4.2) above since using real language pairs makes comparison of different topic models difficult due to a lack of objective

metrics for bilingual data. On the other hand, the unmasked synthetic datasets allow us to use direct similarity metrics used for monolingual topic models to compare the capabilities of the topic models by unmasking the synthetic words.



(a) One-to-Many Synthetic Data

(b) Many-to-One Synthetic Data

Figure 4.2: Comparison of LDA and PLTM using English-Synthetic Bitexts

Figure 4.2 shows results from the comparison of topic-word distributions for the English and Synthetic data using (1) two separate LDA models (one for English and the other for Synthetic respectively) and (2) PLTM. In this experiment, we used JS divergence as the metric to compare the similarities between mutual English-Synthetic topics. Because JS divergence shows the proximity of two distributions, we can consider the JS divergences of topic pairs as distance metrics and easily compare the topic identification capabilities of the two different approaches. As shown, the JS divergences of the PLTM topic pairs are distributed in the lower (better) ranges in both merged and split word cases.

These results indicate that using multilingual topic models can be more suitable than using two monolingual topic models for multilingual topic identification of topically similar data from bilingual parallel corpora. We will further investigate the actual impact on SMT performance in the subsequent experiments.

4.4 Experiment 3: Topic Adaptation in the Idealized Scenario

Based on the findings above, we conducted a series of end-to-end SMT experiments with topically similar data extracted from the Europarl English-French (EN-FR) pair using LDA and PLTM. In these experiments, we make the target side of the test set visible in the data selection process described in Chapter 3 so that the PLTM-based clustering process is available. Concequently, we call this experiment setting the "idealized scenario." We examine the four aspects of topic adaptation listed below under this idealized scenario and discuss each of them in the following subsections:

1. The impact of topic adaptation on SMT performance in a simplified data selection approach
2. The impact of topic counts on SMT performance in a simplified data selection approach
3. The impact of topic counts on SMT performance with a pre-selected test set
4. The impact of the size of topical TUs on SMT performance with a pre-selected test set (learning curve)

4.4.1 SMT System Configuration

For SMT processes, we use the Moses MT toolkit [60]. We choose a standard phrase-based translation model with maximum phrase size of 7. We use 5-gram language models (LMs) smoothed by the modified Knese-Ney method with SRILM [91]. We use MGIZA++ [35] for word alignment and MERT [79] for the tuning process with $n = 100$ for *n*-best lists. For MT performance evaluation, we compute the BLEU score according to [81]:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{4} \log p_n\right) \times 100, \tag{4.1}$$

where BP indicates the brevity penalty which is computed as follows:

$$BP = \min\left(1, \frac{\text{output sentence length}}{\text{reference sentence length}}\right). \tag{4.2}$$

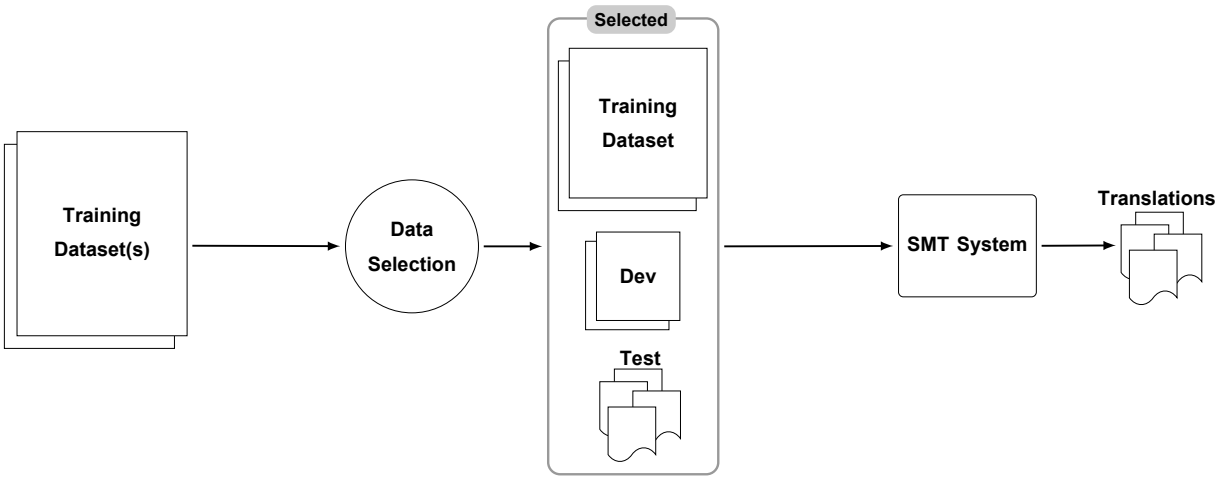


Figure 4.3: Simplified Topic Adaptation

The sentence length is measured by the number of words in the sentence. p_n in Equation 4.1 is the *n*-gram precisions, where the maximum *n* is 4. The BLEU score yielded by Equation 4.1 ranges from 0 to 100.

4.4.2 Topic Adaptation in the Simplified Data Selection Approach

To simplify the topic adaptation process, the SMT experiment discussed in this subsection are conducted under the topic clustering procedure shown in Figure 4.3. As illustrated, we create all three SMT datasets from the training set after the data selection process is completed. The purpose of this experiment is to investigate the impact of topic adaptation on SMT performance with limited confounding factors. The following procedure selects topically similar and out-of-topic portions from the corpus:

1. Assign topics to document pairs by running PLTM to identify 100 topics in EN-FR data. This number is chosen arbitrarily.

2. Characterize each document pair based on their most dominant topics, and group the document pairs according to those topics.

3. Group the topics using the HAC algorithm with JS divergence as a distance metric.

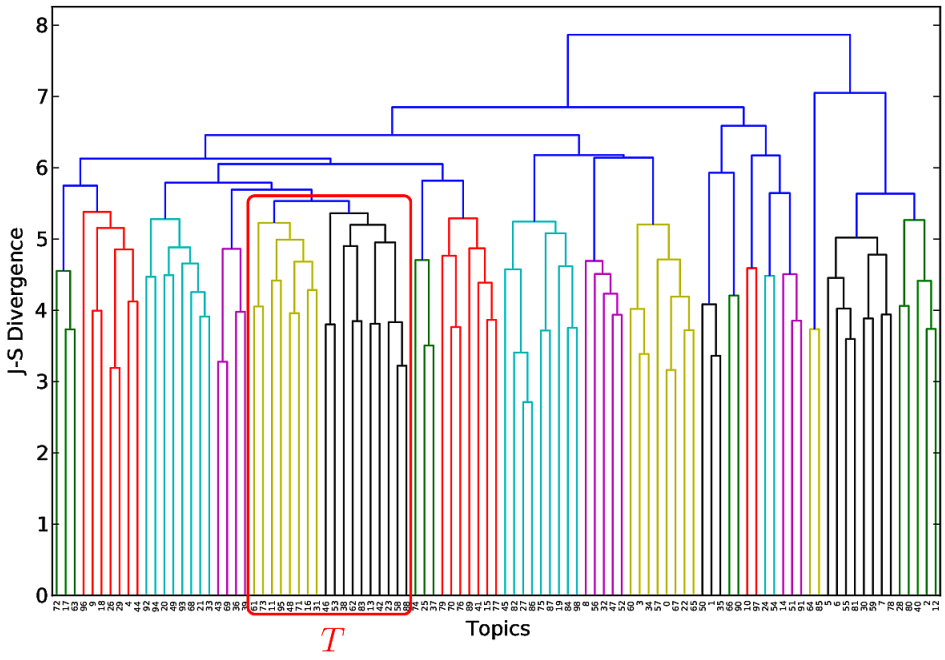


Figure 4.4: Dendrogram of Clustering Process

4. Choose a cluster of similar topics. we choose a cluster of topically similar data and call it T (topical) in this experiment. Figure 4.4 illustrates T chosen based on the clustering process.

5. Collect document pairs which do not belong to cluster T Call the out-of-topic documents NT (non-topical).

6. Generate training (T_{train} and NT_{train}), tuning (T_{dev} and NT_{dev}), and test (T_{test} and NT_{test}) datasets by splitting these topically similar and out-of-topic data, and build phrase-based SMT models with these datasets. Restrict the size of both the tuning and test sets at 2,500 TUs.

7. Tune the models with MERT, translate the test sets with the models, and report respective BLEU scores.

An example bilingual topic discovered by PLTM during the data selection process is shown in Figure 4.5. Through this process, we obtained 300,000 TUs as a topically similar dataset T and just as many out-of-topic TUs in NT. Furthermore, we created two additional datasets C

(a) Topic 33 (English)

(b) Topic 33 (French)

Figure 4.5: An Example Topic Discovered by PLTM

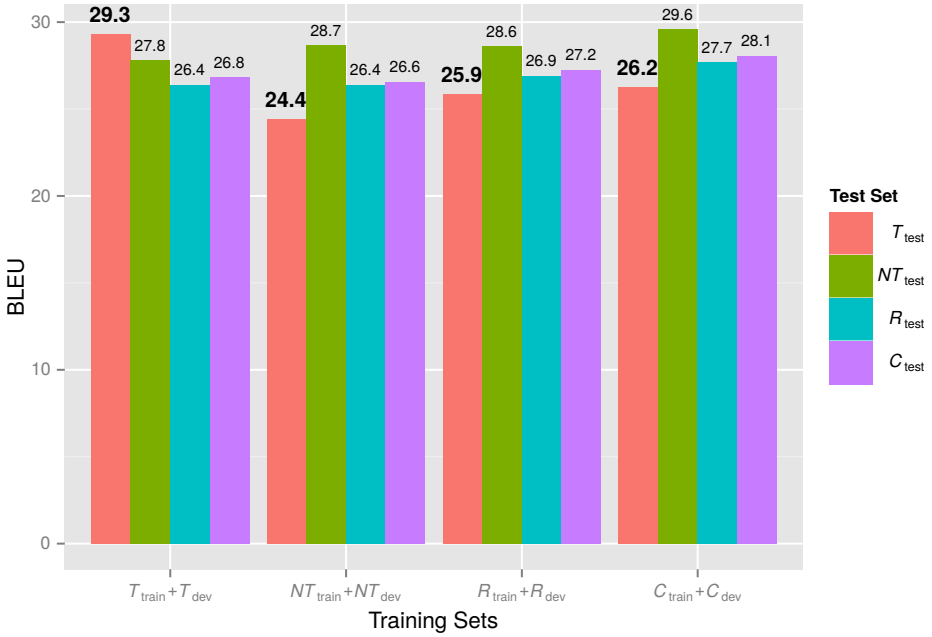


Figure 4.6: SMT Performance Results

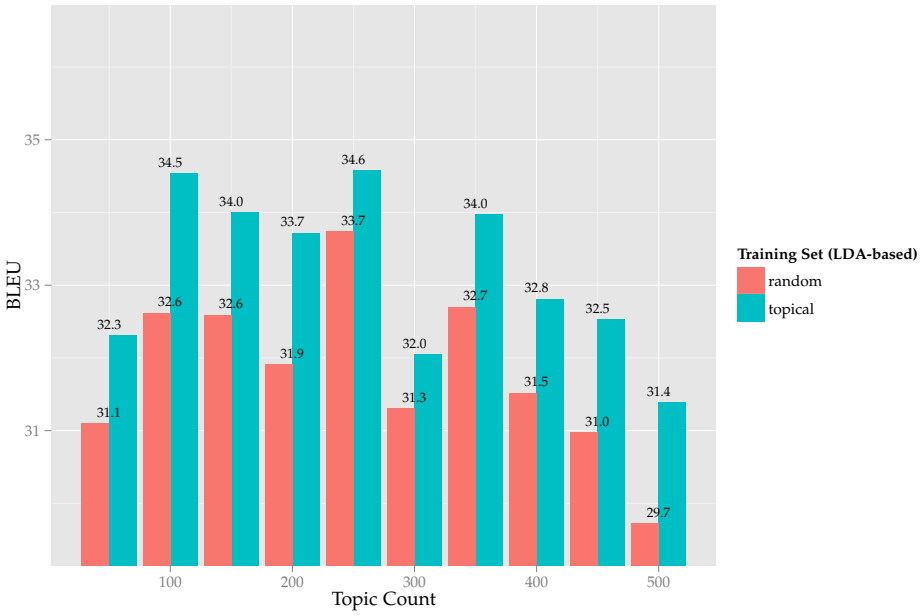
(concatenation of T and NT) and R (a dataset with 300,000 randomly selected TUs — equally — from T and NT) for comparison.

Figure 4.6 shows the SMT performance measured by BLEU with these datasets. Regarding the BLEU scores against T_{test}, $T_{\text{train}} + T_{\text{dev}}$ (i.e., an MT system trained and tuned on T_{train} and T_{dev}) significantly outperformed the other training datasets. On the other hand, the performance on NT_{test} is lower than the other training and tuning sets (approximately 1 to 2 BLEU points). The performance on R_{test} and C_{test} is also lower than the other datasets although the difference is not so different as the case of NT_{test}. This result clearly indicates that topical similarity of training and test data is more beneficial to SMT performance than increasing data quantity through simple concatenation. Also, this result confirms that the topic adaptaion using PLTM effectively identifies similar data from the topically-rich Europarl corpus in an effective manner.

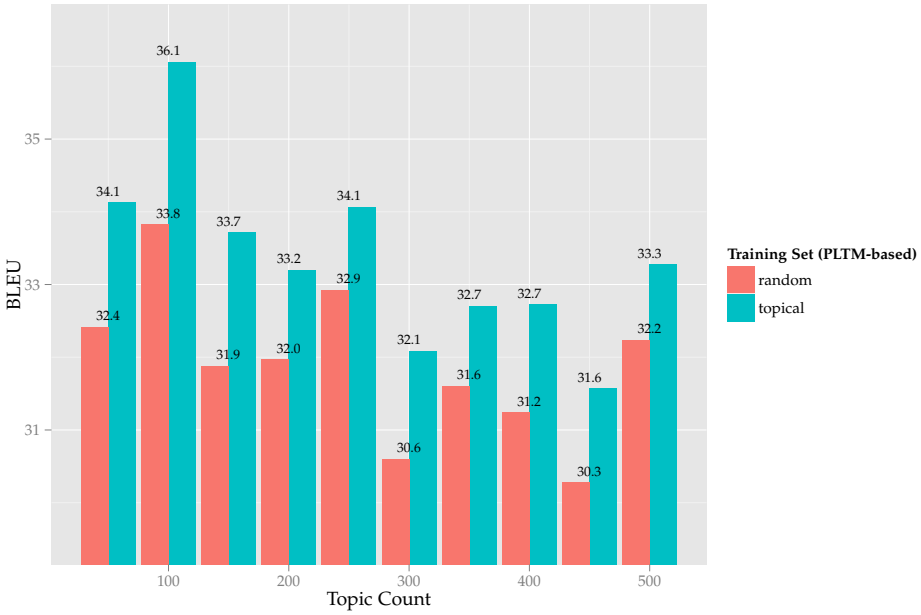
4.4.3 Topic Counts in the Simplified Data Selection Approach

In the experiment above, we fixed the topic count to 100. As in the case of JS divergence in the segment granularity experiment described in Section 4.2, we realize that topic count affects SMT performance and that BLEU scores with topically similar data vary significantly as this parameter changes. In this experiment, we investigate the impact of topic counts under the simplified data selection approach of Figure 4.3. The experiment procedure is essentially the same as that in Subsection 4.4.2, but the topic counts are swept from 50 to 500 in increments of 50. We used both LDA and PLTM for data selection, and p (the topical cluster) and m (the random cluster) are considered in this experiment. The results are shown in Figure 4.7. As seen in each topical and random bar pair in the figure[1], BLEU scores with the topical clusters outperform the random cluster counterparts by more than one BLEU point across all the topic counts. Therefore, the trend found in Subsection 4.4.2 holds across other topic counts with both LDA and PLTM under this simplified data selection approach.

[1]Since the test set is different for each topic count, the BLEU scores of different topic counts are not comparable.



(a) LDA

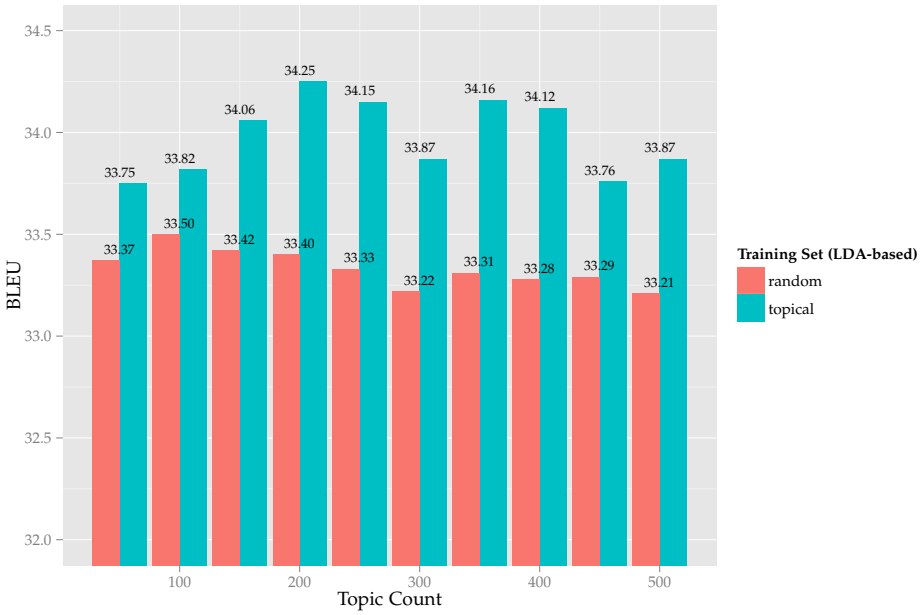


(b) PLTM

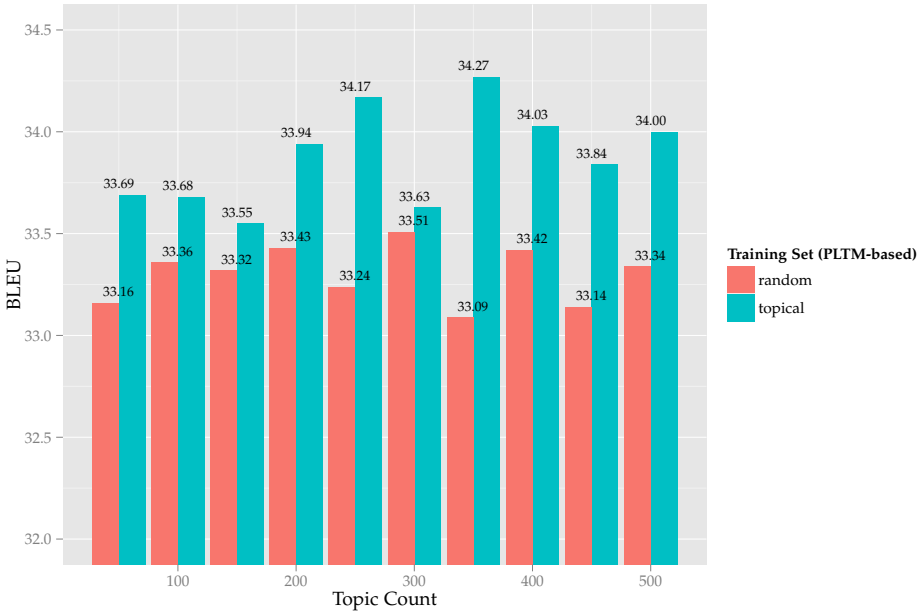
Figure 4.7: Topical data selected with various topic counts and BLEU scores

4.4.4 Topic Counts with Pre-Selected Test Set

The simplified data selection approach (Figure 4.3) used in the two previous experiments selects the test set from the topical cluster. It was a useful approach to demonstrate the impact possible



(a) LDA



(b) PLTM

Figure 4.8: Topical data selected with various topic counts and BLEU scores

in the ideal case. However, this is not realistic because a test set must be provided by the user. In this experiment, we investigate the effectiveness of topic adaptation using a test set selected from Europarl *prior to* the data selection process. Since this experiment deals with a pre-selected test set, the topic adaptation process is similar to the one depicted in Figure 3.2b. However, this experiment

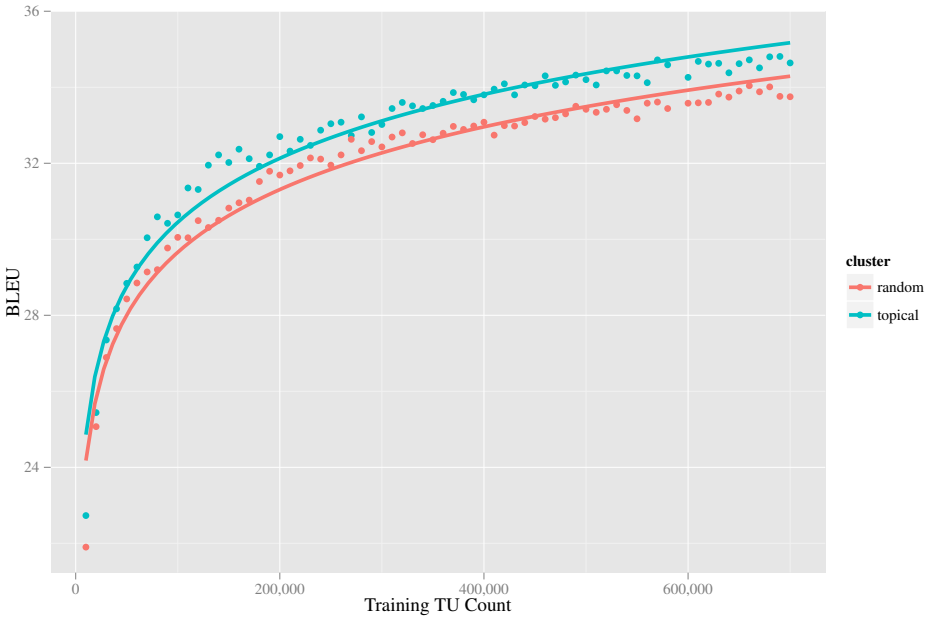
is still idealized, since the target side of test set is still visible during the data selection process. In this experiment, we generate a test set with randomly selected 50 segments from Europarl and exclude them from the training dataset. Then we conduct data selection with LDA and PLTM as described in Subsection 4.4.2. We repeat the same process with different topic counts in the same manner as Subsection 4.4.3. The results are shown in Figure 4.8. As indicated in the figure, the topical clusters outperform the random counterparts across all the topic counts. This shows that the trend seen in Subsection 4.4.3 still holds even when a test set is chosen prior to the data selection process.

4.4.5 Learning Curve

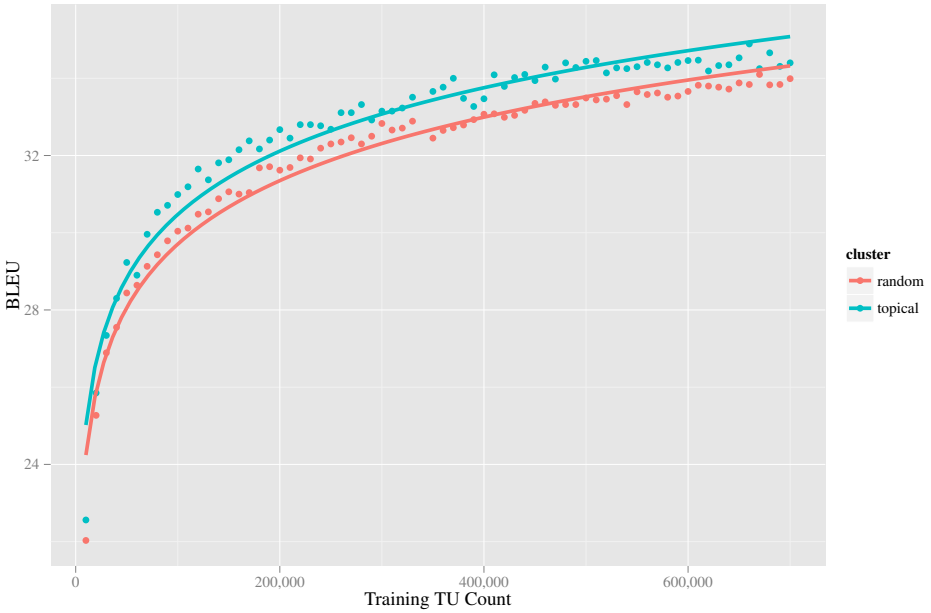
The last experiment in the idealized scenario is to examine the impact of increase in TU counts to SMT performance. The topic adaptation procedure is essentially the same as Subsection 4.4.4, where we used a pre-selected test set. The difference from the previous experiment is that we change the TU count contained in the training set from 10,000 to 70,000 in increments of 10,000 and fix the topic count to 200 because the SMT performance was best between 200 and 400 in the previous experiment. The results are shown in Figure 4.9. These results clearly indicate that topical clusters with both LDA and PLTM always outperform random clusters for all TU counts. Therefore, the results from Subsections 4.4.3 and 4.4.4 still hold across larger training sets.

4.5 Discussion

The results from the four SMT experiments described above indicate that data selection with topic adaptation is effective in the idealized scenario. Regardless of topic counts, TU counts, and types of topic models, topically-clustered training data always outperformed random data. Also, the segments with 50 TUs each generated based on the experiment in Section 4.2 functioned reasonably well although the approach was very simple. The contribution from these findings is that training data, chosen properly, will yield superior translation results. Furthermore, topics identified by topic models can help select relevant training data according to the provided test data.



(a) LDA



(b) PLTM

Figure 4.9: TU count increase and BLEU scores

In the next chapter, we investigate the effectiveness of topic adaptation in the realistic scenario, where no target-language information in the test set is available, using the corresponding topic adaptation approaches discussed in Chapter 3.

Chapter 5

Realistic Scenario

5.1 Overview

In this chapter, we investigate the effectiveness of data selection with topic adataption using a blind test set, where no target-side information is available during the data selection process. To distinguish this scenario from the idealized scenario which we discussed in the previous chapter, we call this experimental setting the *realistic scenario*. We conduct several experiments under the realistic scenario. First, we evaluate the SMT performance with a topical training set selected with the aid of a surrogate set. Second, for the sake of comparison, we use additional features called top-*n* per-document [96] along with per-document topic distributions for selecting training data. Third, we compare the topic adaptation approaches with the other existing data selection approach by [76]. Lastly, we leave behind the standard benchmark datasets and apply the topic adaptation approaches to a real-world dataset.

5.2 Surrogate Selection 1

To conduct data selection with both monolingual and bilingual topic models under the realistic scenario, it is necessary to generate a data set which is similar to the test set but which also has the target-side information (i.e., reference translations) available. In Subsection 3.3.3, we described the approach to generate such a dataset from the training set, and we called this dataset a "surrogate" set because this dataset is used as a substitute for the blind test set in the training data selection process. In this section, we examine several approaches for surrogate set creation. As illustrated in Figure 3.2c, a surrogate set must be generated prior to the training data selection process, when

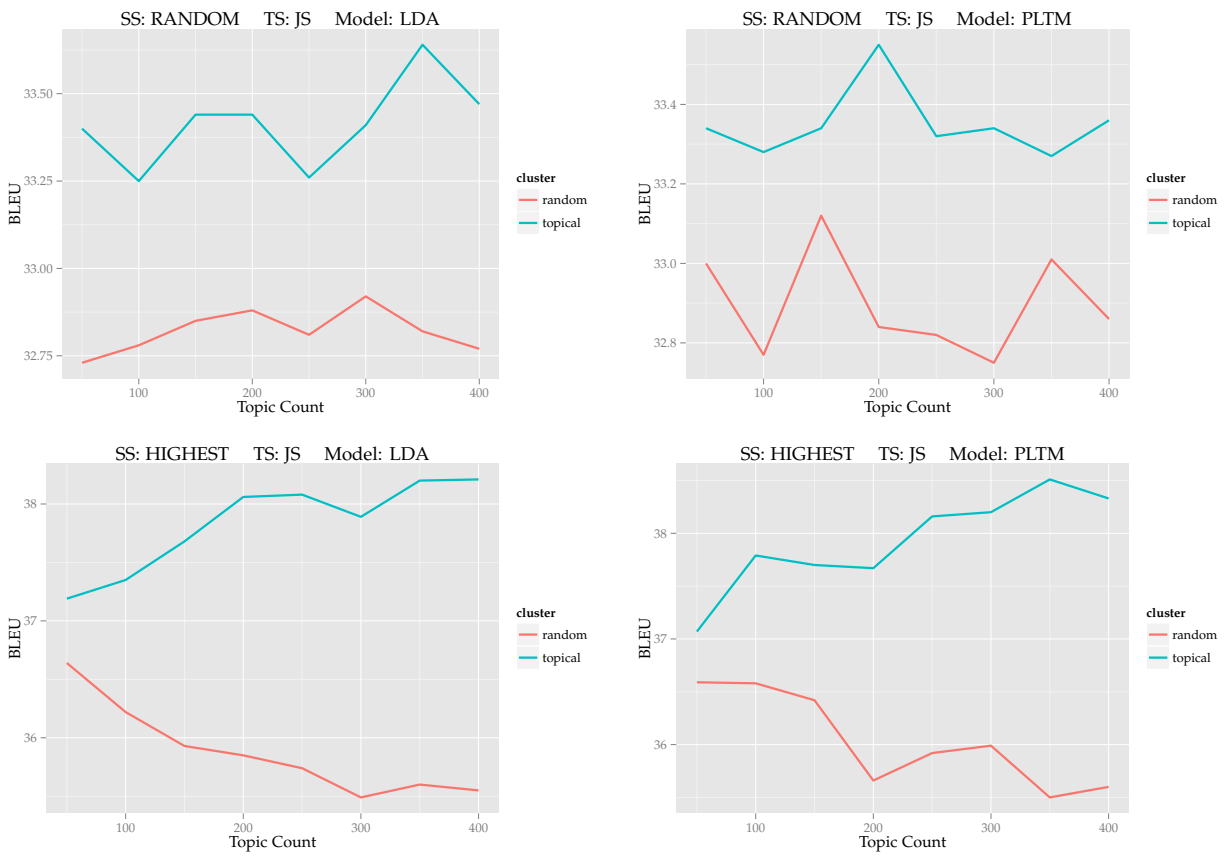


Figure 5.1: SMT results evaluated on the surrogate sets. SS (Surrogate Selection) and TS (Training Selection) indicate the types of dataset selection methods.

using test and training datasets. Since the surrogate set needs to be topically similar to the test set, the surrogate set creation process is conducted in a manner similar to the training data selection process described in the previous chapters. The main difference between the training and surrogate selection processes is that in the training selection process, multiple segments similar to each test document must be found to obtain a desired number of TUs (e.g., 300k), whereas only one surrogate segment corresponding to each test document needs to be found in the surrogate selection process. Because of this simplicity, there are more approaches available to the surrogate selection process than the training selection process. However, since the target side of the blind test set is not visible, surrogate selection must be accomplished only with LDA. Based on these settings, we investigate the following four approaches in this section:

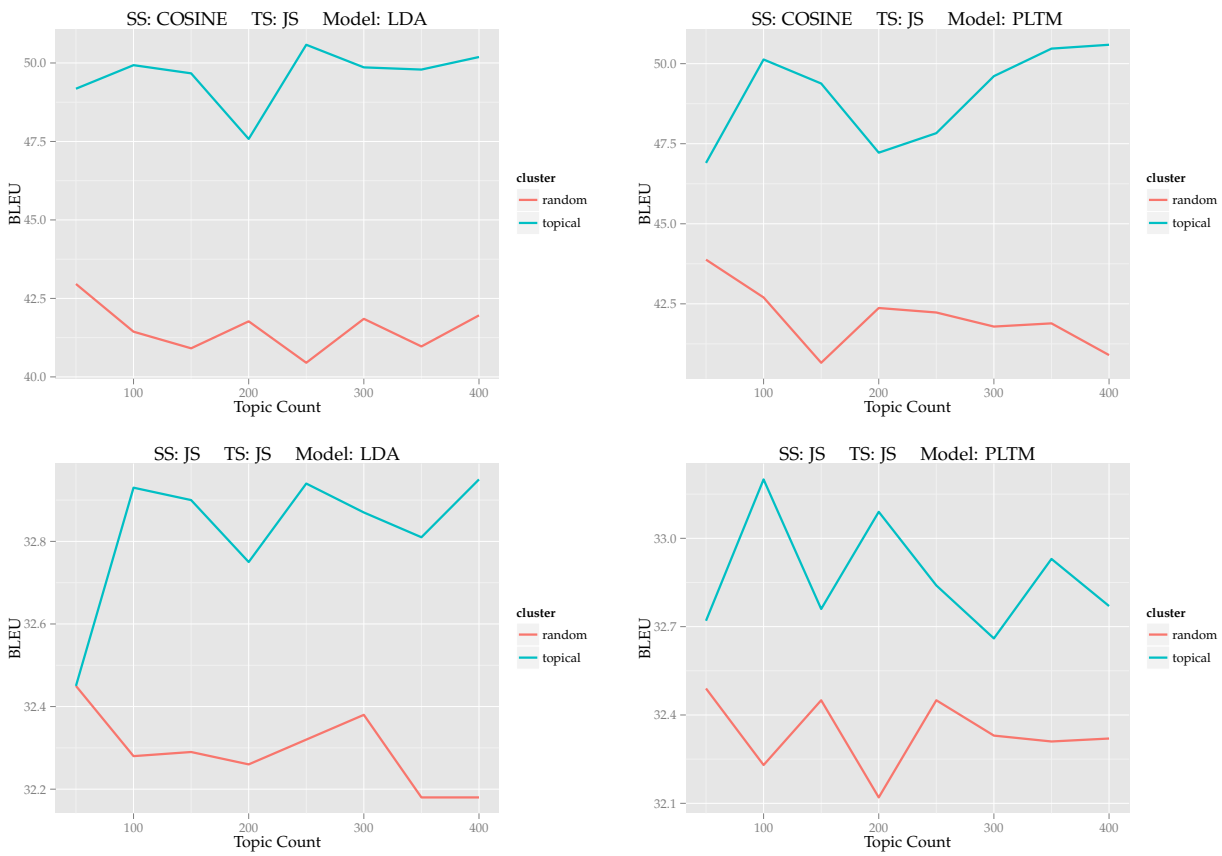


Figure 5.2: SMT results evaluated on the surrogate sets (Cont.)

1. Random: randomly select the same number of segments as that of the blind test documents from the training set.

2. Highest: choose a segment which has a mode at the same topic index in the per-document topic distribution vector (henceforth, θ vector) as the test segment.

3. Cosine: choose a segment which is the most similar to the segment in the training set with the lowest cosine distance measure between the θ vector pairs for the respective documents.

4. JS: Choose a segment in the training set with the lowest JS value computed between θ vector pairs.

For the surrogate selection process, we consistently use the topic count of 100 for LDA. For Highest, Cosine, and JS, there is a possibility that the same training segment is chosen for multiple test

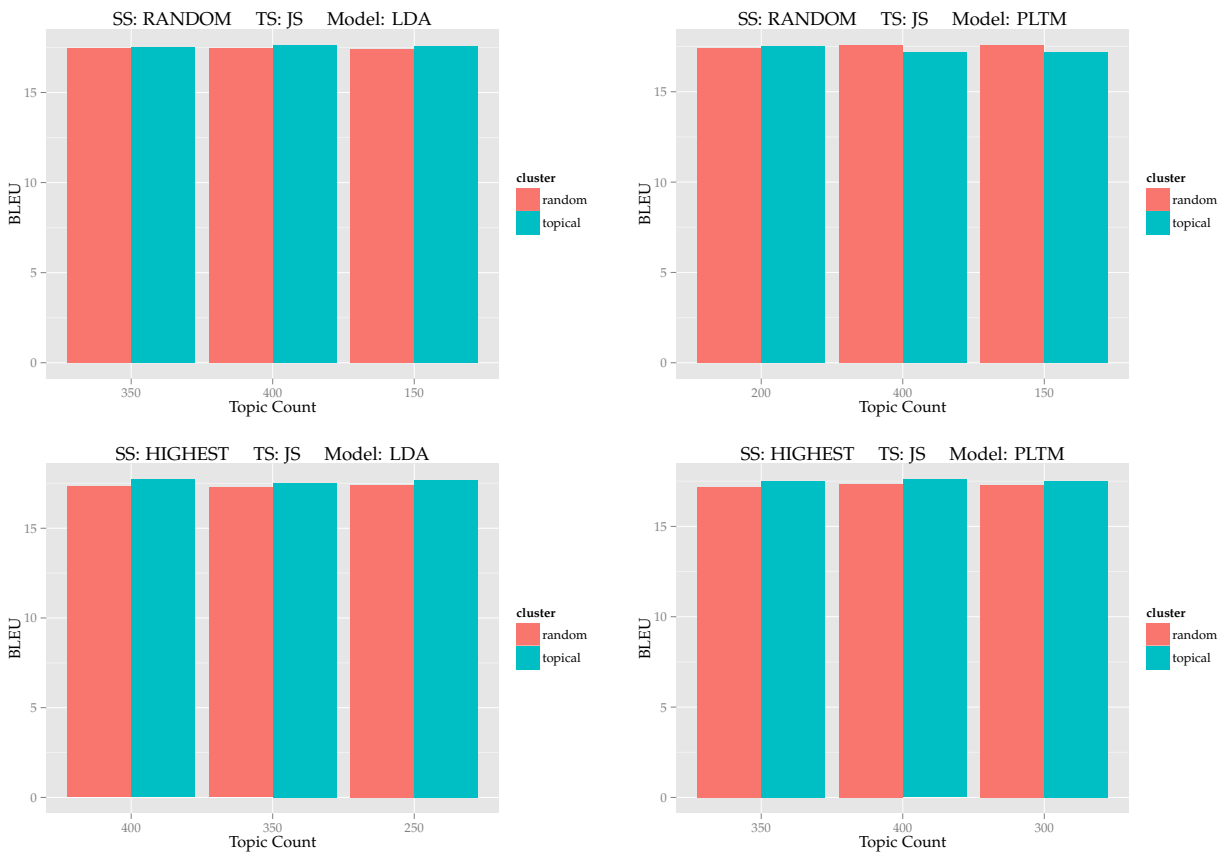


Figure 5.3: Three BLEU scores on the blind dataset with the topic counts of the best three BLEU scores on the surrogate dataset. The topic counts are ordered according to the BLEU scores evaluated on the surrogate set.

documents. In that case, we simply include multiple copies of the same segment (no deduplication). In these experiments, we choose to use `newstest2008`, a blind test set used in the WMT workshop 2008.[1] This blind set contains 2,051 sentences to be translated from various news articles and is segmented in blocks of 50 TUs each as for the other datasets in the data selection process. The experiment is conducted based on the PLTM-Dev method shown in in Figure 3.2c with slight modifications:

1. The generated surrogate set is used as a test set rather than a development set in SMT, in order to sweep topic counts as in the experiment described in the experiment of topic count sweeping with a pre-selected test set in Section 4.4.4. Therefore, the development set for the

[1] Available at `http://www.statmt.org/wmt08/shared-task.html`.

SMT tuning process is generated along with the training set as in the LDA-Dev approach shown in Figure 3.2b.

2. For comparison, we use both LDA and PLTM for the training selection process.

3. Based on sweeping the topic count parameter, we choose the three topic counts yielding the three highest BLEU scores with the surrogate set and create three topical training sets using the topic counts for SMT and evaluate the performance of the three systems with the blind test set.

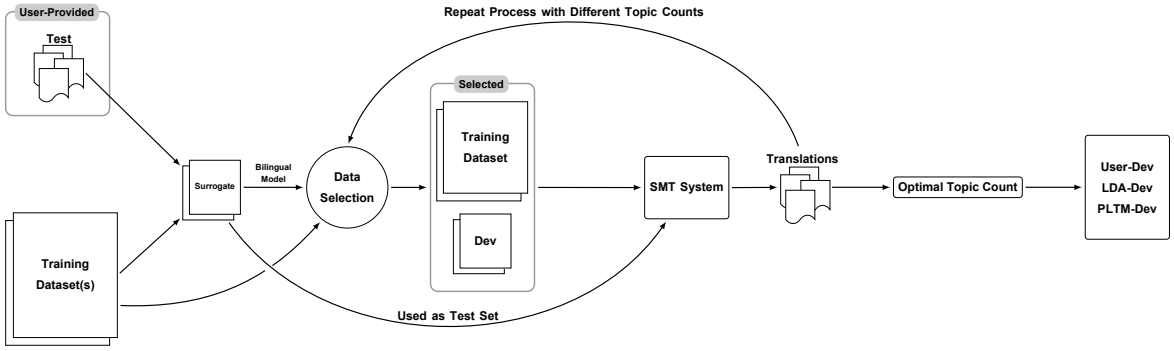


Figure 5.4: PLTM-Dev Modified for Topic Count Sweeping with a Surrogate Set

Figure 5.4 shows the schema of topic count sweeping with a surrogate set. The significant aspect of this approach is that by using a surrogate set for sweeping the topic count parameter, we are able to identify ideal topic counts without examining the target translations of the blind set in the realistic scenario. The other settings for this experiment are the same as Section 4.4.4. Figures 5.1 and 5.2 show the topic counts and the BLEU scores evaluated on the surrogate dataset chosen by the four surrogate selection approaches described above. For clarity, we label surrogate selection as SS and training data selection as TS in the figures. Overall, the topical training sets outperform the random counterparts with all the topic counts, regardless of the type of surrogate selection method. This is a similar trend seen in the results of the experiment in Subsection 4.4.4. Also, The superiority of the performance with the topical training sets is confirmed in both LDA and PLTM for surrogate selection.

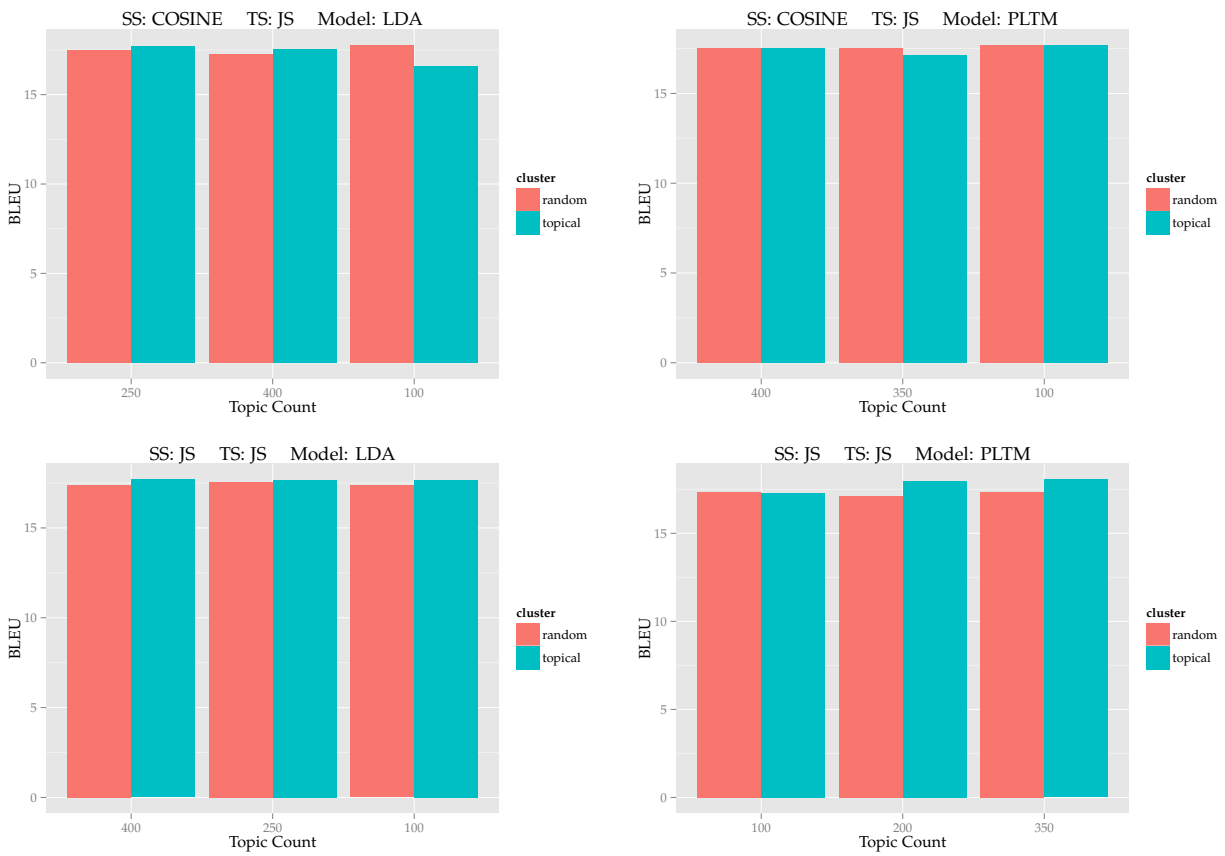


Figure 5.5: Three BLEU scores on the blind dataset with the topic counts of the best three BLEU scores on the surrogate dataset (Cont.)

Figures 5.3 and 5.5 show the BLEU scores on the blind set using topical training sets generated with three topic count values chosen based on the SMT performance on the surrogate sets. Unlike the results on the surrogate sets, topical training sets do not always outperfom the random counterparts on the blind test set. However, Highest for surrogate selection works for all three topic counts with both LDA and PLTM.

5.3 Surrogate Selection 2

The BLEU scores shown in Figures 5.1 and 5.2 are produced with single SMT processes. However, the MERT algorithm used in the tuning process does not lead to consistent BLEU scores because the search space is non-convex, so the training algorithm is highly likely to be caught by different local optima in different runs [17, 32]. Hence, it is difficult to say that the choices in the topic

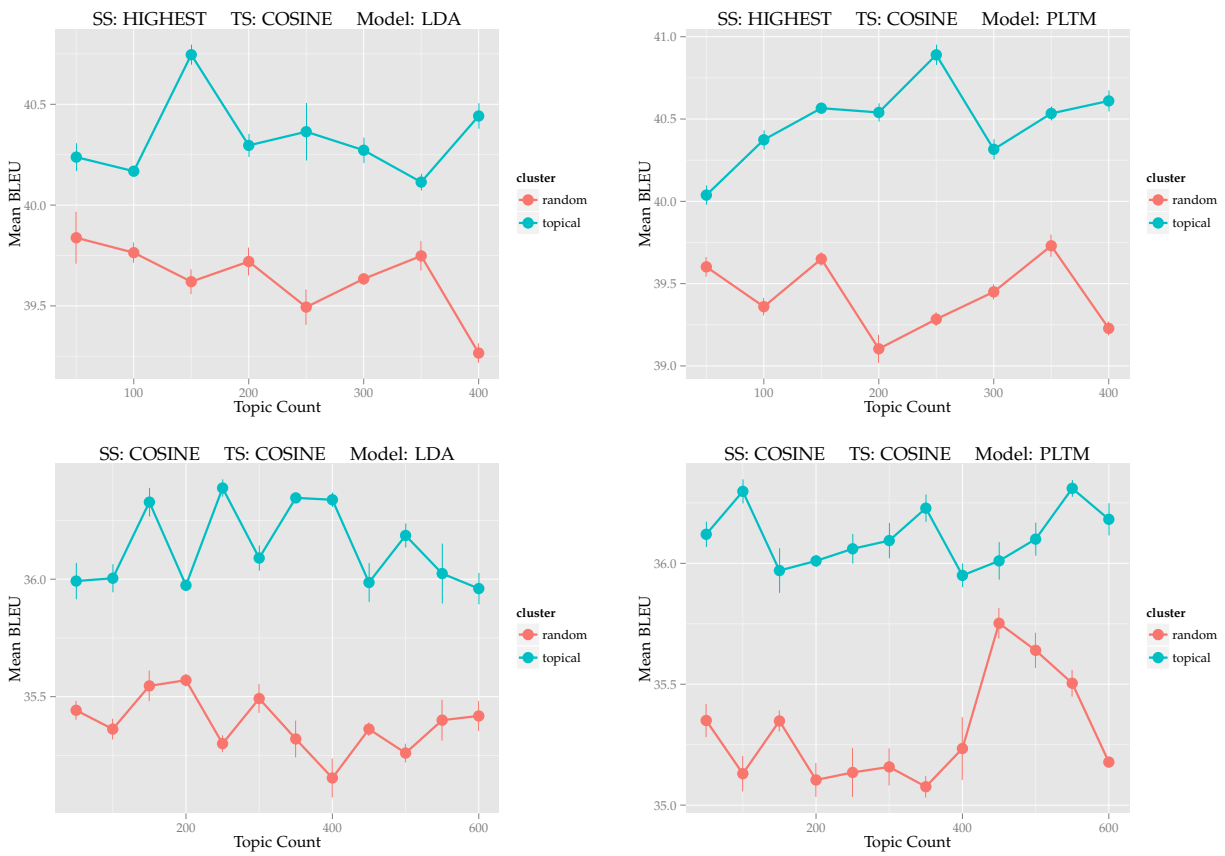


Figure 5.6: SMT results evaluated on the surrogate data. SS (Surrogate Selection) and TS (Training Selection) indicate the types of segment selection methods. The points on the line plots show BLEU scores averaged by five runs, and vertical bars on the points indicate the standard errors.

counts based on Figures 5.1 and 5.2 are optimal. To choose optimal topic counts more reliably, it is necessary to obtain averaged BLEU scores through multiple runs with the same training, tuning and test sets, which is a widely-used SMT evaluation method for system comparison (e.g., Hu et al. 46, Kirchhoff and Bilmes 52). We choose to run SMT system training and tuning five times. The other experiment settings are the same as Section 5.2. However, we omit the random surrogate selection approach to focus on the other three approaches. For training selection, we use Cosine in the case of Highest and Cosine surrogate selection and JS for JS surrogate selection.

Figures 5.6 and 5.7 show the BLEU scores evaluated on the surrogate sets with topic counts from 50 to 600. As in the cases shown in Figures 5.1 and 5.2 in the previous section, topical clusters with all three surrogate selection approaches outperform the random counterparts although the

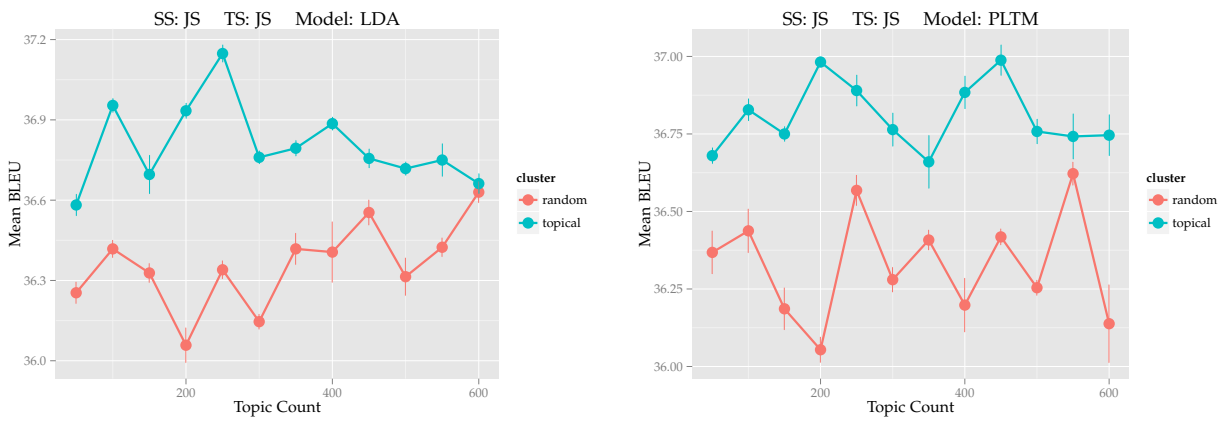


Figure 5.7: SMT results evaluated on the surrogate data (Cont.)

BLEU scores in the JS approach at the high topic counts (600 for LDA and 550 for PLTM) are very close or almost identical.

Figure 5.8 shows the BLEU scores evaluated on the blind set using the three topic counts that yield the best BLEU scores on the surrogate set. As shown, Cosine works well in both LDA and PLTM cases and is able to identify topical clusters that outperform the random counterparts effectively while the other surrogate selection methods exhibit mixed results.

5.4 Top-*N* Per-Document (TNPD) Features

The BLEU scores with Cosine surrogate selection shown in Figure 5.8 are consistently better than those with the random clusters at the selected topic counts. However, the differences in the BLEU scores on the blind set observed in the experiment are rather small ($<$ 1 BLEU point). Seeking to attain better SMT performance with topical training sets, we extract additional features from the training and test segments along with the θ vectors in the training selection process, and we examine if the additional features help identify training segments which are topically similar to the test segments more effectively.

The additional features we examine are top-n per-document (TNPD; Walker 96), features which are a varation of TF-IDF. The conventional TF-IDF score is computed for a word in a

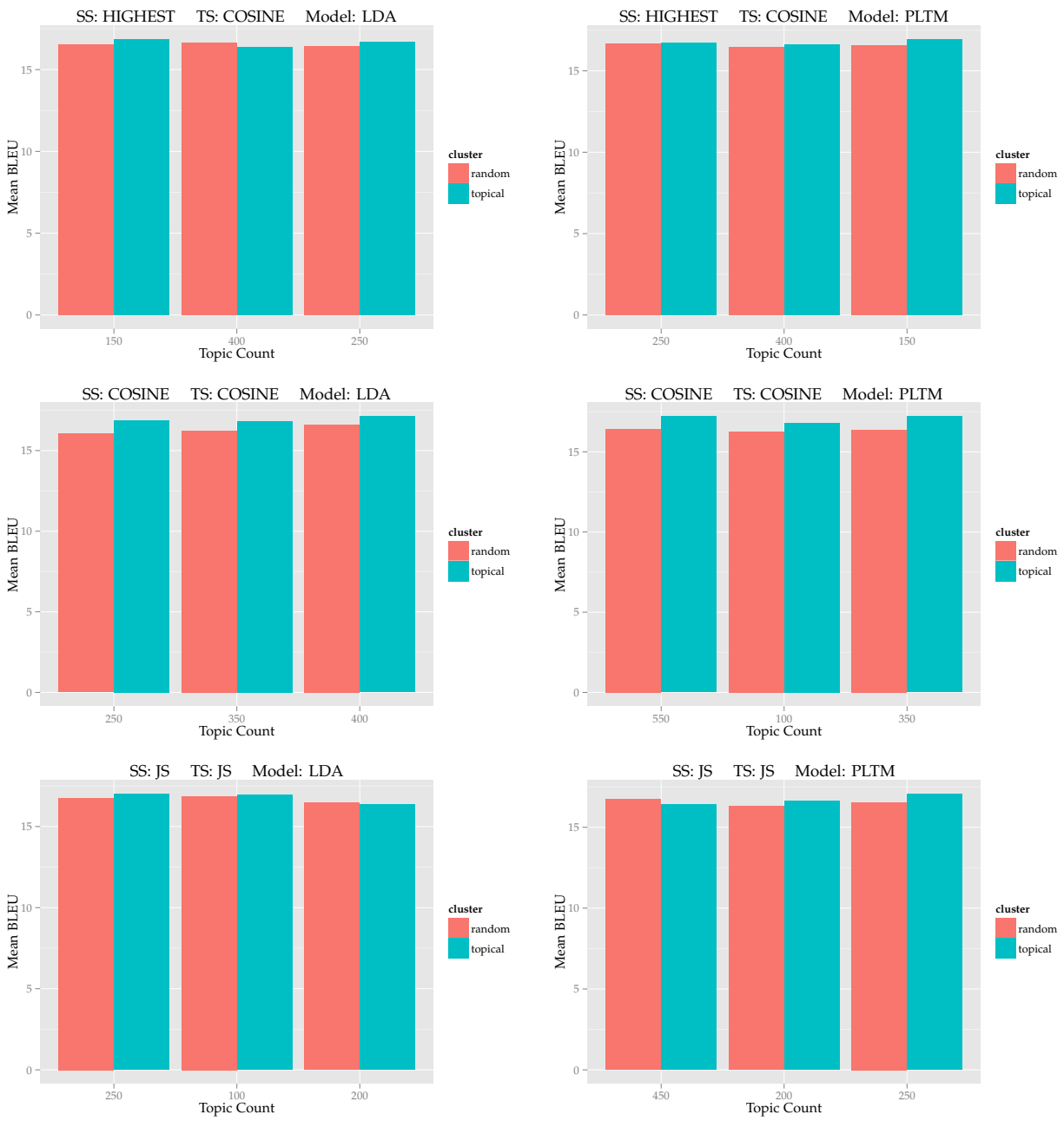


Figure 5.8: Three BLEU scores on blind set at the topic counts of the best three BLEU scores on the surrogate set. Topic counts are ordered according to the BLEU scores evaluated on the surrogate set.

document as follows:

$$\text{TF-IDF}(w, d) = \frac{f(w, d)}{\max_{w' \in d} f(w', d)} \cdot \log_2 \left(\frac{N}{n_w} \right) \tag{5.1}$$

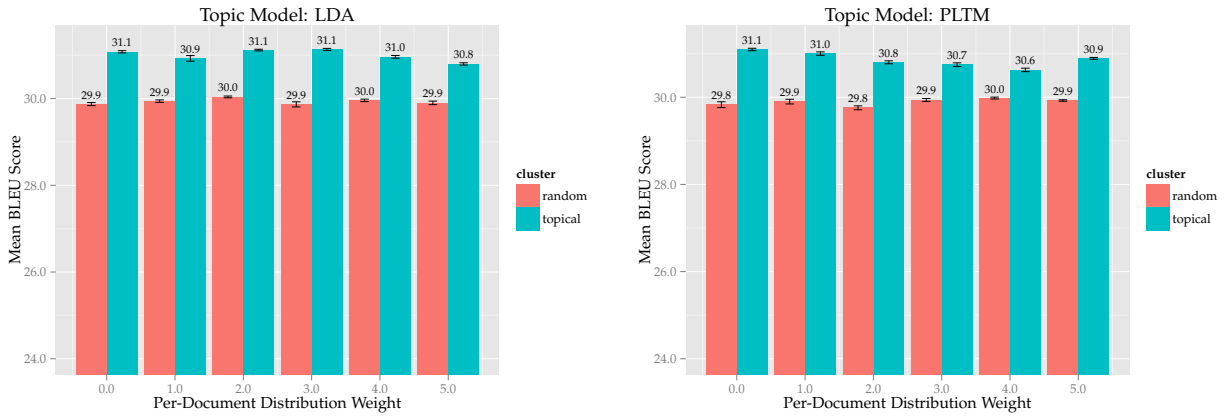


Figure 5.9: Weighted Feature Combination and BLEU scores (Surrogate Set)

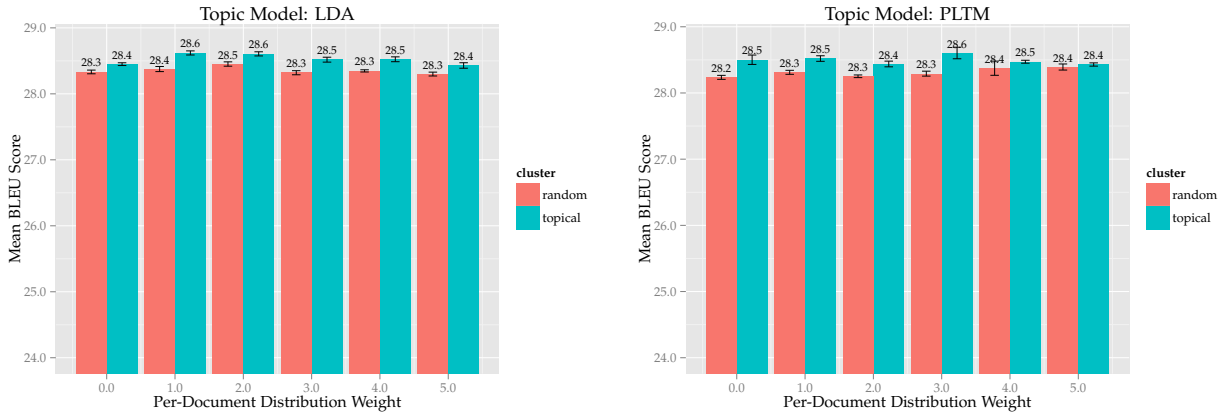


Figure 5.10: Weighted Feature Combination and BLEU scores (Blind Set)

where d is a document in a corpus, w is a word in the document d, N is the total number of documents in the corpus, n_w is the number of documents which contain the word w, and $f(w, d)$ is the number of occurrences of the word w in the document d. The problem with TF-IDF is that the dimensionality of the extracted features is very high. Rather than using all of the vocabulary items in the document collection, we select the top n words from each document to limit the feature dimensionality. Thus we call this feature extraction approach TNPD. In this experiment, we choose $n = 1$ (i.e., one initial feature from each document). This allows us to limit the dimension of features to 5,865 in the segments of the source side of EN-FR Europarl. We extract these features and form a feature vector to create a representation of each segment.

To combine TNPD and per-document topic distribution features, we normalize respective vectors independently to make them unit vectors, set a mixing weight to values in the range from 0.0 to 5.0 in increments of 1.0 on per-document topic distributions, and concatenate them to be a single vector for each segment. These vectors are used for the clustering process in training selection. In this experiment, the topic count is fixed to 200 without sweeping the topic count parameter to examine the impact of the TNPD features on SMT performance. We use the same surrogate data selection approach used in the previous two sections. However, the blind test set used in this experiment is `test2007`, a held-out blind set created from Europarl containing 2,000 sentences to be translated.[2] We replace `newstest2008`, the previous blind set, with `test2007` to examine the effectiveness of the approach under the setting where training, development, and blind sets are generated from the same corpus and the difference in SMT performance between topical and random datasets on the blind set can be more perceivable than that of `newstest2008`. For BLEU computation, we average the scores of five runs as in the experiment of the previous section. Figure 5.9 shows the BLEU scores evaluated on the *surrogate set* with six different feature weights. For both LDA and PLTM, the differences of BLEU scores between topical and random clusters are consistent, and those of the topical training sets are approximately one BLEU point higher than those of random counterparts across all the weight values. Figure 5.10 shows the BLEU scores evaluated on the *blind set* with the six different feature weights. As seen in Figure 5.9, the topical training sets outperform the random counterparts throughout all the feature weights although the differences are much smaller. In the previous section, we observed the same superiority of the topical clusters generated by per-document topic distributions only. Therefore, it is difficult to say that TNPD features enhances the training selection capability. Also, note that the BLEU scores with the weight 0.0 means that training data are selected with TNPD features only. The surrogate set results in Figure 5.9 show that the SMT performance with TNPD features only is the best. However, the blind set results in Figure 5.10 show that the weight 2.0 for LDA and 3.0 for PLTM yield the best performance. These results indicate that using per-document topic distributions only is more

[2] Available at `http://www.statmt.org/wmt07/shared-task.html`.

effective to select topical training data than combining TNPD features with per-document topic distributions.

5.5 Comparison with Cross-Entropy Approach

So far, we have compared the topic adaptation approaches only with random selection. In this section, we compare the three approaches (User-Dev, LDA-Dev, and PLTM-Dev) discussed in Chapter 3 with a data selection method called the cross-entropy approach by [76], along with unadapted SMT systems as baselines. We describe the cross-entropy approach and show the experiments in the following subsections.

5.5.1 Cross-Entropy Approach

The cross-entropy approach by [76] is one of the most popular methods for data selection for SMT. The data selection procedure is as follows:

1. Train a language model with the in-domain training set (LM_{in})
2. Train another language model with the out-of-domain training set (LM_{out})
3. Score each sentence in a data set called $POOL$ with the difference of cross-entropy values using the trained language models.
4. Rank additional sentences according to the scores computed in step #3 and use the n highest-ranked sentences as additional in-domain data.

The additional sentences indicate data contained in available training sets, which do not belog to either the in- or out-of-domain sets used for language model training. The score for ranking is computed as follows:

$$\underset{s \in POOL}{\text{top } n} \quad H(s, LM_{in}) - H(s, LM_{out}), \tag{5.2}$$

where $POOL$ is a large collection of additional sentences potentially similar to the in-domain data, and s is a sentence contained in $POOL$. $H(s, LM)$ means that the cross-entropy of the sentence s according to the language model LM. "Top *n*" indicates selecting top *n* sentences according to the difference in cross-entropy scores: the lower the difference, the higher the rank.

Because this approach is designed for domain adaptation, the user must identify in- and out-of- domain data *a priori*, which does not enable a direct comparison with the topic adaptation approaches. To make the comparison possible, Kirchhoff and Bilmes [52] modify the score computation as follows:

$$\underset{s \in TRAIN}{\text{top } n} \quad H(s, LM_{TEST}) - H(s, LM_{TRAIN}), \tag{5.3}$$

where $TRAIN$ indicates the training set for SMT systems, $TEST$ indicates the blind test set for SMT evalutation. This approach enables us to collect sentences from the provided training set based on similarity to the blind set without pre-specifying in- and out-of-domain sets. Because we determine the data similarity with language models, only the source side of TUs in the parallel corpus is involved in this process; thus, no unjustifiable advantage is to be had.

5.5.2 Experiment Setup

The experiments are conducted based on the setting described below (see also Kirchhoff and Bilmes [52]):

1. Let the blind test set be `test2007`, as in the experiment of the previous section, and let the development set be `dev2006`.[3]

2. Sweep the selected training data size from 10% to 40% of the entire training set.

The experimental configurations are as follows:

[3] Available at `http://www.statmt.org/wmt07/shared-task.html`.

1. Conduct the three topic adaptation approaches (User-Dev, LDA-Dev, and PLTM-Dev) described in Chapter 3. To determine the topic count, we use `test2006`[4] so as not to exclude any data from the original training set for surrogate selection.

2. Compute avaraged BLEU scores of five runs with the selected training sets.

3. Conduct the same experiments with the cross-entropy approach described above.

4. Conduct the same experiments with randomly selected data and the entire dataset as unadapted versions.

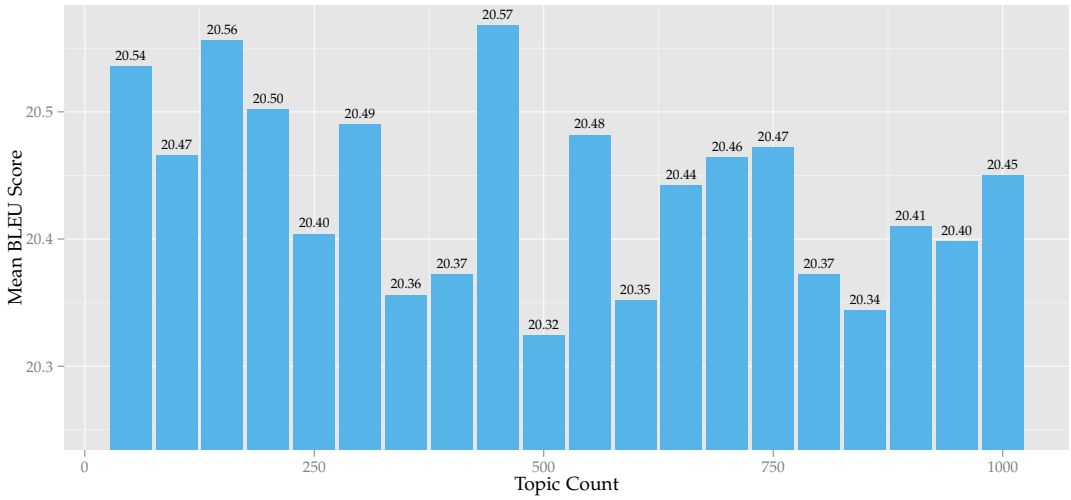
Again, we use the Europarl corpus for the experiments. However we conduct the experiments with the EN-FR and EN-DE pairs in both translation directions (EN $\leftrightarrow$ FR and EN $\leftrightarrow$ DE).

5.5.3 Results

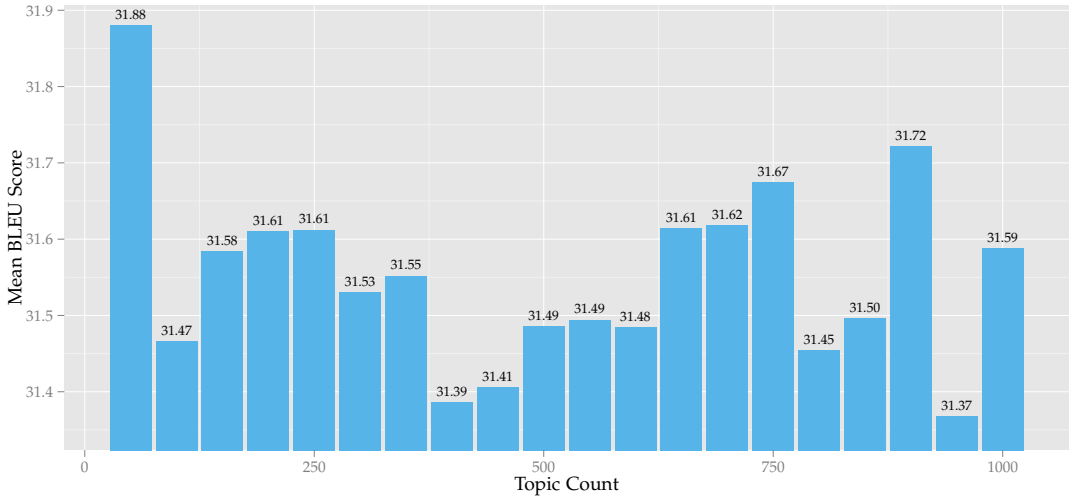
Figure 5.11 shows the BLEU scores with different topic counts (50 to 1,000) evaluated on `test2006` in order to identify the topic count that yields the highest BLEU score. For this process, we fix the TU count for selected training set to 300,000 as in the other experiments. Because of the intensity of computation, we examine the EN $\rightarrow$ DE and EN $\rightarrow$ FR directions only and apply the identified topic counts to the DE $\rightarrow$ EN and FR $\rightarrow$ EN based on the EN $\rightarrow$ DE and EN $\rightarrow$ FR results, respectively, based on the assumption that the identified topic counts are applicable to both translation directions. According to the peaks in Figure 5.11, we choose 450 topics for EN-DE/DE-EN and 50 topics for EN-FR/FR-EN.

Table 5.1 shows the averaged BLEU scores for each of the four language pairs. As shown, the three topic adaptation results on EN-FR outperform those of other approaches significantly (approx. 3 BLEU points) for all the data sizes. For FR-EN, the topic adaptation results are not able to outperform those of the cross-entropy (Xent) approach. For EN-DE, none of the topic adaptation approaches work better than the others. For DE-EN, only PLTM-Dev outperforms Xent at 10%, 30%, and 40%. For both EN-DE and DE-EN, 100% attains the best performance among all.

[4] Available at `http://www.statmt.org/wmt07/shared-task.html` as well.



(a) EN-DE



(b) EN-FR

Figure 5.11: Topic Count Sweeping with `test2006`

The reason that the topic adaptation approaches outperform the others with the EN-FR pair is possibly because we have conducted a series of experiments to examine various components of the topic adaptation approaches, such as distance metrics, surrogate selection methods and so forth, using this language pair, and we have formulated the topic adaptation approaches based on the results. Then we have applied the same formulation of the topic adaptation approaches uniformly to all the other language pairs in this experiment. To obtain more desirable results with other language

Lang. Pair	Method	Data Subset Sizes			
		10%	20%	30%	40%
EN-FR	Rand	25.95	26.22	26.37	26.53
	Xent	26.11	26.47	26.52	26.58
	User-Dev	29.60	30.13	**30.64**	30.57
	LDA-Dev	**29.79**	30.16	30.26	30.62
	PLTM-Dev	**29.79**	**30.91**	30.54	**30.89**
	100%	26.47			
		10%	20%	30%	40%
FR-EN	Rand	28.42	28.94	29.16	29.17
	Xent	31.64	**32.23**	**32.47**	**32.66**
	User-Dev	**31.66**	32.04	32.26	**32.66**
	LDA-Dev	31.47	32.08	32.31	32.58
	PLTM-Dev	31.29	32.09	32.29	32.52
	100%	32.61			
		10%	20%	30%	40%
EN-DE	Rand	20.13	20.59	20.84	21.02
	Xent	**20.42**	**20.85**	**21.03**	**21.19**
	User-Dev	20.13	20.60	20.82	20.99
	LDA-Dev	20.23	20.55	20.92	20.93
	PLTM-Dev	20.08	20.69	20.85	21.01
	100%	21.32			
		10%	20%	30%	40%
DE-EN	Rand	26.43	**27.16**	27.39	27.49
	Xent	26.29	27.15	27.54	27.62
	User-Dev	26.67	26.67	26.96	27.29
	LDA-Dev	**26.98**	26.98	27.45	27.54
	PLTM-Dev	26.90	26.90	**27.76**	**27.65**
	100%	28.14			

Table 5.1: Averaged BLEU scores on the Europarl translation task with `test2007` for random (Rand), cross-entropy (Xent), User-Dev, LDA-Dev, and PLTM-Dev. 100% = system using all of the training data. The bold-faced numbers indicate the best BLEU scores in the data subset sizes.

pairs, we probably need to go through the same formulation process in order to customize the components of the topic adaptation approaches for those language pairs.

5.6 LDS Data Experiment

Lastly, we conduct experiments with the LDS dataset, which is introduced in Section 1.4. As mentioned, this dataset is generated from the translation memory (TM) data collected from a

wide variety of texts and their translations for publication purposes. Also, this dataset is used for the training of SMT systems, which are incorporated as part of their in-house computer-assisted translation framework. We conduct this experiment as an example of the application of topic adaptation to a real-world dataset. We focus on four different language pairs in this section and examine the topic adaptation approaches through the comparison with other approaches as we conducted in the previous section.

5.6.1 Experiment Setup

We use the English-Spanish (EN-ES), English-Japanese (EN-JA), EN-FR and EN-DE TMs extracted from the data storage server in November 2014 for this experiment. This server is used at the LDS Church to store the original English documents and the translations generated by human translators using their in-house computer-assisted translation (CAT) system, which supports large-scale translation processes (worth 85 million words in a year). For data preparation, we clean and extract TUs from the TMX[5] files using Okapi,[6] a bilingual data processing framework. For Japanese sentence tokenization, we use MeCab[7], a widely-used, open-source morphological analyzer for Japanese. Unlike Europarl, the order of TUs is not necessarily retained properly in the database. Therefore, neighboring TUs may be generated from totally unrelated documents. Therefore, the simple corpus segmentation approach used for Europarl is not directly applicable. To group TUs before creating segments, we use item IDs attached to each TU as metadata. An item ID is assigned to each translation project when it is initiated by a human translation team. A translation project can be composed of articles in an issue of a magazine, a webpage, subtitles of a video clip, sections of an instruction manual, and so on, depending on the type and scale of assigned materials to be translated. First we create segments according to the assigned item IDs, assuming the TUs bundled by a specific item ID are topically related. Table 5.2a summarizes the spread of TU counts in the created TU groups based on the item IDs. As shown, the distributions of the TU counts are very

[5] An XML specification for TM data. See `http://www.gala-global.org/lisa-oscar-standards`.
[6] `http://okapi.opentag.com/`
[7] `http://mecab.sourceforge.net`

Lang. Pair	#TU / Item ID	
EN-ES	Max	40,964
	Min	1
	Mean	226
EN-FR	Max	36,139
	Min	1
	Mean	217
EN-DE	Max	39,328
	Min	1
	Mean	198
EN-JA	Max	27,338
	Min	1
	Mean	193

(a) Original TU Count per Item ID in Dataset

Lang. Pair	Dataset	# TU	# Seg.
EN-ES	Train	1,708,186	37,943
	Dev	2,259	50
	Test	2,298	50
EN-FR	Train	1,445,004	31,946
	Dev	2,213	50
	Test	2,347	50
EN-DE	Train	1,155,235	25,632
	Dev	2,309	50
	Test	2,223	50
EN-JA	Train	1,137,728	25,143
	Dev	2,140	50
	Test	2,167	50

(b) Final Number of TU and Segment Counts

Table 5.2: Summary of LDS Dataset

skewed with this TU-grouping approach. To rectify the skewness in the TU counts, we simply divide the large segments into 50 TUs each, as we did for Europarl. Finally, we randomly select 100 segments to use 50 as the test set and the remaining 50 as the user-provided development set for User-Dev in Figure 3.2. The final data sizes are shown in Table 5.2b. Other experiment settings are the same as in the previous section, except that we sweep from 10% to 50% of the data size. For training selection, we uniformly set the topic count to 200 to avoid the heavy computation associated with the topic count sweeping for these four language pairs.

5.6.2 Results

Figure 5.12 shows the SMT performance with the three topic adaptation approaches along with the baselines. The result of the unadapted baseline `all` (i.e., trained on the entire dataset) is repeatedly shown in all data sizes with the other results for comparison. As shown, User-Dev, LDA-Dev and PLTM-Dev significantly outperform Random for all data sizes and language pairs (6 to 16 BLEU points). This comparison clearly indicates that the TUs contained in the LDS dataset do not contribute equally to the SMT performance with the specific test set and that the data selection

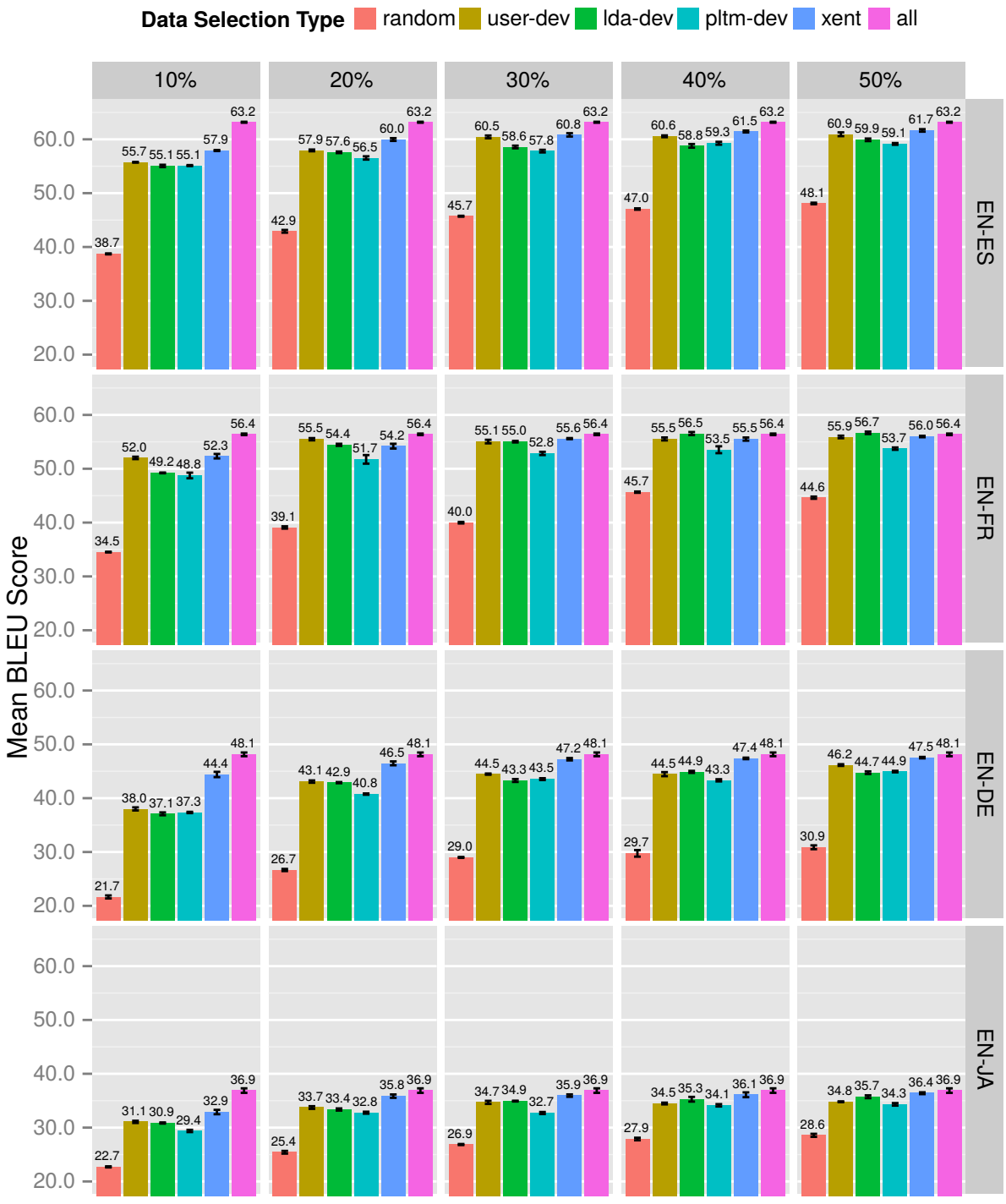


Figure 5.12: LDS Dataset Results

approaches with topic adaptation effectively captured segments containing TUs that enhance the SMT performance for this dataset.

For EN-FR, LDA-Dev outperforms both Xent and `all` with 40% and 50%. The BLEU score difference between LDA-Dev of 50% and `all` is statistically significant at $p < 0.05$ and between LDA-Dev of 50% and Xent at $p < 0.01$ based on the paired bootstrap resampling analysis [54].[8] For other language pairs, the topic adaptation approaches underperform Xent and/or `all`, as seen in the Europarl experiment of the previous section.

5.7 Discussion

In this chapter, we examined the effectiveness of topic adaptation in the realistic scenario. As seen in the experiment results in the sections above, we cannot conclude that the topic adaptation approaches are consistently effective in terms of outperforming the cross-entropy and unadapted approaches. However, we have observed that the SMT performance for EN-FR in the Europarl and LDS datasets is superior to other baselines. This possibly indicates that various parameters to be estimated such as segment sizes and topic counts should be found specifically for a language pair and/or corpus. Such careful parameter estimation was avoided because of the intensiveness of the required computation in some of the experiments in this chapter. We will re-examine the effectiveness of per-language pair topic adaptation with careful parameter estimation in future work.

[8] The script is available at `http://www.ark.cs.cmu.edu/MT/`.

Chapter 6

Conclusions

6.1 Contributions of This Work

In this thesis, we have presented three data selection methods using topic models for better SMT performance and examined their effectiveness in the idealized scenarios, in which the target side of the test set is visible in the data selection process, and realistic scenario, in which no target language information in the test set is available. The following list summarizes the contributions of this work discussed in this thesis.

Data selection approaches that are applicable to any SMT methods. We have introduced three data selection approaches (User-Dev, LDA-Dev, and PLTM-Dev) in Chapter 3 and investigated their effectiveness in Chapters 4 and 5. As depicted in Figure 3.2, all of these three approaches are accomplished independently of the subsequent SMT training process. Although we have focused on the effectiveness of the approaches in the experiments only with the standard phrase-based translation model training, these approaches are easily applied to any type of translation models, such as hierarchical phrase-based models [18, 67] and syntax-based models [83, 102]. Such flexibility is not necessarily available to data weighting approaches, because they need to be incorporated directly into the SMT training framework and should be implemented for a specific training method.

Topical training data selection using topic models and clustering algorithms. We have presented the data clustering process based on topical similarities between training and test data in Chapter 3 . Since topic models do not cluster documents provided for inference of hidden

topics, an additional process for document clustering is required. We have proposed a two-step process to cluster segments with the HAC algorithm after discovering topic distributions with monolingual and bilingual topic models, using JS divergence and cosine distance as similarity metrics.

Using a surrogate set for rigorous experimentation. In this study, we have used a surrogate set for two purposes. One is using a surrogate set for topic count sweeping. In this case, we used it as a test set in the SMT process to identify the best topic count based on the BLEU scores evaluated on the surrogate set, and the identified topic count is used for training selection and the final BLEU score is reported with the blind set. The other is using a surrogate set as a substitute for the blind set for the PLTM-Dev approach. In both cases, the surrogate set enables us to use parallel data which are similar to the target task for setting a topic count and using a bilingual topic model without looking at the target side of the blind data in the training selection process.

An investigation of the impact of topical training data on SMT performance. In Chapters 4 and 5 we have examined the effectiveness of the data selection approaches in the idealized and realistic scenarios. In the idealized scenario, we have shown that topical training sets consistently outperform unadapted training sets in the topic count sweeping and learning curve experiments of Section 4.4. These results constitute an upper bound showing what is possible. In the realistic scenario experiments, we have observed that topic adaptation approaches outperform the cross-entropy approach and unadapted (100%) SMT systems with EN-FR Europarl and LDS datasets, although such results are not observed in the other language pairs. These results indicate that training data, if chosen properly from a topically diverse dataset, will yield superior translation results even when the size of the topical training set is smaller than the entire dataset.

6.2 Limitations of This Work

The main drawback of these data selection approaches is the intensive computation associated with the training selection and SMT processes. Our data selection methods rely on the HAC algorithm, the complexity of which is $O(n^3)$. This makes the clustering process significantly slow even in a high-performance computing environment, when a large training set is provided. Also, the SMT training and tuning processes require significant amounts of time if training and tuning datasets are large. Because of this constraint, we did not conduct topic count sweeping in some of the realistic scenario experiments. For future work, we need to investigate a computationally efficient approach to search for optimal topic counts without relying on computationally heavy processes.

6.3 Future Work

There are several directions to extend the data selection methods discussed in this thesis. Among others, we can further our investigation of document segmentation granularity discussed in Chapter 4. In our experiments, we consistently use 50 TUs in a segment based on the JS divergence experiment in Section 4.2. However, this segmentation approach is not necessarily applicable to all the parallel corpora and/or language pairs. Several previous studies investigate various approaches to segment documents based on topical coherence in the monolingual setting (e.g., [8, 24, 30, 44, 88]). These methods can be applied using the source side of parallel data prior to our current data selection approaches.

Also, identifying topics at the TU level rather than the segment level is another research direction. Co-clustering models [37, 87] are good candidates for this direction because these models can assign topics to sentences as well as words, which is desirable for SMT due to the fact that SMT requires TUs as the atomic unit of training data (see also [41, 98]). These co-clustering models can be extended to accommodate bilingual data in a manner similar to PLTM as a multilingual extension of LDA.

The purpose of using topic models for data selection is creating vector representations of training data. As shown in the TNPD experiment of Section 5.4, this purpose can be accomplished with other types of feature extraction approaches. For example, the paragraph vector [63] is one possible candidate. The paragraph vector is an extension of the word vector representations extracted with neural networks [72]. This approach extracts features of documents or sentences and maps them to a continuous vector space. Using paragraph vector representations of training segments or TUs in lieu of per-document topic distributions is also promising avenue of future work.

Appendix A

Derivation of the PLTM Complete Conditional Distribution for Gibbs Sampling

This appendix shows the derivation of complete conditional distributions for Gibbs sampling inference in the PLTM, the bilingual topic model used in this thesis, based on [73] and [94]. Although PLTM can handle any number of languages, the following derivation focuses on the bilingual case because SMT deals only with two languages. However, the derivation can be modified for the multilingual case by expanding the upper limit of the language count l from 2 to L (i.e., an arbitrary number of language counts). Also, the LDA complete conditional can be derived by confining the language count l to 1.

Let $\boldsymbol{w}^{(l)}$ and $\boldsymbol{z}^{(l)}$ be the word tokens and word topics in the language l respectively. The joint distribution $p\left(\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ \boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)} | \alpha, \beta^{(1)}, \beta^{(2)}\right)$ is expressed as follows:

$$p\left(\boldsymbol{w}^{(1)}, \ \boldsymbol{w}^{(2)}, \boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)} \middle| \alpha, \beta^{(1)}, \beta^{(2)}\right) = \tag{A.1}$$

$$p\left(\boldsymbol{z}^{(1)} \middle| \alpha\right) p\left(\boldsymbol{z}^{(2)} \middle| \alpha\right) p\left(\boldsymbol{w}^{(1)} \middle| \boldsymbol{z}^{(1)}, \beta^{(1)}\right) p\left(\boldsymbol{w}^{(2)} \middle| \boldsymbol{z}^{(2)}, \beta^{(2)}\right).$$

By the independence assumptions of the model,

$$= \int p\left(\boldsymbol{z}^{(1)}\middle|\,\Theta\right) p\left(\boldsymbol{z}^{(2)}\middle|\,\Theta\right) p\left(\Theta\middle|\,\alpha\right) d\Theta \int p\left(\boldsymbol{w}^{(1)}\middle|\,\boldsymbol{z}^{(1)}, \Phi^{(1)}\right) p\left(\Phi^{(1)}\middle|\,\beta^{(1)}\right) d\Phi^{(1)}$$
$$\times \int p\left(\boldsymbol{w}^{(2)}\middle|\,\boldsymbol{z}^{(2)}, \Phi^{(1)}\right) p\left(\Phi^{(2)}\middle|\,\beta^{(2)}\right) d\Phi^{(2)}, \tag{A.2}$$

where Θ is a vector of $\boldsymbol{\theta}$ vectors and $\Phi^{(l)}$ is a $K \times V$ matrix (K: topic count; V: vocabulary size). To iterate over all the possible values of the random variables and given constants $D, K, \boldsymbol{N}^{(1)}, \boldsymbol{N}^{(2)}$ ($\boldsymbol{N}^{(l)}$ are a vector of word token counts in each document d),

$$= \int \prod_{l=1}^{2} \prod_{d=1}^{D} \prod_{n=1}^{N_d^{(l)}} p\left(z_{d,n}^{(l)}\middle|\,\boldsymbol{\theta}_d\right) p\left(\boldsymbol{\theta}_d\middle|\,\alpha\right) d\boldsymbol{\theta}_d$$
$$\times \int \prod_{d=1}^{D} \prod_{n=1}^{N_d^{(1)}} p\left(w_{d,n}^{(1)}\middle|\,z_{d,n}^{(1)}, \phi_{z_{d,n}^{(1)},n}^{(1)}\right) \prod_{k=1}^{K} p\left(\boldsymbol{\phi}_k^{(1)}\middle|\,\beta^{(1)}\right) d\boldsymbol{\phi}_k^{(1)} \tag{A.3}$$
$$\times \int \prod_{d=1}^{D} \prod_{n=1}^{N_d^{(2)}} p\left(w_{d,n}^{(2)}\middle|\,z_{d,n}^{(2)}, \phi_{z_{d,n}^{(2)},n}^{(2)}\right) \prod_{k=1}^{K} p\left(\boldsymbol{\phi}_k^{(2)}\middle|\,\beta^{(2)}\right) d\boldsymbol{\phi}_k^{(2)}.$$

Since $p\left(z_{d,n}^{(l)}\middle|\,\boldsymbol{\theta}_d\right) = \theta_{d,z_{d,n}^{(l)}}$, and $p\left(w_{d,n}^{(l)}\middle|\,z_{d,n}^{(l)}, \phi_{z_{d,n}^{(l)},n}^{(l)}\right) = \phi_{z_{d,n}^{(l)},w_{d,n}^{(l)}}^{(l)}$,

$$= \int \prod_{l=1}^{2} \prod_{d=1}^{D} \prod_{n=1}^{N_d^{(l)}} \theta_{d,z_{d,n}^{(l)}} p\left(\boldsymbol{\theta}_d\middle|\,\alpha\right) d\boldsymbol{\theta}_d$$
$$\times \int \prod_{d=1}^{D} \prod_{n=1}^{N_d^{(1)}} \phi_{z_{d,n}^{(1)},w_{d,n}^{(1)}}^{(1)} \prod_{k=1}^{K} p\left(\boldsymbol{\phi}_k^{(1)}\middle|\,\beta^{(1)}\right) d\boldsymbol{\phi}_k^{(1)} \tag{A.4}$$
$$\times \int \prod_{d=1}^{D} \prod_{n=1}^{N_d^{(2)}} \phi_{z_{d,n}^{(2)},w_{d,n}^{(2)}}^{(2)} \prod_{k=1}^{K} p\left(\boldsymbol{\phi}_k^{(2)}\middle|\,\beta^{(2)}\right) d\boldsymbol{\phi}_k^{(2)}.$$

Because all the $\boldsymbol{\theta}$ vectors are independent of each other and all the $\boldsymbol{\phi}_k^{(l)}$ vectors are also independent of each other,

$$
\begin{aligned}
&= \prod_{d=1}^{D} \int \prod_{l=1}^{2} \prod_{n=1}^{N_d^{(l)}} \theta_{d,z_{d,n}^{(l)}} \, p\left(\boldsymbol{\theta}_d \middle| \, \alpha\right) d\boldsymbol{\theta}_d \\
&\qquad \times \prod_{k=1}^{K} \int p\left(\boldsymbol{\phi}_k^{(1)} \middle| \, \beta^{(1)}\right) \prod_{d=1}^{D} \prod_{n=1}^{N_d^{(1)}} \phi_{z_{d,n}^{(1)}, w_{d,n}^{(1)}} \, d\boldsymbol{\phi}_k^{(1)} \\
&\qquad \times \prod_{k=1}^{K} \int p\left(\boldsymbol{\phi}_k^{(2)} \middle| \, \beta^{(2)}\right) \prod_{d=1}^{D} \prod_{n=1}^{N_d^{(2)}} \phi_{z_{d,n}^{(2)}, w_{d,n}^{(2)}} \, d\boldsymbol{\phi}_k^{(2)}.
\end{aligned}
\tag{A.5}
$$

Since $\boldsymbol{\theta}_d | \alpha \sim \text{Dir}(\alpha)$ and $\boldsymbol{\phi}_k^{(l)} | \beta^{(l)} \sim \text{Dir}\left(\beta^{(l)}\right)$,

$$
\begin{aligned}
&= \prod_{d=1}^{D} \int \frac{\Gamma\left(\sum_{k=1}^{K} \alpha\right)}{\prod_{k=1}^{K} \Gamma(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha-1} \prod_{l=1}^{2} \prod_{n=1}^{N_d^{(l)}} \theta_{d,z_{d,n}^{(l)}} \, d\boldsymbol{\theta}_d \\
&\qquad \times \prod_{k=1}^{K} \int \frac{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)}\right)}{\prod_{v=1}^{V^{(1)}} \Gamma(\beta^{(1)})} \prod_{v=1}^{V^{(1)}} \phi_{k,v}^{\beta^{(1)}-1} \prod_{d=1}^{D} \prod_{n=1}^{N_d^{(1)}} \phi_{z_{d,n}^{(1)}, w_{d,n}^{(1)}} \, d\boldsymbol{\phi}_k^{(1)} \\
&\qquad \times \prod_{k=1}^{K} \int \frac{\Gamma\left(\sum_{v=1}^{V^{(2)}} \beta^{(2)}\right)}{\prod_{v=1}^{V^{(2)}} \Gamma(\beta^{(2)})} \prod_{v=1}^{V^{(2)}} \phi_{k,v}^{\beta^{(2)}-1} \prod_{d=1}^{D} \prod_{n=1}^{N_d^{(2)}} \phi_{z_{d,n}^{(2)}, w_{d,n}^{(2)}} \, d\boldsymbol{\phi}_k^{(2)}.
\end{aligned}
\tag{A.6}
$$

Let $z_{d,n}^{(l)}$ and $w_{d,n}^{(l)}$ be k and v respectively :

$$
\begin{aligned}
&= \prod_{d=1}^{D} \int \frac{\Gamma\left(\sum_{k=1}^{K} \alpha\right)}{\prod_{k=1}^{K} \Gamma(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha-1} \prod_{k=1}^{K} \theta_{d,k}^{c_{d,k}^{(1)} + c_{d,k}^{(2)}} \, d\boldsymbol{\theta}_d \\
&\qquad \times \prod_{k=1}^{K} \int \frac{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)}\right)}{\prod_{v=1}^{V^{(1)}} \Gamma(\beta^{(1)})} \prod_{v=1}^{V^{(1)}} \phi_{k,v}^{\beta^{(1)}-1} \prod_{v=1}^{V^{(1)}} \phi_{k,v}^{t_{k,v}^{(1)}} \, d\boldsymbol{\phi}_k^{(1)} \\
&\qquad \times \prod_{k=1}^{K} \int \frac{\Gamma\left(\sum_{v=1}^{V^{(2)}} \beta^{(2)}\right)}{\prod_{v=1}^{V^{(2)}} \Gamma(\beta^{(2)})} \prod_{v=1}^{V^{(2)}} \phi_{k,v}^{\beta^{(2)}-1} \prod_{v=1}^{V^{(2)}} \phi_{k,v}^{t_{k,v}^{(2)}} \, d\boldsymbol{\phi}_k^{(2)},
\end{aligned}
\tag{A.7}
$$

where the counter variable $c_{d,k}^{(l)}$ indicates the number of times that topic with index k has been generated from the multinomial distribution specific to document $d^{(l)}$. $t_{k,v}^{(l)}$ is another counter which counts the number of times the word v has been sampled by topic k.

Then simplify the products:

$$
\begin{aligned}
= \prod_{d=1}^{D} \int \frac{\Gamma\left(\sum_{k=1}^{K} \alpha\right)}{\prod_{k=1}^{K} \Gamma(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)} - 1} d\boldsymbol{\theta}_d \\
\times \prod_{k=1}^{K} \int \frac{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)}\right)}{\prod_{v=1}^{V^{(1)}} \Gamma(\beta^{(1)})} \prod_{v=1}^{V^{(1)}} \phi_{k,v}^{\beta^{(1)} + t_{k,v}^{(1)} - 1} d\boldsymbol{\phi}_k^{(1)} \\
\times \prod_{k=1}^{K} \int \frac{\Gamma\left(\sum_{v=1}^{V^{(2)}} \beta^{(2)}\right)}{\prod_{v=1}^{V^{(2)}} \Gamma(\beta^{(2)})} \prod_{v=1}^{V^{(2)}} \phi_{k,v}^{\beta^{(2)} + t_{k,v}^{(2)} - 1} d\boldsymbol{\phi}_k^{(2)}.
\end{aligned}
\tag{A.8}
$$

Next, multiply by constants to integrate to one:

$$
\begin{aligned}
&= \prod_{d=1}^{D} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha\right)}{\prod_{k=1}^{K} \Gamma(\alpha)} \frac{\prod_{k=1}^{K} \Gamma\left(\alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)}\right)} \\
&\quad \times \underbrace{\int \frac{\Gamma\left(\sum_{k=1}^{K} \alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)}\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)}\right)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)} - 1} d\boldsymbol{\theta}_d}_{=1} \\
&\quad \times \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)}\right)}{\prod_{v=1}^{V^{(1)}} \Gamma\left(\beta^{(1)}\right)} \frac{\prod_{v=1}^{V^{(1)}} \Gamma\left(\beta^{(1)} + t_{k,v}^{(1)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t_{k,v}^{(1)}\right)} \\
&\quad \times \underbrace{\int \frac{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t_{k,v}^{(1)}\right)}{\prod_{v=1}^{V^{(1)}} \Gamma\left(\beta^{(1)} + t_{k,v}^{(1)}\right)} \prod_{v=1}^{V^{(1)}} \phi_{k,v}^{\beta^{(1)} + t_{k,v}^{(1)} - 1} d\boldsymbol{\phi}_k^{(1)}}_{=1} \\
&\quad \times \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V^{(2)}} \beta^{(2)}\right)}{\prod_{v=1}^{V^{(2)}} \Gamma\left(\beta^{(2)}\right)} \frac{\prod_{v=1}^{V^{(2)}} \Gamma\left(\beta^{(2)} + t_{k,v}^{(2)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(2)}} \beta^{(2)} + t_{k,v}^{(2)}\right)} \\
&\quad \times \underbrace{\int \frac{\Gamma\left(\sum_{v=1}^{V^{(2)}} \beta^{(2)} + t_{k,v}^{(2)}\right)}{\prod_{v=1}^{V^{(2)}} \Gamma\left(\beta^{(2)} + t_{k,v}^{(2)}\right)} \prod_{v=1}^{V^{(2)}} \phi_{k,v}^{\beta^{(2)} + t_{k,v}^{(2)} - 1} d\boldsymbol{\phi}_k^{(2)}}_{=1}.
\end{aligned}
\tag{A.9}
$$

Next, drop the constants which only depend on the hyperparameters α and β:

$$
\propto \prod_{d=1}^{D} \frac{\prod_{k=1}^{K} \Gamma\left(\alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)}\right)} \prod_{k=1}^{K} \frac{\prod_{v=1}^{V^{(1)}} \Gamma\left(\beta^{(1)} + t_{k,v}^{(1)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t_{k,v}^{(1)}\right)} \prod_{k=1}^{K} \frac{\prod_{v=1}^{V^{(2)}} \Gamma\left(\beta^{(2)} + t_{k,v}^{(2)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(2)}} \beta^{(2)} + t_{k,v}^{(2)}\right)}.
\tag{A.10}
$$

To simplify the derivation, we focus on the langage 1. In this case, the third term in Equation A.10 can be dropped:

$$\propto \prod_{d=1}^{D} \frac{\prod_{k=1}^{K} \Gamma\left(\alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)}\right)} \prod_{k=1}^{K} \frac{\prod_{v=1}^{V^{(1)}} \Gamma\left(\beta^{(1)} + t_{k,v}^{(1)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t_{k,v}^{(1)}\right)}. \tag{A.11}$$

Next, split the product to pull out the terms dependent on the current sample position $d = a$ and $n = b$:

$$\begin{aligned} \propto \prod_{d \neq a} \frac{\prod_{k=1}^{K} \Gamma\left(\alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha + c_{d,k}^{(1)} + c_{d,k}^{(2)}\right)} \frac{\prod_{k=1}^{K} \Gamma(\alpha + c_{a,k}^{(1)} + c_{d,k}^{(2)})}{\Gamma\left(\sum_{k=1}^{K} \alpha + c_{a,k}^{(1)} + c_{d,k}^{(2)}\right)} \\ \times \prod_{k=1}^{K} \frac{\prod_{v \neq w_{a,b}^{(1)}} \Gamma\left(\beta^{(1)} + t_{k,v}^{(1)}\right) \times \Gamma\left(\beta^{(1)} + t_{k,w_{a,b}^{(1)}}^{(1)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t_{k,v}^{(1)}\right)}. \end{aligned} \tag{A.12}$$

Then drop terms that do not depend on (a, b):

$$= \frac{\prod_{k=1}^{K} \Gamma\left(\alpha + c_{a,k}^{(1)} + c_{d,k}^{(2)}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha + c_{a,k}^{(1)} + c_{d,k}^{(2)}\right)} \prod_{k=1}^{K} \frac{\Gamma\left(\beta^{(1)} + t_{k,w_{a,b}^{(1)}}^{(1)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t_{k,v}^{(1)}\right)}. \tag{A.13}$$

Now we compute the complete conditional probability $\left[z_{a,b}^{(1)}\right]$. We separate the count without the topic at the position (a, b) and the one with it at the position (a, b). Therefore,

$$\begin{aligned} \propto \frac{\prod_{k \neq z_{a,b}^{(1)}} \Gamma\left(\alpha + c_{a,k,\neg b}^{(1)} + c_{d,k}^{(2)}\right) \times \Gamma\left(\alpha + c_{a,z_{a,b}^{(1)},\neg b}^{(1)} + c_{d,k}^{(2)} + 1\right)}{\Gamma\left(1 + \sum_{k=1}^{K} \alpha + c_{a,k,\neg b}^{(1)} + c_{d,k}^{(2)}\right)} \\ \times \prod_{k \neq z_{a,b}^{(1)}} \frac{\Gamma\left(\beta^{(1)} + t_{k,w_{a,b}^{(1)},\neg}^{(1)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t_{k,v}^{(1)}\right)} \times \frac{\Gamma\left(\beta^{(1)} + t_{z_{a,b}^{(1)},w_{a,b}^{(1)},\neg}^{(1)} + 1\right)}{\Gamma\left(1 + \sum_{v=1}^{V^{(1)}} \beta^{(1)} + t_{z_{a,b}^{(1)},v,\neg}^{(1)}\right)}, \end{aligned} \tag{A.14}$$

where $c^{(l)}_{a,k,\neg b}$ indicates the number of times that topic with index k has been generated from the multinomial distribution specific to document a but the current topic k is excluded from the count. $t^{(l)}_{z^{(l)}_{a,b},w^{(l)}_{a,b},\neg}$ indicates the number of times $w^{(l)}_{a,b}$ has been sampled by topic k, but not counting the $w^{(l)}_{d,n}$ (i.e., $t^{(l)}_{k,w^{(l)}_{d,n}} - 1$).

Expand the Gamma terms using the property $\Gamma(n+1) = n\Gamma(n)$:

$$= \frac{\prod_{k \neq z^{(1)}_{a,b}} \Gamma\left(\alpha + c^{(1)}_{a,k,\neg b} + c^{(2)}_{d,k}\right) \times \Gamma\left(\alpha + c^{(1)}_{a,z^{(1)}_{a,b},\neg b} + c^{(2)}_{d,k}\right) \times \left(\alpha + c^{(1)}_{a,z^{(1)}_{a,b},\neg b} + c^{(2)}_{d,k}\right)}{\Gamma\left(1 + \sum_{k=1}^{K} \alpha + c^{(1)}_{a,k,\neg n} + c^{(2)}_{d,k}\right)}$$
$$\times \prod_{k \neq z^{(1)}_{a,b}} \frac{\Gamma\left(\beta^{(1)} + t^{(1)}_{k,w^{(1)}_{a,b},\neg}\right)}{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t^{(1)}_{k,v}\right)} \times \frac{\Gamma\left(\beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},w^{(1)}_{a,b},\neg}\right) \times \left(\beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},w^{(1)}_{a,b},\neg}\right)}{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},v,\neg}\right) \times \left(\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},v,\neg}\right)}. \tag{A.15}$$

Then merge the Γ terms back to the products:

$$= \frac{\prod_{k=1}^{K} \Gamma\left(\alpha + c^{(1)}_{a,k,\neg b} + c^{(2)}_{d,k}\right) \times \left(\alpha + c^{(1)}_{a,z^{(1)}_{a,b},\neg b} + c^{(2)}_{d,k}\right)}{\Gamma\left(1 + \sum_{k=1}^{K} \alpha + c^{(1)}_{a,k,\neg b} + c^{(2)}_{d,k}\right)}$$
$$\times \prod_{k=1}^{K} \frac{\Gamma\left(\beta^{(1)} + t^{(1)}_{k,w^{(1)}_{a,b},\neg}\right)}{\Gamma\left(\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t^{(1)}_{k,v}\right)} \times \frac{\beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},w^{(1)}_{a,b},\neg}}{\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},v,\neg}}. \tag{A.16}$$

Eliminate the products because they are constants:

$$= \frac{\alpha + c^{(1)}_{a,z^{(1)}_{a,b},\neg b} + c^{(2)}_{d,k}}{\Gamma\left(1 + \sum_{k=1}^{K} \alpha + c^{(1)}_{a,k,\neg b} + c^{(2)}_{d,k}\right)} \times \frac{\beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},w^{(1)}_{a,b}},\neg}{\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},v,\neg}}. \tag{A.17}$$

Expand the remaining Γ term using $\Gamma(n+1) = n\Gamma(n)$ and drop the Γ term because it is constant:

$$= \frac{\alpha + c^{(1)}_{a,z^{(1)}_{a,b},\neg b} + c^{(2)}_{d,k}}{\Gamma\left(\sum_{k=1}^{K} \alpha + c^{(1)}_{a,k,\neg b} + c^{(2)}_{d,k}\right)\left(\sum_{k=1}^{K} \alpha + c^{(1)}_{a,k,\neg n} + c^{(2)}_{d,k}\right)} \times \frac{\beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},w^{(1)}_{a,b},\neg}}{\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},v,\neg}} \tag{A.18}$$

$$= \frac{\alpha + c^{(1)}_{a,z^{(1)}_{a,b},\neg b} + c^{(2)}_{d,k}}{\sum_{k=1}^{K} \alpha + c^{(1)}_{a,k,\neg b} + c^{(2)}_{d,k}} \times \frac{\beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},w^{(1)}_{a,b},\neg}}{\sum_{v=1}^{V^{(1)}} \beta^{(1)} + t^{(1)}_{z^{(1)}_{a,b},v,\neg}} \tag{A.19}$$

Finally, expand the summations and substitute the general indices k and n for $z^{(1)}_{a,b}$ and $w^{(1)}_{a,b}$ respectively:

$$\left[z^{(1)}_{d,n} = k\right] = \frac{\alpha + c^{(1)}_{d,k,\neg n} + c^{(2)}_{d,k}}{K\alpha + c^{(1)}_{d,\cdot,\neg n} + c^{(2)}_{d,\cdot}} \times \frac{\beta^{(1)} + t^{(1)}_{k,n,\neg}}{V\beta^{(1)} + t^{(1)}_{k,\cdot,\neg}}, \tag{A.20}$$

where the dot $(\cdot)$ denotes summation over all values of the variable whose index the dot takes.

In the same manner, $\left[z^{(2)}_{d,k} = k\right]$ is derived as below:

$$\left[z^{(2)}_{d,k} = k\right] = \frac{\alpha + c^{(2)}_{d,k,\neg n} + c^{(1)}_{d,k}}{K\alpha + c^{(2)}_{d,\cdot,\neg n} + c^{(1)}_{d,\cdot}} \times \frac{\beta^{(2)} + t^{(2)}_{k,n,\neg}}{V\beta^{(2)} + t^{(2)}_{k,\cdot,\neg}}. \tag{A.21}$$

References

[1] Amittai Axelrod. *Data Selection for Statistical Machine Translation*. PhD thesis, University of Washington, 2014.

[2] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, 2011.

[3] Amittai Axelrod, QingJun Li, and William D. Lewis. Applications of data selection via cross-entropy difference for real-world statistical machine translation. In *Proceedings of IWSLT*, pages 201–208, 2012.

[4] Arianna Bisazza, Nick Ruiz, and Marcello Federico. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, 2011.

[5] David Blei. Probabilistic topic models, 2012. URL `http://icml.cc/2012/tutorials/`.

[6] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[7] David M. Blei and John Lafferty. Topic models. *Text mining: Classification, clustering, and applications*, 10:71–93, 2009.

[8] David M Blei and Pedro J Moreno. Topic segmentation with an aspect hidden markov model. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 343–348. ACM, 2001.

[9] David M. Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet Allocation. *JMLR*, 2003.

[10] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *9th Workshop on Statistical Machine Translation*, 2014.

[11] Jordan Boyd-Graber, David Mimno, and David Newman. Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of Mixed Membership Models and Their Applications*, 2014.

[12] Fabienne Braune and Alexander Fraser. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 81–89. Association for Computational Linguistics, 2010.

[13] Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning sentences in parallel corpora. In *Meeting of the Association for Computational Linguistics*, pages 169–176, 1991.

[14] Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *EMNLP-CoNLL*, pages 61–72, 2007.

[15] Marine Carpuat, Hal Daumé III, Alexander Fraser, Chris Quirk, Fabienne Braune, Ann Clifton, Ann Irvine, Jagadeesh Jagarlamudi, John Morgan, Majid Razmara, Aleš Tamchyna, Katharine Henry, and Rachel Rudinger. Domain adaptation in machine translation: Final report. In *2012 Johns Hopkins Summer Workshop Final Report*. 2012. URL `http://hal3.name/damt/`.

[16] Marine Carpuat, Hal Daumé III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. SenseSpotting: Never let your parallel data tie you to an old domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1435–1445, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P13-1141`.

[17] Daniel Cer, Daniel Jurafsky, and Christopher D Manning. Regularization and search for minimum error rate training. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 26–34. Association for Computational Linguistics, 2008.

[18] David Chiang. Hierarchical phrase-based translation. *computational linguistics*, 33(2): 201–228, 2007.

[19] David Chiang. Hope and fear for discriminative training of statistical translation models. *The Journal of Machine Learning Research*, 13(1):1159–1187, 2012.

[20] David Chiang, Steve DeNeefe, and Michael Pust. Two easy improvements to lexical weighting. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 455–460. Association for Computational Linguistics, 2011.

[21] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, 7:551–585, 2006.

[22] Hal Daumé III and Jagadeesh Jagarlamudi. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, 2011.

[23] Yonggang Deng, Shankar Kumar, and William Byrne. Segmentation and alignment of parallel text for statistical machine translation. *Natural Language Engineering*, 13(3):235, 2007.

[24] Lan Du, Wray Buntine, and Mark Johnson. Topic segmentation with a structured topic model. In *Proceedings of NAACL-HLT*, pages 190–200, 2013.

[25] Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Analysis of translation model adaptation in statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT'10), Paris, France*, 2010.

[26] Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. Adaptation data selection using neural language models: Experiments in machine translation. In *ACL (2)*, pages 678–683, 2013.

[27] Matthias Eck, Yura Zemlyanskiy, Joy Zhang, and Alex Waibel. Extracting translation pairs from social network content. In *Proceedings of IWSLT*, 2014.

[28] Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 115–119. Association for Computational Linguistics, 2012.

[29] Andreas Eisele and Yu Chen. MultiUN: A multilingual corpus from united nation documents. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA), 5 2010.

[30] J. Eisenstein and R. Barzilay. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 334–343. Association for Computational Linguistics, 2008.

[31] George Foster and Roland Kuhn. Mixture-model adaptation for SMT. In *Proceedings of Association for the Second ACL Workshop on Statistical Machine Translation*, pages 128–136, 2007.

[32] George Foster and Roland Kuhn. Stabilizing minimum error rate training. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 242–249, 2009.

[33] George Foster, Cyril Goutte, and Roland Kuhn. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459. Association for Computational Linguistics, 2010.

[34] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184, 1991.

[35] Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. Association for Computational Linguistics, 2008.

[36] Kevin Gimpel and Noah A Smith. Structured ramp loss minimization for machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 221–231. Association for Computational Linguistics, 2012.

[37] Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. Hidden topic markov models. In *International Conference on Artificial Intelligence and Statistics*, pages 163–170, 2007.

[38] Barry Haddow and Philipp Koehn. Analysing the effect of out-of-domain data on SMT systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 422–432. Association for Computational Linguistics, 2012.

[39] Eva Hasler. *Dynamic Topic Adaptation for Improved Contextual Modelling in Statistical Machine Translation*. PhD thesis, University of Edinburgh, 2014.

[40] Eva Hasler, Barry Haddow, and Philipp Koehn. Margin infused relaxed algorithm for Moses. *The Prague Bulletin of Mathematical Linguistics*, 96:69–78, 2011.

[41] Eva Hasler, Barry Haddow, and Philipp Koehn. Sparse lexicalised features and topic adaptation for SMT. In *Proceedings of the International Workshop on Spoken Language Translation, Hong Kong, HK*, 2012.

[42] Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. Dynamic topic adaptation for phrase-based MT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. EACL, apr 2014.

[43] Eva Hasler, Barry Haddow, and Philipp Koehn. Dynamic topic adaptation for SMT using distributional profiles. In *Proceedings of WMT*, jul 2014.

[44] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64, 1997.

[45] Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of EAMT*, pages 133–142, 2005.

[46] Yuening Hu, Ke Zhai, Vladimir Edelman, and Jordan Boyd-Graber. Topic models for translation domain adaptation. In *Topic Models: Computation, Application, and Evaluation*. NIPS Workshop, 2013.

[47] Yuening Hu, Ke Zhai, Vladimir Eidelman, and Jordan Boyd-Graber. Polylingual tree-based topic models for translation domain adaptation. In *Proceedings of Association for Computational Linguistics*, 2014.

[48] Ann Irvine. Statistical machine translation in low resource settings. *NAACL HLT*, page 54, 2013.

[49] Ann Irvine and Chris Callison-Burch. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 2013.

[50] Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 2013. URL `https://aclweb.org/anthology/Q/Q13/Q13-1035.pdf`.

[51] Ann Irvine, Chris Quirk, and Hal Daumé III. Monolingual marginal matching for translation model adaptation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[52] Katrin Kirchhoff and Jeff Bilmes. Submodularity for data selection in statistical machine translation. In *Proceedings of EMNLP*, 2014.

[53] Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 130–140. Association for Computational Linguistics, 2012.

[54] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395, 2004.

[55] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, volume 5, 2005.

[56] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2010.

[57] Philipp Koehn and Hieu Hoang. Factored translation models. In *EMNLP-CoNLL*, pages 868–876, 2007.

[58] Philipp Koehn and Kevin Knight. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, pages 9–16. Association for Computational Linguistics, 2002.

[59] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics, 2003.

[60] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[61] Phillipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 224–227, 2007.

[62] Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability: Practical Approaches to Hard Problems*, 3:19, 2012.

[63] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[64] William Lewis and Sauleh Eetemadi. Dramatically reducing training data size through vocabulary saturation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 281–291, 2013.

[65] Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, pages 176–186. Association for Computational Linguistics, 2013.

[66] Le Liu, Yu Hong, Hao Liu, Xing Wang, and Jianmin Yao. Effective selection of translation model training data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 569–573. Association for Computational Linguistics, 2014.

[67] Adam Lopez. Hierarchical phrase-based translation with suffix arrays. In *EMNLP-CoNLL*, pages 976–985, 2007.

[68] Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):8, 2008.

[69] Saab Mansour, Joern Wuebker, and Hermann Ney. Combining translation and language model scoring for domain-specific data filtering. In *Proceedings of IWSLT*, 2011.

[70] Spyros Matsoukas, Antti-Veikko I Rosti, and Bing Zhang. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 708–717. Association for Computational Linguistics, 2009.

[71] Mohammed Mediani, Joshua Winebarger, and Alexander Waibel. Improving in-domain data selection for small in-domain sets. In *Proceedings of IWSLT*, 2014.

[72] Tomas Mikolov, Anoop Deoras, Daniel Povey, Lukas Burget, and Jan Cernocky. Strategies for training large scale neural network language models. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 196–201. IEEE, 2011.

[73] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 14th Conference on Empirical Methods in Natural Language Processing*, pages 880–889, Singapore, 2009.

[74] Shachar Mirkin and Laurent Besacier. Data selection for compact adapted SMT models. In *Proceedings of Conference the Association for Machine Translation in the Americas (AMTA)*, pages 301–314, 2014.

[75] Robert C. Moore. Fast and accurate sentence alignment of bilingual corpora. *Machine Translation: From Research to Real Users*, pages 135–144, 2002.

[76] Robert C. Moore and William D. Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224. Association for Computational Linguistics, 2010.

[77] Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL 2004: Main Proceedings*, pages 265–272, 2004.

[78] Jan Niehues and Alex Waibel. Domain adaptation in statistical machine translation using factored translation models. In *Proceedings of EAMT*, 2010.

[79] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics, 2003.

[80] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[81] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[82] Mārcis Pinnis, Radu Ion, Dan Ştefănescu, Fangzhong Su, Inguna Skadiņa, and Bogdan Babych. Accurat toolkit for multi-level alignment and information extraction from comparable corpora. In *Proceedings of the ACL 2012 System Demonstrations*, pages 91–96. Association for Computational Linguistics, 2012.

[83] Chris Quirk, Arul Menezes, and Colin Cherry. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of ACL*. Association for Computational Linguistics, June 2005. URL `http://research.microsoft.com/apps/pubs/default.aspx?id=69192`.

[84] Chris Quirk, Raghavendra Udupa, and Arul Menezes. Generative models of noisy translations with applications to parallel fragment extraction. *Proceedings of the Machine Translation Summit XI*, pages 377–384, 2007.

[85] Salim Roukos, David Graff, and Dan Melamed. Canadian Hansard French/English. Linguistic Data Consortium (LDC), 1995.

[86] Rico Sennrich. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549. Association for Computational Linguistics, 2012.

[87] M. Mahdi Shafiei and Evangelos E. Milios. Latent Dirichlet co-clustering. In *IEEE International Conference on Data Mining (ICDM)*, Hong Kong, December 2006.

[88] M. Mahdi Shafiei and Evangelos E. Milios. A statistical model for topic segmentation and clustering. *Advances in Artificial Intelligence*, pages 283–295, 2008.

[89] Jason R Smith, Chris Quirk, and Kristina Toutanova. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411. Association for Computational Linguistics, 2010.

[90] Jason R Smith, Philipp Koehn, Herve Saint-Amand, Chris Callison-Burch, Magdalena Plamada, and Adam Lopez. Dirt cheap web-scale parallel text from the Common Crawl. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2013.

[91] Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *Proceedings of INTERSPEECH*, 2002.

[92] Jörg Tiedemann. News from OPUS – A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248, 2009.

[93] Jörg Tiedemann. Bitext alignment. *Synthesis Lectures on Human Language Technologies*, 4 (2):1–165, 2011.

[94] Ivan Vulic, Wim De Smet, Jie Tang, and Marie-Francine Moens. Probabilistic topic modeling in multilingual settings: A short overview of its methodology with applications. In *Proceedings of the NIPS Workshop on Cross-Lingual Technologies (xLiTe)*, pages 1–11, 2012.

[95] Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, 51(1):111–147, 2015.

[96] Daniel David Walker. *Bayesian Test Analytics for Document Collections*. PhD thesis, Brigham Young University, 2012.

[97] Chris Wendt and Federico Garcea. Use case: Customization and collaboration to enhance MT for a knowledge base online portal. *Proceedings of the Machine Translation Summit XIV*, pages 401–405, 2013.

[98] Deyi Xiong and Min Zhang. A topic-based coherence model for statistical machine translation. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13), Bellevue, Washington, USA, July*, 2013.

[99] Deyi Xiong and Min Zhang. A sense-based translation model for statistical machine translation. In *Proceedings of ACL*, pages 1459–1469, 2014.

[100] Deyi Xiong and Min Zhang. *Linguistically Motivated Statistical Machine Translation*. Springer, 2015.

[101] Jia Xu, Rirchard Zens, and Hermann Ney. Sentence segmentation using IBM word alignment model 1. In *Proceedings of EAMT 2005 (10th Annual Conference of the European Association for Machine Translation)*, pages 280–287, 2005.

[102] Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics, 2001.

[103] Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. Method of selecting training data to build a compact and efficient translation model. In *IJCNLP*, pages 655–660, 2008.