Computer Speech & Language 65 (2021) 101138

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Part-of-speech tagging for Arabic tweets using CRF and Bi-LSTM

Wasan AlKhwiter, Nora Al-Twairesh*

College of Computer and Information Sciences, King Saud Univeristy, Riyadh, 11543, Saudi Arabia

ARTICLE INFO

Article History:
Received 20 April 2020
Revised 21 July 2020
Accepted 28 July 2020
Available online 31 July 2020

Keywords:
Part-of-speech (POS) tagging
Conditional random fields
Bidirectional Long Short-Term Memory
Arabic Tweets

ABSTRACT

Over the past few years, Twitter has experienced massive growth and the volume of its online content has increased rapidly. This content has been a rich source for several studies that focused on natural language processing (NLP) research. However, Twitter data pose numerous challenges and obstacles to NLP tasks. For the English language, Twitter has an NLP tool that provides tweet-specific NLP tasks, which present significant opportunities for English NLP research and applications. Part-of-speech (POS) tagging for English tweets is one of the tasks that is offered and facilitated by such a tool. In contrast, only a few attempts have been made to develop POS taggers for Arabic content on Twitter. In this paper, we consider POS tagging, which is one of the NLP tasks that directly affects the performance of other subsequent text processing tasks. We introduce three manually annotated datasets for the POS tagging of Arabic tweets: the 'Mixed,' 'MSA,' and 'GLF' datasets with 3000, 1000, and 1000 Arabic tweets, respectively. In addition, we present an exploratory analysis of the behavior of using hashtags in Arabic tweets, which is a phenomenon that affects the task of POS tagging. We also present two supervised POS taggers that are developed based on two approaches: Conditional Random Fields and Bidirectional Long Short-Term Memory (Bi-LSTM) models. We conclude that the Bi-LSTM-based POS tagger achieves the state-of-the-art results for the 'Mixed' dataset with 96.5% accuracy. However, the specific-dialect taggers trained on the 'MSA' and 'GLF' datasets achieve an accuracy of 95.6% and 95%, respectively. The results for the 'Mixed' dataset indicate the effectiveness of developing a joint POS tagger without the need for a dialect-specific POS tagger.

1. Introduction

Natural language processing (NLP) involves several tasks and applications. Part-of-speech (POS) tagging is one of the first processes that directly affect the performance of other subsequent text processing tasks in NLP applications (Albared et al., 2011). The performance of most NLP tasks and applications depends on the genre of the text being processed. Recently, the popular microblogging service "Twitter" has experienced a significant growth rate in the last few years; where it encourages people to post millions of messages. Thus, Twitter is a rich and fruitful source of data to study the evolution of various issues. However, Twitter data pose numerous challenges due to the nature of text in microblogs, such as the restricted length, which leads to a substantial number of abbreviations, and noisy and informal content. In addition, grammar and correct spellings are usually not properly adhered to in Twitter (Farghaly and Shaalan, 2009). Hence, processing Twitter data differ from other genres of text.

Twitter-based POS taggers and NLP tools provide POS tagging for the English language, and this presents significant opportunities for English NLP research and applications. In contrast, the lack of Twitter-based POS taggers for Arabic is a clear result of the lack of Arabic annotated datasets for POS tagging. To date, only a few studies have investigated this problem and developed

*Corresponding author.
E-mail addresses: iwasan.kh@gmail.com (W. AlKhwiter), twairesh@ksu.edu.sa (N. Al-Twairesh).

https://doi.org/10.1016/j.csl.2020.101138
0885-2308/© 2020 Elsevier Ltd. All rights reserved.

POS taggers for Arabic tweets (Al-Sabbagh and Girju, 2012) (Albogamy and Ramasy, 2015) (Darwish et al., 2018) (Alharbi et al., 2018). Furthermore, although the problem of POS tagging has been solved using different approaches in literature, deep learning-based studies are still relatively scarce, and hence, the use of deep learning approaches is explored for this task in this paper.

The Arabic language belongs to the Semitic language family, and is the official language of more than twenty countries in Africa and the Middle East. It is considered to be the fourth most used language on the web.[1]The Arabic language has different variants, which are: Classical Arabic (CA), Modern Standard Arabic (MSA), and Colloquial or Dialectal Arabic (DA). CA was the language used in ancient days, and MSA is the primary written language used by the media and in education, these days. DA, however, is the everyday spoken language that exists in different varieties according to the country or Arab region the speaker is from. DA involves different dialects based on geographical locations in Arab countries (El-Beltagy and Ali, 2013). Therefore, DA varies geographically and socially and is not standardized (Zitouni, 2016). Dialects differ from MSA phonologically, morphologically and syntactically (Habash, 2010). Moreover, dialects do not have standard orthographies. This makes the task of building morphological analyzers and POS taggers for dialects a big challenge. Hence, these varieties of the Arabic language require advanced processing for Arabic text. Until recently, DA was mostly spoken and was never found in written form. The proliferation of social media has changed this trend, as Arab users tend to use DA in these new venues. Hence, DA is now also found in written form. This paper focuses on one of the Arabic dialects, namely Gulf (GLF). According to Habash (Habash, 2010), GLF Arabic includes the dialects of Kuwait, United Arab Emirates, Bahrain, Qatar, and Saudi Arabia.

In this paper, we aim to build POS taggers for tweets written in both MSA and the GLF dialect. We present two tagging models by using Conditional Random Fields (CRFs) and Bidirectional Long Short-Term Memory (Bi-LSTM). To train these models, we have constructed three datasets: 'Mixed,' 'MSA,' and 'GLF' with 3000, 1000, and 1000 Arabic tweets, respectively. A series of experiments were conducted to evaluate our trained models. We also investigated the effect of features derived from the morphological analyzer MADAMIRA (Pasha et al., 2014) on the tagging performance. To the best of our knowledge, the 'Mixed' dataset that has 3000 Arabic tweets is, till now, the largest annotated dataset extracted from Twitter for POS tagging.

The contributions of this paper are as follows:

- We present a POS tagger for Arabic tweets using a deep learning approach that achieves a state-of-the-art performance.
- POS taggers are developed for MSA and GLF variants of the Arabic language.
- The gold standard annotated datasets that have been constructed for POS tagging are made accessible to the research community.
- We present an exploratory analysis of the behavior of using hashtags in Arabic tweets, and this can be leveraged in future studies.

The paper is structured as follows. Section 2 presents related work. Section 3 introduces the dataset for POS tagging for Arabic tweets. Section 4 presents hashtag analysis. Section 5 demonstrates the adopted features. Section 6 presents the tagging methods. Section 7 describes experiments. Section 8 discusses the obtained results and presents error analysis. Finally, Section 9 concludes the paper and discusses future work.

2. Related work

POS tagging is a well-studied problem in NLP over the past decades. Several studies have been conducted to develop POS taggers that are tailored for social media text. Gimpel et al. (2011) presented one of the preliminary POS tagging methods for English tweets included in a web-based CMU Twitter NLP toolkit (ArkNLP). They developed a tagset of 25 tags and used it to annotate a corpus consisting of 1827 tweets (26,436 tokens). The corpus was divided into training/development/test sets of 1000/327/500 tweets, respectively. Ark POS tagger adopted the CRF model developed by (Lafferty et al., 2001) by using a feature-based sequence tagging model and achieved 89.37% tagging accuracy. Ritter et al., al.(2011) presented another Twitter POS tagger (T-POS). It used a tagset based on the Penn Treebank set and adopted the CRF approach, it achieved 88.4% tagging accuracy. Owoputi et al. (2012) presented an improved Ark POS tagger by extracting several word features from large-scale word clusters. These improvements resulted in an overall performance gain of 3.6%, thus increasing the accuracy to 92.8%. Derczynski et al. (2013) presented a detailed error analysis of existing POS taggers for English tweets. In addition, they presented a novel approach for system combination for the case in which the available taggers used different tagsets and achieved 88.7% tagging accuracy.

For the Arabic language, a few efforts have been proposed to build POS taggers tailored to Twitter text or similar text genres. Al-Sabbagh and Girju (2012) presented an implementation of Brill's transformation-based POS tagging algorithm trained on a manually-annotated Twitter-based Egyptian Arabic corpus of 423,691 tokens. They used a function-based annotation scheme in which words were labeled based on their grammatical function rather than morpho-syntactic structure. The tagger achieved 87.6% tagging accuracy. However, it did not cover other Arabic dialects. Furthermore, tweet-specific tags were not provided for the different phenomena found in tweets, such as hashtags, URLs, and mentions. Albogamy and Ramsay (2015) evaluated three state-of-the-art POS taggers for Arabic -AMIRA (Diab, 2009), MADA (Habash et al., 2009), and Stanford (Toutanova et al., 2003) after applying them to Arabic tweets. Based on the observed errors, they presented their approach, which achieved 79% tagging

[1] http://www.internetworldstats.com/ Accessed: 10 Feb, 2018

accuracy on a relatively small corpus of 390 tweets (5454 tokens). The same authors (Albogamy and Ramsay, 2016) introduced a fast and robust POS tagger using agreement-based bootstrapping to avoid the noisy behavior of the domain for Arabic tweets and achieved 74% tagging accuracy.

Recently, Darwish et al. (2018) presented a multi-dialect CRF based POS tagger for Arabic tweets trained on a corpus composed of 1400 tweets from four dialects: Egyptian, Levantine, Gulf, and Maghrebi. To validate their approach, they have manually segmented a set of 350 tweets in each dialect and used a tagset of 24 tags that includes six new tags for Twitter and dialect items. The tagger achieved 89.3% average tagging accuracy for all dialects. Alharbi et al. (2018) presented Gulf POS taggers using SVM and a deep learning approach using Bi-LSTM. They used the gold annotated dataset of (Samih et al., 2017) that consisted of 343 Tweets (6844 tokens), and used a tagset of 21 tags. The Bi-LSTM-based POS tagger achieved 91.2% tagging accuracy. These studies present the efforts to develop preliminary POS taggers for Arabic tweets. However, the size of corpora that were used and on which the models were trained, were considerably small. This, when coupled with the coarse tagsets, provided a generic POS tag for a given word. In this paper, we aim to address this gap by presenting a larger and gold standard dataset of Arabic tweets manually annotated with a POS tagset of 44 tags that account for the peculiarities of Arabic dialects and Twitter text.

3. Datasets

This section presents the datasets that have been constructed for the POS tagging of Arabic tweets. We first describe the annotation process, including how the data was collected and annotated, and present the adopted tagset in our POS tagging process. We then present statistical information on the datasets.

3.1. Annotation

We used a supervised learning approach that required an annotated corpus for training the classifiers. However, most of the publicly available annotated corpora for Arabic were either sampled from non-Twitter text, which do not contain the characteristics of tweets, or were built specifically for variants of NLP tasks other than POS tagging. Therefore, we constructed our own corpus by harvesting Twitter to sample a collection of Arabic tweets that preserved the tweet characteristics. To avoid overfitting to time-specific phenomena, as observed in the work of (Gimpel et al., 2011), we collected Arabic tweets on a daily basis in August 2018. In total, 7750 tweets were collected.

Twitter text is known to be informal and noisy, with links, hashtags, emojis, etc. Therefore, these tweets should be filtered, cleaned, and prepared for the annotation process. We filtered the initial set of 7750 tweets by, first, excluding tweets with a length of less than seven words to avoid overly concise tweets, which could introduce ambiguous context. Second, spam tweets were excluded, and the ASA spam list (Al-Twairesh et al., 2016) was used to detect the spam tweets. Third, tweets that were written in dialectal Arabic other than the Gulf dialect were excluded. After the filtering phase, 5133 tweets remained, which represented the dataset size that was prepared for annotation. Thereafter, preprocessing was done in different steps: normalization, tokenization, and POS tagging. We normalized the pre-annotated tweets by removing the tatweel or kasheeda (ـ) symbol, the repeating letters in elongated words to reduce the words to their compressed forms, the diacritics (nine zero-width symbols that can be written optionally and appear above or below the Arabic letters (Habash, 2010)), and letter normalization for the letters, which may have multiple forms such as: أ(>) 'Alef', ة(p) 'Ta-Marbuta', ي(y) 'Ya'a', و(w) 'wāw' (Habash, 2010). This was done by uploading the cleaned tweets into MADARi (Obeid et al., 2018), which is a web application for morphological annotation and spelling correction for texts in Standard and Dialectal Arabic. Tweets were processed by running the morphological analyzer MADAMIRA (Pasha et al., 2014), which is in the core of MADARi. MADAMIRA converts the multiple forms of أ, إ, آ (>) 'Alef' into ا (>) 'Alef', the letter ة(p) 'Ta-Marbuta' into ه (h)'Ha' and the different forms of ي, ى (y) 'Ya'a' into ي (y) 'Ya'a'. Additionally, it tokenizes tweets and automatically tags each token. (throughout the paper we will be using the Buckwalter transliteration (Habash et al., 2007) for Arabic letters and words.

Corpus annotation is defined as "the practice of adding interpretative linguistic information to an electronic corpus" (Garside et al., 1997). We recruited four native Arabic annotators; who held a bachelor's degree in Arabic linguistics. The annotators were presented with annotation guidelines of the Conventional Orthography for Dialectal Arabic (CODA*) (Habash et al., 2018). CODA* is an extension of CODA (Habash et al., 2012), which is oriented for Arabic dialects and designed primarily for the purpose of developing computational models. It provides a set of guidelines. The annotators received a tutorial in video graphic form which had been recorded to demonstrate the POS tagging process and the interface. Additionally, the annotators were provided with an Arabic version of the user manual for the MADARi interface to facilitate the annotation process.

The annotation process was accomplished in several stages. **Stage 1** included manual annotation. Due to a few constraints, we minimized the number of tweets to 3000, which were distributed to four annotators via the MADARi interface. The dataset was split such that two annotators each were responsible for one part. The automatic POS tagging produced by MADARi was presented to the annotators and they were asked to approve or correct the POS tags. In **Stage 2**, we calculated Cohen's kappa to measure the Inter-Annotator Agreement (IAA) for the 3000 tweets that were annotated. The Cohen's κ value was found to be 0.90 on average which reflected 'perfect' agreement according to the common interpretation of the Kappa value (Landis and Koch, 1977). In **Stage 3**, an expert of the Arabic language (Ph.D. in Arabic linguistics) reviewed the tagging decisions that were disagreed upon by the annotators and made a decision on them. A final sweep was conducted to correct errors and improve the consistency of tags across the dataset. The annotation process took approximately two months. All experiments used the output of this final stage of annotation. Fig. 1 shows the annotation procedure.

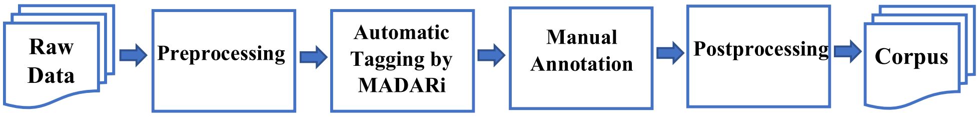

Fig. 1. Annotation Procedure.

For POS annotation, MADARi uses The CAMEL POS tagset and features described by K The CAMEL POS (Khalifa et al., 2018) is designed as a single tagset for both MSA and the dialects and is in the format: POS.features, where POS represents the baseword's POS tag (39 tags), and 'features' is the possible morphological feature combination that is suitable for the POS tag. We used the new CAMEL POS tagset adapted from (Khalifa et al., 2018; Obeid et al., 2018) since it was designed for adaptation of both MSA and dialects and supports backward compatibility with previously annotated resources (Khalifa et al., 2018). We also introduced five tweet-specific tags: # for Twitter hashtags, @ for Twitter at-mentions, RT for retweet, EM for emoticons, and URL for links or email addresses. Table 1 lists the full tagset of 44 tags. The annotation interface 'MADARi' does not allow the adding of new tags for tagging; hence, we had to tag Twitter items with their corresponding tag, which was done as postprocessing. For tagging hashtags, analysis was conducted to tag them based on their role in the tweet, as shown in Section 5. We retained the assigned POS if the hashtag served as a part of the tweet; otherwise, the underscores were reinserted in their original place and subsequently, the full hashtag string was tagged with (#) tag.

3.2. Corpus statistics

Our dataset contains 75,677 annotated tokens (13,914 unique types). We refer to it as 'Mixed' dataset in the following sections since it contains tweets that are in both the MSA and Gulf dialect. Table 2 shows the frequency of all tags in our tagset.

An interesting question was whether it would be effective to isolate the MSA tweets from the GLF tweets. We recruited two annotators to classify the tweets to MSA or GLF. The guidelines were simple, and the tweet was considered GLF if one of the following statements was satisfied (Alharbi et al., 2018); otherwise, it was classified as MSA. The given tweet was considered GLF if it:

- Had one or more dialectal words. E.g. أبي (>by) 'I want' and راح (rAH) 'I will'
- Had one or more words written as it is pronounced. E.g. بئر (b}r) 'well' written as بير (byr) 'well'.
- Had one or more words with dropped letters. E.g. ي رب (y rb) 'oh God' and ع الطاولة (E AlTAwlp) 'on the table.'

Finally, 1017 tweets were labeled as GLF, and the remaining 1983 tweets were labeled as MSA. To construct a balanced dataset for the experiments, we used 1000 GLF tweets and 1000 MSA tweets. In the experiments, we refer to these datasets as the 'GLF' dataset and 'MSA' dataset, respectively. Table 3 illustrates the statistics of all the datasets.

Table 1
The full POS tagset used to annotate tweets.

Old tags from (CAMEL POS Arabic)				New tags	
Tag	Description	Tag	Description	Tag	Description
NOUN	Common Noun	PART_DET	Determiner Particle	#	Hashtag
NOUN_NUM	Cardinal Number	PART_EMPHATIC	Emphatic Particle	@	At_mention
NOUN_PROP	Proper Noun	PART_FOCUS	Focus Particle	RT	Retweet
NOUN_QUANT	Noun Quantifier	PART_FUT	Future Particle	EM	Emoticon or Emoji
ADJ	Adjective	PART_INTERROG	Interrogative Particle	URL	Link
ADJ_COMP	Comparative Adjective	PART_NEG	Negative Particle		
ADJ_NUM	Ordinal Numbers	PART_PROG	Progressive Particle		
ADV	Adverb	PART_RC	Response Conditional Particle		
ADV_INTERROG	Interrogative Adverb	PART_RESTRICT	Restrictive Particle		
ADV_REL	Relative Adverb	PART_VERB	Verb Particle		
VERB	Verb	PART_VOC	Vocative Particle		
VERB_FROZEN	Non-inflectional Verb	CONJ_SUB	Subordinating Conjunction		
VERB_PSEUDO	Pseudo Verb	PREP	Prepositions		
PRON	Bound Pronoun	CONJ	Coordinating Conjunction		
PRON_DEM	Demonstrative Pronoun	DIGIT	Digit		
PRON_EXCLAM	Exclamative Pronoun	ABBREV	Abbreviation		
PRON_INTERROG	Interrogative Pronoun	INTERJ	Interjections		
PRON_REL	Relative Pronoun	FOREIGN	Foreign		
PART	Particle	PUNC	Punctuation		
PART_CONNECT	Connective Particle				

Table 2
The frequency of tags in the dataset. The last column indicates each tag's relative frequency against the total (75,677 tokens).

Tag	*#Tokens*	*%*	*Tag*	*#Tokens*	*%*
#	742	1.0	PRON_REL	879	1.2
@	722	1.0	PART	167	0.2
RT	1845	2.4	PART_CONNECT	27	0.0
EM	1375	1.8	PART_DET	7458	9.9
URL	727	1.0	PART_EMPHATIC	172	0.2
NOUN	17,122	22.6	PART_FOCUS	22	0.0
NOUN_NUM	206	0.3	PART_FUT	210	0.3
NOUN_PROP	2225	2.9	PART_INTERROG	64	0.1
NOUN_QUANT	616	0.8	PART_NEG	1550	2.0
ADJ	2984	3.9	PART_PROG	30	0.0
ADJ_COMP	463	0.6	PART_RC	31	0.0
ADJ_NUM	105	0.1	PART_RESTRICT	169	0.2
ADV	415	0.5	PART_VERB	80	0.1
ADV_INTERROG	181	0.2	PART_VOC	394	0.5
ADV_REL	26	0.0	CONJ_SUB	1401	1.9
VERB	8228	10.9	PREP	6976	9.2
VERB_FROZEN	35	0.0	CONJ	3777	5.0
VERB_PSEUDO	111	0.1	DIGIT	300	0.4
PRON	7645	10.1	ABBREV	133	0.2
PRON_DEM	414	0.5	INTERJ	65	0.1
PRON_EXCLAM	11	0.0	FOREIGN	58	0.1
PRON_INTERROG	69	0.1	PUNC	5447	7.2
Total				***75,677***	***100%***

4. Hashtag analysis

Twitter users can use hashtags in different ways. In fact, the use of hashtags differs across cultures. For example, in English, underscores are not used, and hashtags with more than one word are written by capitalizing each word, such as #FirstSecond. In contrast, in Arabic, the underscores are inserted to delimit the parts of the word. It was observed that there were many hashtags in our dataset that are used as a part of the text in tweets and this affected their suitability to be tagged as a hashtag. Table 4 shows an example.

In the first example, sequencing of the context must be maintained; hence, this case cannot be handled as a hashtag. For this reason, we conducted an exploratory analysis of the behavior of using hashtags for all hashtags in our dataset. The analysis involved observing the role of the hashtag (as a hashtag or as a part of the tweet), position of the hashtag (beginning, middle or end), and length (one-word or more than one word). Fig. 2 shows the analysis of hashtags in our dataset.

Among the 742 hashtags in our data, 254 hashtags (34%) were served as a part of the tweet, while 488 hashtags (66%) as a hashtag. It was inferred that neglecting this percentage of hashtags, which served as a part of the tweet, would affect the sequence labeling process such that the true sequencing of POSs would not be determined. Hence, we retained the assigned POS to the words in the hashtag and retained the (#) and the boundary of the hashtag to indicate that a hashtag had appeared here. Overall, it seems that there is no significant preference for using hashtags that had more than a word over those that had one word, which made up 55% and 45%, respectively, of the total.

Regarding the position, the hashtag was considered 'at the beginning' if it was not preceded with a non-hashtag word or if it occurred after a hashtag that was at the beginning of the tweet. The hashtag was considered 'at the end' if it occurred at the end of the tweet (last word) or if it occurred before a hashtag that was at the end of the tweet. The hashtag was considered 'in the middle' if it occurred before or after non-hashtag words. Fig. 3 shows the analysis.

Among the 742 hashtags in our data, the hashtag positions were (33%), (24%), and (43%) at the beginning, middle and end, respectively. This indicates that hashtags in Arabic tweets do not have strong positional preference unlike English tweets, where hashtags tend to occur near the end (Gimpel et al., 2011). Most of the hashtags that were a part of the text of tweets appeared in the middle (98%). In contrast, 82% of those appearing at the beginning and 89% of those appearing at the end served as a hashtag. Thus, a few rules could be derived: any hashtag that appeared in the middle of the tweet could be considered as a part of its text

Table 3
Dataset statistics.

Datasets	Tokens	Tweets
Mixed Dataset	75,677	3000
MSA dataset	25,460	1000
GLF Dataset	22,474	1000

Table 4
Role of hashtag.

Tweet	Hashtag's role
اخترنا تدشين طقمنا الجديد مع فئة غالية علينا جدا من جمهورنا الكبير تشرفنا بزيارتهم و #تدشين_طقم_الأهلي معهم، فشكرا لهم (AxtrnA td$yn TqmnA Aljdyd mE f}p gAlyp ElynA jdA mn jmhwrnA Alkbyr t$rfnA bzyArthm w #td$yn_ Tqm_Al>hly mEhm, f$krA lhm) 'We chose to launch our new kit with a very expensive class of our great audience honored to visit them and #LaunchAlAhliKit with them, so thanks to them'	Part of the tweet's text
ربى انى مسنى الضر وانت ارحم الراحمين. #ادعيه_نبويه (rbY AnY msnY AlDr wAnt ArHm AlrAHmyn. #AdEyh_nbwyh('my lord truly distress has seized me, but thou are the most merciful of those that are merciful. #PropheticPrayers'	As a hashtag

and its words could be tagged to their corresponding POS tags to maintain the sequence of labels. In contrast, neglecting the words of hashtags that appeared at the beginning or end of tweets would not affect the sequencing, in which they were more likely to be used as a hashtag. Concerning the length of the hashtags, 70% of the hashtags that appeared in the middle of the tweet were one-word hashtags, while 74% of the hashtags that appeared at the beginning were more than a single word. From the analysis results, we formulated a few rules to be followed for hashtag handling in POS tagging in future studies.

5. Features

Since, the MADARi (Obeid et al., 2018) annotation interface, which runs the morphological analyzer MADAMIRA (Pasha et al., 2014) in its core, was used; the form (POS.Features) was obtained using MADARi through automatic tagging. These features refer to specific morpho-syntactic aspects of the word. They represent aspect, person, gender, and number. Each feature has several values as follows:

Aspect: with the values Perfective (P), Imperfective (I) and Command (C).
Person: with the values 1st (1), 2nd (2), and 3rd (3).
Gender: with values Masculine (M) and Feminine (F).
Number: with values Singular (S), Dual (D) and Plural (P).

The specified values of the different features are represented in combinations in the following order: <*A*><*P*><*G*><*N*>. Not all POS tags have these features; for example, PREP POS has no features. In the experiments, all these morphological features produced by MADARi were extracted as a feature set, and only the POS was used as a label.

6. Tagging methods

6.1. CRFs

Conditional Random Fields CRFs (Lafferty et al., 2001) have proven to achieve state-of-the-art performance in several sequence labeling tasks. CRFs estimate the probabilities of possible label sequences for a given observation sequence. A previous study has shown that the CRF-based tagger that was developed for English tweets achieved high accuracy (Gimpel et al., 2011).

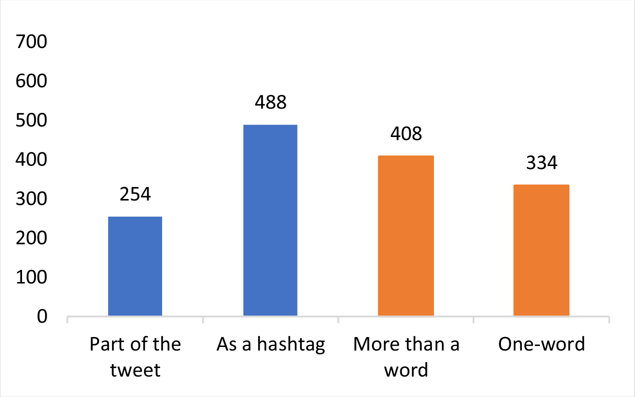

Fig. 2. Hashtag analysis based on the role and length.

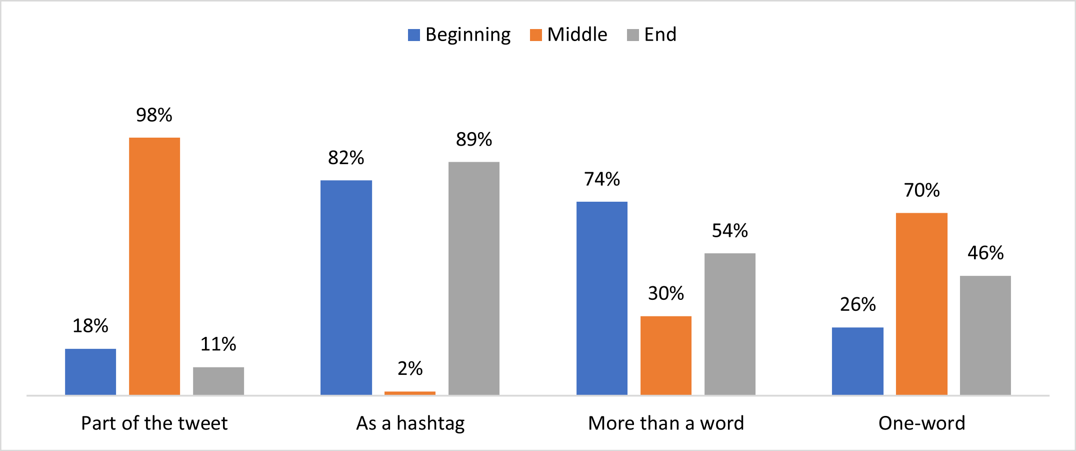

Fig. 3. Hashtag analysis based on the position.

Hence, we used this sequence labeling method on our gold standard dataset which is much larger than the corpus of the previous study on POS tagging of Arabic tweets in (Darwish et al., 2018).

6.2. Bi-LSTM

Over the past few years, different sequence modeling tasks have been accomplished successfully by using Bi-LSTM networks. Furthermore, due to its advantage of mitigating the need for specific feature extraction, the Bi-LSTM-based POS tagging achieved 97.36% accuracy without using any features on English ‘Wikipedia articles’ (Ling et al., 2015). Hence, we used the Bi-LSTM deep learning approach for POS tagging to investigate its performance on another genre of text i.e. Twitter data comprised of Arabic tweets.

7. Experiments

7.1. Experimental setting

We divided all three datasets, namely, ‘Mixed,’ ‘MSA,’ and ‘GLF’ into 80/10/10 train, development, and test sets, respectively. The test set consisted of tweets that were not used to set up the model, while the train and development sets were used to tune the classifier to the optimal value of the parameters that reported the best performance.

Our evaluation tested the efficiency of the proposed models for the task of POS tagging. We have experimentally evaluated the performance of the two proposed methods: CRFs and Bi-LSTM by measuring the accuracy which represents the percentage of tags correctly tagged on the test set.

$$Accuracy = \frac{Number\ of\ words\ correctly\ tagged}{Total\ number\ of\ Words}$$

The experimental results were compared against our gold standard dataset to choose the model that achieved the best performance.

7.1.1. CRFs

We used one of the well-known implementations of CRFs, namely CRF++ (“CRF++: Yet Another CRF toolkit,” n.d.). The training and development sets were used to obtain the best classifier having generalization parameter “C” as 0.1 with L2 regularization algorithm. All experiments were conducted under this setting.

The first set of experiments (Set 1), established the baseline for the three datasets to be compared with. The baseline contained only the basic features. In these experiments, we tested the model with different size contexts, namely unigram and bigram features. This was done to evaluate the impact of adding the features in the next set of experiments. **In the second set of experiments (Set 2)**, we used a combination of the morphological features: aspect, person, gender, and number that corresponded to each token. The aim was to determine whether using these features would improve the performance in POS tagging. **In the third set of experiments (Set 3),** we extended the size window to ‘8′ (which by neglecting the boundaries, referred to the two previous and next tokens). The sets of experiments are summarized in Table 5. The experiments on ‘Mixed’ dataset tested the efficiency of POS tagging for mixed tweets (MSA and GLF). If the classifiers achieved good results, this could indicate that a joint model could be developed for POS tagging, instead of a dialect-specific model. In the experiments conducted on MSA and GLF datasets, an objective was to observe how the performance of a dialect-specific POS tagging model could be improved.

Table 5
CRF set of experiments.

Set	Experiment
Set1	BL_{Mixed}, BL_{GLF}, BL_{MSA}: unigram + bigram combination of clitic.
Set2	BL_{Mixed}, BL_{GLF}, BL_{MSA} + morph features (aspect, person, gender, and number)
Set3	BL_{Mixed}, BL_{GLF}, BL_{MSA} + window size = ‘8′.
	BL_{Mixed}, BL_{GLF}, BL_{MSA} + morphological features + window size = ‘8′.

For our experiments, we assumed that perfect tokenization for a given sequence of words would result in a sequence of clitics: $c_n...c_{-2}, c_{-1}, c_0, c_1, c_2...c_m$. The clitic c_0 would have the unigram and bigram features in the simplest form of baseline, and a combination of morphological features. For all datasets, the CoNLL format (“CoNLL-U Format,” n.d.) was applied. To retain the boundary of a word after tokenizing to several clitics, we used the following boundaries: token boundary (TB) tag, word boundary (WB) tag and end of sentence (EOS) tag, respectively.

7.1.2. Bi-LSTM

The model was trained using the experimental values for the hyper-parameters, which yielded the best performance on the development set. The most efficient model was obtained using the following hyper-parameter values: dimensions of the word embeddings were 128, an LSTM hidden layer was used with a bidirectional modifier, composed of 30 neurons, training was performed with a mini-batch size of 100, the learning rate was 0.003, number of epochs were 100, and the activation function was set to ‘softmax’ due to our multi-class classification. The model was composed of three layers: the *input layer* contained the word embeddings, *hidden layer* (bi-LSTM) mapped word representations to hidden sequences, and *dense layer* picked the appropriate POS tag. The ‘TimeDistributed’ identifier was added to run on each element of the sequence. Our task of sequence modeling required a fixed length for the input and output sequences to allow the LSTM model to efficiently perform batch matrix operations. Thus, we followed the general approach for sentence padding, where each sequence was padded at the borders to the length of the longest sequence in the dataset. However, these paddings should be ignored when the accuracy is computed. Therefore, we computed our model accuracy after ignoring the padding predictions.

Word embeddings refer to encoding the text data to numeric values. In RNNs, word embeddings can be derived via an embedding layer and trained through backpropagation along with the rest of the network. Hence, we added the embedding layer, and encoded each word and POS tag in our data to a unique value, and substituted with this value in our dataset to perform pointwise operations on the data. In the LSTM layer, using the bidirectional modifier inputted the next and previous values in the sequence. We performed the experiment for the Bi-LSTM POS tagging model without any features. This is because, in the work of Ling et al. (Ling et al., 2015), an accuracy of 97.36% had been achieved using the Bi-LSTM tagger on English tweets without any features.

8. Results and discussion

8.1. CRF model results

In these set of experiments, the effectiveness of the CRF approach was evaluated for POS tagging on the development sets of the three annotated datasets, namely ‘Mixed,’ ‘MSA,’ and ‘GLF’

8.1.1. Cross validation method results

We created 5-fold partitions for cross-validation with 80/10/10 train/dev/test splits for each dataset. We conducted the set of experiments using our CRF-based model for training and testing on variant datasets. Since we used 5-fold cross validation, we report on the average across all folds. The results obtained from all the sets of experiments are shown in Table 6.

In the first set of experiments (**Set 1**), the baselines were established. As observed in Table 6, the highest accuracy of 83.7% is achieved for the ‘Mixed’ dataset. From **Set 1** and **Set 2**, it is inferred that adding the morphological features improves the accuracy for all datasets. The performance increases for the ‘Mixed,’ ‘MSA,’ and ‘GLF’ datasets by + 6.5%, +7.9% and +9.3%, respectively. In **Set 3**, the extension of the window size in BL marginally improves the performance on all, whereas the accuracy of extending window size with morph features is slightly decreased on all datasets.

Table 6
CRF experiment results using 5-folds cross validation.

		Accuracy		
		Mixed	MSA	GLF
Set 1	BL	83.7	80.6	76.1
Set 2	+ morph features	**90.2**	**88.5**	**85.4**
Set 3	BL + window size=8	84.1	81.2	76.2
	+ morph features + window size=8	90.0	88.1	84.6

Table 7
CRF experiment results using held-out method.

		Accuracy		
		Mixed	MSA	GLF
Set 1	BL	86.8	84.4	82. 8
Set 2	+ morph features	**91.3**	**91.0**	89.0
Set 3	BL + window size=8	87.8	86.0	83.6
	+ morph features + window size=8	91.2	90.7	**89.1**

8.1.2. Held-out method results

We conducted the set of experiments using our CRF-based model for training and testing on variant datasets. The results obtained from all the sets of experiments are shown in Table 7.

In the first set of experiments (**Set 1**), the baselines were established. As observed in Table 7, the highest accuracy of 86.8% is achieved for the 'Mixed' dataset. From **Set 1** and **Set 2**, it is inferred that adding the morphological features improves the accuracy for all datasets. The performance increases for the 'Mixed,' 'MSA,' and 'GLF' datasets by + 4.5%, +6.6% and +6.2%, respectively. In **Set 3**, the extension of the window size marginally improves the performance in the 'GLF' dataset, whereas the accuracy on both Mixed and MSA datasets is slightly decreased.

8.2. Bi-LSTM results

Although we conducted basic experiments using Bi-LSTM without any features, the obtained results were superior on all datasets as shown in Table 8. The Bi-LSTM-based model achieves 96.5% accuracy on 'Mixed' dataset, and a closely matching performance is achieved on both the 'MSA' and 'GLF' datasets with 95.6% and 95.0% accuracy, respectively.

For the CRF-based model, it is observed that the Experiment **Set 2** achieves the highest accuracies of 91.3%, 91.0%, and 89.1% for the 'Mixed,' 'MSA,' and 'GLF' datasets, respectively. Therefore, we adopted the setting of Experiment **Set 2** for testing. Regarding the Bi-LSTM-based model, excellent results have been obtained over all datasets; precisely 96.5% was achieved for the 'Mixed' dataset. This supports our hypothesis that because the Bi-LSTM approach achieved 97.36% accuracy in English POS tagging without any features (Ling et al., 2015), a close result was expected for the Arabic language as well.

The results achieved by the two proposed methods: CRF when performing the classification on the test set, and Bi-LSTM are shown in Table 9. We can see from this table that the best performance for all the datasets was using Bi-LSTM.

In Table 10, we compare the results of our POS taggers for each method with the state-of-the-art results in literature for the same method. Both (Darwish et al., 2018) and (Alharbi et al., 2018), considered the dialects, and hence, our results for only the same dialects are used for comparison. Although we note here that their tagset and datasets are different than ours as we showed in the related work section.

Concerning the CRF, our CRF-based tagger for GLF outperforms the model presented in (Darwish et al., 2018) by +3.4% accuracy. While for the MSA, our tagger achieves a close result but did not add an improvement on the top of their tagger's performance. However, it is worth mentioning that we have excluded the boundary prediction accuracy from the overall accuracy of our model results to reflect the actual results of POS tagging. The results for 'MSA' and 'GLF' without excluding boundary predictions are 96.4% and 95.6%, respectively. Regarding the Bi-LSTM model, our results show a 4% improvement over the accuracy of the GULF-specific POS tagger presented in (Alharbi et al., 2018). In general, and to the best of our knowledge, our Bi-LSTM-based POS tagger achieves the state-of-the-art results with 96.5% accuracy on the 'Mixed' dataset having 3000 tweets. Furthermore,

Table 8
Bi-LSTM results.

Dataset	Accuracy
Mixed dataset	96.5
MSA dataset	95.6
GLF dataset	95.0

Table 9
Results of CRF and Bi-LSTM approaches.

	CRF-based POS Tagger '5-fold cross validation'	CRF-based POS Tagger 'held out'	Bi-LSTM-base POS Tagger
Mixed dataset	90.4	91.6	96.5
MSA dataset	87.4	92.6	95.6
GLF dataset	86	91.2	95.0

Table 10
Comparison of the results of the proposed methods with previous attempts.

	Previous work		Our results	
	CRF (Darwish et al., 2018)	Bi-LSTM (Alharbi et al., 2018)	CRF	Bi-LSTM
MSA	**93.6**	–	92.6	**95.6**
GLF	87.8	91	**90.0**	**95.0**

dialect-specific Bi-LSTM-based POS taggers outperform the existing dialect-specific models with accuracies of 95.6% for ‘MSA’ and 95.0% for ‘GLF.’

Error analysis is important in POS tagging to analyze the nature of incorrectly classified tags. We conducted error analysis for the predicted POS tags by the CRF-based model. The top 10 error types were captured for each dataset. Results are shown in Table 11.

The most common error over all datasets was to confuse NOUN with ADJ, and NOUN with NOUN_PROP. This might have occurred due to the similarity in the attached features for nouns, and adjectives, which only represent the gender and number features. Additionally, it may be noted that this type of error resulted from decisions that were deemed to be the most difficult to make, according to the disagreements between the annotators. Thus, if it is difficult for the human to predict, it will be the same for the machine. The ambiguity of proper nouns might be a result of the lack of common names in the gazetteers used by MADAMIRA. Furthermore, in comparison to nouns in English, Arabic has no similar structure to differentiate the proper nouns as done using capitalization in English. As observed in Table 11, the tagger also struggles with twitter-specific tags (@, #, URL, RT). This can be explained by the loss of a pattern of the surrounding context, along with the loss of meta-types that can identify these entities. Finally, the misclassification of PART_NEG and PRON_REL can be explained by the fact that the function wordما (mA) ‘what’ has 13 different functions in Arabic. Additionally, it may be due to the frequent use ofما (mA) ‘not’ as a negation word in Dialectal Arabic (DA); hence, greater similarity with PRON_REL in terms of context could have caused this type of error.

9. Conclusion

We have introduced new datasets for POS tagging that are constructed from Arabic tweets. A supervised approach is used to train two different models, namely CRF and Bi-LSTM on such annotated datasets. It is shown that the proposed Bi-LSTM-based POS tagger achieves the state-of-the-art results over the existing dialect-specific models with 96.5% accuracy on a ‘Mixed’ dataset of 3000 tweets. However, the specific-dialect taggers for MSA and Gulf achieve an accuracy of 95.6% and 95%, respectively on our datasets. The results on ‘Mixed’ dataset indicate the effectiveness of developing a joint POS tagger without the need for a dialect-specific POS tagger.

For the CRF-based model, it has been observed from error analysis that the tagger is unable to tag twitter-specific items accurately. Hence, in future, our objective is to use meta-types to identify whether a specific clitic is a retweet, mention, hashtag, or an URL. Regarding the Bi-LSTM-based model, our future work is to use pre-trained language models such as fastText or BERT. Further, we plan to develop a tokenizer that can be packaged with our existing POS tagger in a joint tool for NLP tasks for Arabic tweets. Additionally, we plan to investigate building a joint model capable of POS tagging for the MSA and Gulf dialects with minimal loss of accuracy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table 11
Top 10 types of errors produced by the CRF model for each dataset.

Mixed dataset		MSA dataset		GLF dataset	
Error type	Counts	Error type	Counts	Error type	Counts
NOUN -> ADJ	174	NOUN -> ADJ	114	NOUN -> ADJ	129
NOUN -> NOUN_PROP	88	NOUN -> NOUN_PROP	60	NOUN -> NOUN_PROP	60
RT -> @	38	#->@	19	RT -> @	43
CONJ_SUB-> PART_EMPHATIC	31	RT -> @	13	CONJ -> PREP	16
CONJ -> PREP	29	ADJ -> NOUN	13	ADJ -> NOUN	15
ADJ -> NOUN	23	NOUN -> ADJ_COMP	11	URL -> EM	12
@ -> #	22	URL -> #	9	CONJ -> CONJ_SUB	12
PART_NEG -> PRON_REL	20	CONJ -> PREP	8	@ -> #	11
NOUN_PROP -> NOUN	18	NOUN -> PRON_DEM	7	NOUN -> ADJ_COMP	11
CONJ -> CONJ_SUB	15	ADJ -> NOUN_PROP	5	PART_NEG -> PRON_REL	9

Acknowledgements

We would like to thank Nizar Habash, Salam Khalifa, and Osama Obeid for providing us with access to MADARi for use in the annotation process. This research project was supported by a grant from the “Research Center of the Female Scientific and Medical Colleges,” Deanship of Scientific Research, King Saud University. The authors thank the Deanship of Scientific Research and RSSU at King Saud University for their technical support.

References

Al-Sabbagh, R., Girju, R., 2012. A Supervised POS Tagger for Written Arabic Social Networking Corpora. KONVENS, Viennapp. 39–52.

Al-Twairesh, N., Al-Tuwaijri, M., Al-Moammar, A., Al-Humoud, S., 2016. Arabic Spam Detection in Twitter. The 2nd Workshop on Arabic Corpora and Processing Tools 2016, pp. 38–43.

Albared, M., Omar, N., Ab Aziz, M.J., 2011. Developing a Competitive HMM Arabic POS Tagger Using Small Training Corpora. Asian Conference on Intelligent Information and Database Systems. Berlin, Heidelberg. Springer, pp. 288–296. https://doi.org/10.1007/978-3-642-20039-7_29.

Albogamy, F., Ramasy, A., 2015. Towards POS Tagging for Arabic Tweets. In: Proceedings of the ACL 2015 Workshop on Noisy User-Generated Text, Beijing, Chinapp. 167–171.

Albogamy, F., Ramsay, A., 2016. Fast and Robust POS tagger for Arabic Tweets Using Agreement-based Bootstrapping. LREC.

Albogamy, F., Ramsay, A., 2015. POS Tagging for Arabic Tweets. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgariapp. 1–8.

Alharbi, R., Magdy, W., Darwish, K., Abdelali, A., Mubarak, H., 2018. Part-of-Speech Tagging for Arabic Gulf Dialect Using Bi-LSTM. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

CoNLL-U Format[WWW Document], n.d. URLhttps://universaldependencies.org/format.html(accessed 4.11.19).

CRF++: Yet Another CRF toolkit[WWW Document], n.d. URLhttps://taku910.github.io/crfpp/(accessed 4.11.19).

Darwish, K., Mubarak, H., Eldesouki, M., Abdelali, A., Samih, Y., Alharbi, R., Attia, M., Magdy, W., Kallmeyer, L., 2018. Multi-Dialect Arabic POS Tagging: a CRF Approach. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

Derczynski, L., Ritter, A., Clark, S., Bontcheva, K., 2013. Twitter Part-of-Speech Tagging for All: overcoming Sparse and Noisy Data. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, pp. 198–206.

Diab, M.T., 2009. Second Generation AMIRA Tools for Arabic Processing: fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. 2nd International Conference on Arabic Language Resources and Tools.

El-Beltagy, S., Ali, A., 2013. Open issues in the sentiment analysis of Arabic social media: a case study. 2013 9th International Conference on Innovations in Information Technology (IIT), pp. 215–220.

Farghaly, A., Shaalan, K., 2009. Arabic Natural Language Processing: challenges and Solutions. ACM Transactions on Asian Language Information Processing (TALIP), p. 8. https://doi.org/10.1145/1644879.1644881.

Garside, R., Leech, G., McEnery, T., 1997. Corpus annotation: Linguistic Information from Computer Text Corpora. Taylor & Francis.

Gimpel, K., Schneider, N., O 'connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A., 2011. Part-of-Speech Tagging for Twitter: annotation, Features, and Experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:Shortpapers, Portland, Oregonpp. 42–47.

Habash, N., Diab, M., Rambow, O., 2012. Conventional Orthography for Dialectal Arabic. In: Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbulpp. 711–718.

Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., Al-Shargi, F., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., Saddiki, H., 2018. Unified Guidelines and Resources for Arabic Dialect Orthography. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

Habash, N., Rambow, O., Roth, R., 2009. MADA+TOKAN: a Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In: Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, pp. 102–109.

Habash, N., Soudi, A., Buckwalter, T., 2007. On Arabic Transliteration. Arabic Computational Morphology. Springer, pp. 15–22. https://doi.org/10.1007/978-1-4020-6046-5_2.

Habash, N.Y., 2010. Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. https://doi.org/10.2200/S00277ED1V01Y201008HLT010.

Khalifa, S., Habash, N., Eryani, F., Obeid, O., Abdulrahim, D., Kaabi, M.Al, 2018. A Morphologically Annotated Corpus of Emirati Arabic. LREC, pp. 3839–3846.

Lafferty, J., Mccallum, A., Pereira, F.C.N., Pereira, F., 2001. Conditional Random Fields: probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA, USA. Morgan Kaufmann Publishers Inc, pp. 282–289.

Landis, J.R., Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics 33, 159. https://doi.org/10.2307/2529310.

Ling, W., Luı́s, T., Marujo, L., Fernandez, A., Amir, S., Dyer, C., Black, A., Trancoso, I., 2015. Finding Function in Form: compositional Character Models for Open Vocabulary Word Representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp. 17–21.

Obeid, O., Khalifa, S., Habash, N., Bouamor, H., Zaghouani, W., Oflazer, K., 2018. MADARi: a Web Interface for Joint Arabic Morphological Annotation and Spelling Correction. In: Proceedings of the Language Resources and Evaluation Conference (LREC), Miyazaki, Japan.

Owoputi, O., O 'connor, B., Dyer, C., Gimpel, K., Schneider, N., 2012. Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances.

Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A.El, Eskander, R., Habash, N., Pooleery, M., Rambow, O., Roth, R.M., 2014. MADAMIRA: a Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. Language Resources and Evaluation Conference (LREC), pp. 1094–1101.

Ritter, A., Clark, S., Etzioni, O., 2011. Named Entity Recognition in Tweets: an Experimental Study. In: Proceeding of Empirical Methods for Natural Language Processing (EMNLP), pp. 1524–1534.

Samih, Y., Eldesouki, M., Attia, M., Darwish, K., Abdelali, A., Mubarak, H., Kallmeyer, L., 2017. Learning from Relatives: unified Dialectal Arabic Segmentation. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 432–441.

Toutanova, K., Klein, D., Manning, C.D., Singer, Y., 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 173–180.

Natural Language Processing of Semitic Languages. In: Zitouni, I. (Ed.), Theory and Applications of Natural Language Processing. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-45358-8.