Improving POS Tagging for Arabic Dialects on Out-of-Domain Texts

Noor Abo Mokh
Indiana University*
noorabom@iu.edu

Daniel Dakota
Indiana University
ddakota@iu.edu

Sandra Kübler
Indiana University
skuebler@indiana.edu

Abstract

We investigate part of speech tagging for four Arabic dialects (Gulf, Levantine, Egyptian, and Maghrebi), in an out-of-domain setting. More specifically, we look at the effectiveness of 1) upsampling the target dialect in the training data of a joint model, 2) increasing the consistency of the annotations, and 3) using word embeddings pre-trained on a large corpus of dialectal Arabic. We increase the accuracy on average by about 20 percentage points.

1 Introduction

Although POS tagging has achieved high results across languages and benchmarks (Bohnet et al., 2018; Heinzerling and Strube, 2019; Wang et al., 2021), there are still challenges, particularly across different domains and for languages with rich morphology, especially in terms of handling rare and unknown words (Plank et al., 2016; Yasunaga et al., 2018). For languages such as Arabic, their diglossic nature adds additional complexity, as POS tagging models must capture a plethora of lexical and syntactic variation plus orthographic differences. For Arabic, the majority of available POS taggers are trained on Modern Standard Arabic (MSA), such as MADAMIRA (Pasha et al., 2014) and Farasa (Darwish and Mubarak, 2016). There is however a growing interest in developing tools specifically for dialectal Arabic (described in Shoufan and Alameri (2015) and Elnagar et al. (2021)), given its preferred use in daily communication, especially on social media platforms and integrated into voice systems.

Our ultimate goal is the computational analysis of syntactic differences across Arabic dialects, which requires syntactically annotated parallel data. However, the existing dialectal parallel corpus, MADAR (Bouamor et al., 2018), does not provide any linguistic annotation. Thus we need access to a POS tagger (and ultimately a parser) that provides reliable analyses across different dialects, in an out-of-domain settings, since all existing POS tagged corpora are from domains different from that of MADAR. In this challenging setting, we investigate methods to improve POS tagging accuracy for the dialects. We investigate solutions that create a single tagger across all dialects as well as individual taggers for each of the four dialects of interest.

The paper is organized as follows: Section 2 gives a short description of dialectal differences, section 3 explains our research questions, section 4 provides a survey of related work. Section 5 describes the corpora and the experimental setup, in sections 6–9, we present the results and an error analysis. We conclude in section 10.

2 Arabic Dialects

Dialects of Arabic show a wide range of linguistic differences, within the dialects themselves and compared to MSA. MSA is mostly used in formal writing such as books and news articles while dialects are used for most other daily communications. Arabic dialects are interesting because much of the variation involves function words, providing strong signals of the presence of syntactic differences.

In Table 1 we provide an example sentence in four dialects that exhibit three instances of syntactic variation. The first example is the complementizer أن 'that'. أن is used in MSA, Levantine (LEV) and Egyptian (EGY) but not in Maghrebi (MAG). In MAG, the complementizer is optional, resulting in different syntactic structures. Another example is found in the use of the interrogatives across dialects. MSA and MAG use an interrogative pronoun (the hamza-alef أ in أتظن in MSA

*The work was done prior to joining Amazon.

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 238 - 248

Dialect	Sentence	Buckwalter[1]
MSA	مرَاد اتظن ان المشكلة ستحل هنَاك	mrAd AtZn An Alm$klp stHl hnAk
LEV	مرَاد انت مفكر انه المشكلة رح تنحل هنَاك	mrAd Ant mfkr Anh Alm$klp rH tnHl hnAk
EGY	مرَاد فَاكر ان المشكلة هتنحل هنَاك	mrAd fAkr An Alm$klp htnHl hnAk
MAG	مرَاد وَاش كتظن المشكلة غَادي تحل تمَاك	mrAd wA$ ktZn Alm$klp gAdy tHl tmAk
Eng.	Do you think (that) this is the solution for the problem Murad?	

Table 1: A parallel sentence selected from MADAR

and وَاش in MAG) while no question word is used in the other dialects. A final difference concerns the future marking. While in all dialects, future marking is obligatory and precedes the verb تحل, each dialect uses a different marker (سَ in MSA, غَادي in MAG, هَ in EGY (in هتنحل), and رح in LEV). Note that in EGY, the particle is realized as a clitic variant as opposed to a separate word in MAG and LEV. Additionally, for MAG, the future marker غَادي, is inflected and carries agreement, unlike in EGY and LEV.

3 Research Questions

Our main question is how we can improve POS tagging for dialectal Arabic when testing on out-of-domain data. To address this question, we break it down into four sub-questions: 1) Does the POS tagger profit more from having access to a large training set even though the majority of training examples are from a different dialect, or is a smaller, dialect specific training set more appropriate? 2) Does upsampling help mitigate the data imbalance in a joint dialectal model? 3) Can we increase consistency in annotations, using minimal effort? And will increased consistency yield an increased accuracy? 4) Can using pre-trained embeddings improve POS tagging performance?

4 Related Work

4.1 Arabic POS Tagging

Many of the currently available POS taggers are trained on MSA, such as MADAMIRA and Farasa (Pasha et al., 2014; Abdelali et al., 2016) (MADAMIRA also supports Egyptian). Recently, more attention has been given to POS tagging for dialectal Arabic. One approach for dialectal Arabic has been to adapt an MSA model. For example, Zribi et al. (2017) adapted an MSA morphological analyzer, which includes a POS tagger, to Tunisian Arabic by integrating a Tunisian-based lexicon, containing roots and patterns. While they report the system's accuracy as 87.3%, such adaptation methods are less effective than dialect-specific taggers. Alharbi et al. (2018); Alharbi and Lee (2020), e.g., found that a tagger designed for a specific dialect, in this case Gulf, performed better than an adapted MSA tagger. Other dialect specific taggers include the tagger by Al-Shargi et al. (2016) for Moroccan and Sanaani and the one by Khalifa et al. (2018) for Emirati[2]. The difficulty of adaption can be attributed to the diglossic nature of Arabic, which makes it challenging for such systems to process colloquial Arabic (Farghaly and Shaalan, 2009; Diab and Habash, 2007). Arabic has the standard form (MSA), and the spoken forms of Arabic (in addition to other varieties such as Classical Arabic), which coexist and are used by speakers in distinct situations. Each of those varieties has its own linguistic features.

A problem concerning dialect specific taggers is that they do not use uniform annotation schemes. Thus, they may be ineffective in a cross-dialectal setting. Darwish et al. (2018) approach this problem by introducing a multi-dialectal POS tagger for the dialects of Gulf, Levantine, Egyptian and Maghrebi by developing a CRF tagger, which is extended by Darwish et al. (2020) to using bi-LSTM layers as input. Their system provided state-of-the-art performance for POS tagging of dialectal Arabic.

4.2 Domain Adaptation for POS Tagging

Domain adaptation has been pivotal in attempts to handle the differences in data distributions between a source and target domain. Kübler and Baucom (2011) use an ensemble of three POS taggers

[1] http://www.qamus.org/transliteration.htm

[2] See Duh and Kirchhoff (2005); Habash et al. (2013) for overviews of dialect specific POS taggers and NLP tools.

Dialect	No. words: train	No. words: test
Gulf	74 162	21 208
Levantine	80 940	23 090
Egyptian	83 908	23 986
Maghrebi	71 090	20 234

Table 2: Size of the Darwish corpus per dialect.

trained on the source corpus to annotate sentences in the target domain; they then select identically predicted sentences and add them to the training data. These data selection techniques yielded improvements when POS tagging target domain data.

Kuncham et al. (2014) adapt a Hindi morphological analyzer for a domain specific use by adding domain specific words to the lexicon. Another approach is creating POS tagging experts. Mukherjee et al. (2017) create genre experts for POS tagging by using topic modeling in both the training and test set, where they train an expert for each topic and then use the expert to POS tag the same topic. They then assign new test sentences to the genre expert by using similarity metrics.

The importance of including small amounts of target data is attested by Attia and Elkahky (2019). This is further supported by Behzad and Zeldes (2020) who find that a tagger trained on a small amount of Reddit data can outperform taggers trained on much larger out-of-domain corpora.

5 Experimental Setup

5.1 Multidialectal POS-Tagged Corpus

For training, we use the multi-dialectal POS-tagged corpus by Darwish et al. (2018, henceforth the Darwish corpus) since, to the best of our knowledge, it is the only publicly available, POS tagged multi-dialectal corpus for Arabic. The sentences in this corpus are selected from a large collection of Arabic tweets. The corpus includes four major dialects (350 sentences each): Gulf, Levantine, Egyptian, and Maghrebi (representing sub-varieties spoken in Morocco, Algeria and Tunisia). To extract dialectal sentences without code-switching with MSA, Samih et al. (2017b) used a list of exclusively dialectal words such as Maghrebi كيمّا (Eng.: like/as) and Levantine هيك (Eng.: like this). A detailed description of the tweet selection methodology is provided by Eldesouki et al. (2017); Samih et al. (2017b). Table 2 gives an overview of the corpus. Since the sentences are taken from Twitter, they are mostly comments on events, conversations, and attitudes. The corpus was morphologically analyzed using a dialectal morphological analyzer (Samih et al., 2017b).

The POS tagset is derived from the MSA corpus described by Darwish et al. (2017), it includes 18 MSA POS tags, plus two additional dialect specific tags: Prog_Part (tense marker) and Neg_Part (negation marker). A native speaker of each dialect annotated the corpus for POS.

5.2 MADAR

For testing, we use MADAR (Bouamor et al., 2018). The corpus is based on the (English) Basic Traveling Expression Corpus (BTEC) by Takezawa et al. (2007). The English text was translated into dialects of Arabic. This means that we have a significant difference in domains between MADAR and the Darwish corpus.

MADAR is a collection of parallel sentences from different dialects representing the Arabic varieties of 25 cities[3] in addition to MSA, i.e., the information in MADAR is more fine-grained. For compatibility with the Darwish corpus, we group the MADAR data into four major dialects: Egyptian (EGY), Gulf (GLF), Levantine (LEV), and Maghrebi (MAG).

Our initial preprocessing consists of normalizing all Hamzas in all dialects to Alifs and Yaas and then converting to Buckwalter transliteration[4]. Additionally, we removed all hashtags, URLs, and handles from the data since (1) they are not necessary for the purposes of this study (2) this was necessary since the POS tagger does not seem to be able to handle URLs, etc.

5.3 Designing the Gold Standard

Since MADAR is not annotated for POS tags, we selected 100 sentences per dialect to annotate manually. Since we have several translations of each original sentence per dialect (one per city), we ensure that only one version of an original sentence is chosen for a dialect, thus ensuring lexical and syntactic variation in the test sentences. Table 3 shows an overview of the test set.

[3]The following cities are covered: Aleppo, Alexandria, Algiers, Amman, Aswan, Baghdad, Basra, Beirut, Benghazi, Cairo, Damascus, Doha, Fes, Jeddah, Jerusalem, Khartoum, Mosul, Moscut, Rabat, Riyadh, Sanaa, Salt, Sfax, Trupoli, Tunis.

[4]We use the conversion to Buckwalter transliteration from https://github.com/KentonMurray/Buckwalter/blob/master/buckwalter.py

Dialect	No. words
GLF	699
LEV	666
EGY	754
MAG	727
Total	2 846

Table 3: Size of our MADAR test set per dialect

In order to obtain a segmentation close to the one in the Darwish corpus, we used the multi-dialectal Arabic morphological analyzer by Samih et al. (2017b); Eldesouki et al. (2017); Samih et al. (2017a). This morphological analyzer uses a unified segmentation model for the four major dialects Gulf, Levantine, Egyptian, Maghrebi.

We then used the multi-dialectal Arabic POS tagger by Darwish et al. (2018) to automatically POS tag the sentences. Each dialect was corrected by two speakers of Arabic. We then examined inter-annotator agreement: Across all dialects, the annotators showed high agreement (95% for Egyptian, 90% for Levantine, 90% for Gulf, and 85% for Maghrebi). This was followed by an additional pass to resolve differences between annotators. We used the Camel POS tagging guidelines[5] to guide our decisions[6]. For instance, some negation markers were marked as PART, when they are supposed to be marked as NEG_PART.

5.4 Part-of-Speech Tagger

We train the POS tagger using the Bi-LSTM architecture introduced by Darwish et al. (2018, 2020); Alharbi et al. (2018)[7] for tagging dialects of Arabic.

A sentence is fed into the bi-LSTM with a final forward LSTM layer. The neural network of the tagger by Darwish et al. (2018) uses embeddings of stems and affixes trained on the training data, rather than pretrained models. For example, for the word مدخلتوش, the vector represents the stem and the clitics: م, دخل, تو, and ش.

6 First Experiments

6.1 Reproducing Prior Results

We first reproduce the results reported by Darwish et al. (2018) for the joint dialectal experimental setup[8]. Following Darwish et al. (2018), we train on the Darwish corpus using the concatenation of the training sets of all dialects. We then test on each dialect separately using the dialect's test section from the same corpus. Results are shown in Table 4. The first row reports the results by Darwish et al. (2018), and the second row are our results using our preprocessing (see section 5.2). Our results show a higher accuracy than the results reported by Darwish et al. (2018). This may be due to improvements in the POS tagger or the additional preprocessing step, in which we removed Twitter specific tags: hashtags, URLs, and handles.

6.2 Testing on MADAR

We now train the POS tagger using the training sections from the Darwish corpus for all dialects (the joint model)[9] and test on each dialect from MADAR separately. In this setup, target and source data are from the same dialects[10], but different in terms of domains. The results are shown in row 3 in Table 4. The accuracy is lower for all dialects than for the in-domain data in row 2. For instance, the accuracy for GLF is 59.5% for MADAR, but 97.7% when tested on Darwish. We expected the accuracy to be lower for the out-of-domain test data, but the drop in accuracy is rather extreme, between 27.3 and 38.2 percent points. The OOV rates between the Darwish corpus and the MADAR test set range from 36% to 44%, which at least partly explains the results.

These results lead to the question whether training a joint dialectal model is the best solution. The joint model has the advantage of a large training size, but 3/4 of the training data are from dialects other than the one that we are testing on. For this reason, we experiment with training and testing on each dialect separately, to see whether a smaller but dialectally more similar training set results in higher accuracies. In this experiment, we train, e.g., on the Egyptian dialect training data from Darwish and test on Egyptian from MADAR. The results are

[5] https://camel-guidelines.readthedocs.io/en/latest/morphology/

[6] Camel uses a different tagset from that in the Darwish corpus, but it offers guidelines on how to annotate specific phenomena.

[7] Available from https://github.com/qcri/dialectal_arabic_pos_tagger.

[8] Note that the currently available version is different from the one used by Darwish et al. (2018, 2020); Alharbi et al. (2018).

[9] We experimented with adding the MSA section to the training set. Results were considerably lower.

[10] Or as close as possible based on the two corpora.

	Model	Train	Test set	GLF	LEV	EGY	MAG
1)	(Darwish et al., 2018)	Darwish joint	Darwish	87.2	88.6	93.2	87.7
2)	ours + preprocessing	Darwish joint	Darwish	97.7	96.6	95.6	94.4
3)	ours + preprocessing	Darwish joint	MADAR	59.5	61.3	68.3	61.4
4)	ours + preprocessing	Darwish single dialect	MADAR	66.3	67.6	74.4	67.4
5)	ours + preprocessing	Darwish joint upsampled	MADAR	72.8	75.0	81.1	74.2

Table 4: Summary of POS tagging results.

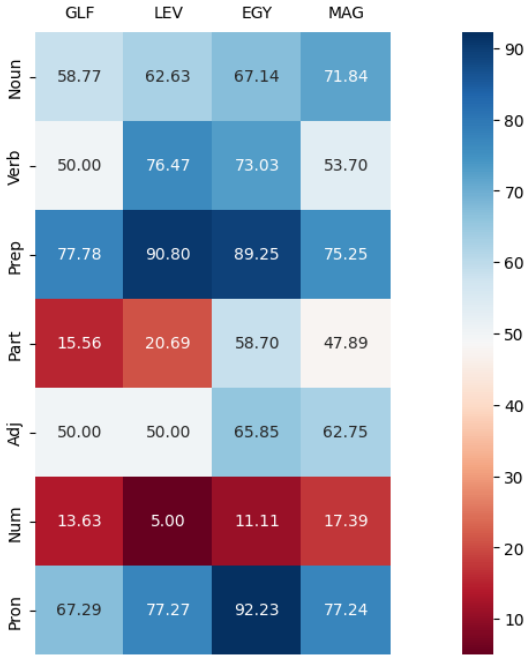

Figure 1: Results per dialect (precision) for MADAR.

reported in row 4 of Table 4. This setting performs worse than testing in-domain (row 2) but improves over the results of using the joint model. For instance, the accuracy for Gulf (GLF) increases to 66.30%, compared to 59.5% for the joint model. The accuracy gain is similar across all dialects. The increase in accuracy despite the smaller training set may be due to the fact that the Darwish corpus focuses on highly dialectal data, which maximizes dialectal differences.

6.3 Error Analysis

We further examine the tagging errors for each dialect: In Figure 1, we show a heatmap for tagging quality for MADAR; we show precision per tag and dialect for the experiment in row 3 in Table 4 (e.g., PREP was correct 75.25%). We focus on the tags which produced the majority of the errors: Pronouns (PRON), Nouns, Numbers (NUM), Adjectives (ADJ), Particles (PART), Prepositions (PREP), Verbs.

Numbers have the lowest tagging precision rate across all dialects, it ranges from 17.39% for MAG to 5.00% for LEV. This low accuracy is due to inconsistencies in annotations in the training set, where numbers sometimes are tagged as nouns and in other cases as NUM. For instance, the number three in the phrase ثلاث دقايق (Eng.: 'three minutes') and in ثلاث سنوات (Eng.: 'three years') are assigned NUM and NOUN respectively. Another issue, which also applies to other POS tags, is the inconsistency in spelling across speakers, for instance, تاني is sometimes spelled with ت, but in other instances with ث. This is an issue for LEV and EGY, where variation in spelling is more likely to occur due to phonetic variation.

Spelling variation may also result in ambiguity in POS tagging. For example, particles (PART) show a remarkably low accuracy for GLF and LEV because of homographic words shared across dialects, resulting in ambiguity. As an example, the word وش, (Eng.: which) in GLF is a particle, the same orthographic form is a noun in LEV وشّ (Eng.: face). Since the model is trained on all dialects, the LEV وشّ is incorrectly assigned the tag PART.

We also notice that future and negation markers show different performance across dialects. For LEV, for instance, the system fails to assign the future marker to any future clitic. A closer examination shows that the same future marker (رح) is marked as FUT_PART in the LEV training data but marked as PART in the MAG data, indicating annotation inconsistency across dialects. Such inconsistencies will be addressed in section 8.

7 Addressing the Data Imbalance

One drawback of using a joint model of the four dialects is that it is trained on only 25% examples of the target dialect, which means that dialect specific, correct decisions may be overruled by other dialects. In section 6.2, we showed that creating

Word	Original POS	New POS
Negation markers	PART	NEG_PART
Interrogatives	PART	ADV
Rel. Pronouns	PART	PRON
FUT and PROG markers	inconsistent	fixed
unmarked CONJ	PART/NOUN	CONJ
Adverbs	NOUN	ADV
Verbal suffixes	PRON	concatenated Verb and suffix
Nominal suffixes	NSUFF	concatenated Noun and suffix

Table 5: Annotation changes

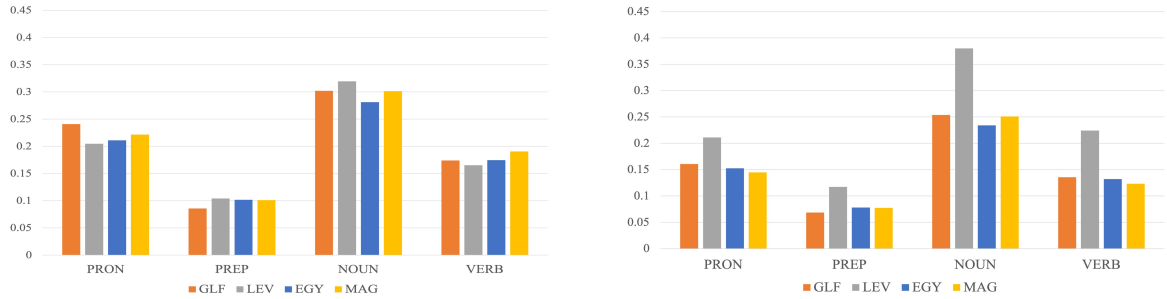

Figure 2: Distribution of POS tags across dialects before (left) and after (right) annotation changes. For instance, EGY has 16% PRON of all POS tags.

individual POS tagger models per dialect improves results. Here, we investigate whether we can use upsampling to further improve results. Upsampling is a standard method for handling data imbalance, for example in shared tasks on Arabic dialect identification (Zitouni et al., 2020; Habash et al., 2021) and for POS tagging non-standardized web data (Neunerdt et al., 2014; Horsmann and Zesch, 2015). For this approach, we duplicate instances of the target domain in the training data. For example, if the target domain is Egyptian, then in the training data (consisting of the four dialects), we duplicate the Egyptian examples, creating a more balanced training set by increasing the number of examples of the target class.

The results of this experiment are shown in Table 4, row 5. These results show an improvement overall across all dialects in comparison to the joint model (row 3) and the single dialect model. The best performance was achieved for LEV, its accuracy increased by 13.7 percent points over the joint model; for EGY, it increased by 12.8 percent points.

This shows that upsampling can successfully combine the advantages of having a large training set and a focused one.[11]

8 Annotation Changes

A closer examination of the POS tagger errors shows that in some cases, the problems derive from the gold annotations of the training data. Apart from the expected annotation errors due to lack of attention, which are mostly random, we also find more systematic inconsistencies, partly across dialects.

One such inconsistency concerns dialect-specific POS tags, such as negation, progressive, and future markers. For instance, in the GLF data, none of the negation markers were annotated with the negation-specific POS tag.

Systematic inconsistencies can potentially be corrected semi-automatically. To address the annotation inconsistencies, we further experiment with annotation changes on the Darwish corpus (our training data; Darwish et al. (2018)). We created a list of annotation inconsistencies, focusing on those which can be found and corrected automatically. We used the Camel Lab guidelines[12] as a reference since they provide specific and consistent POS tagging guidelines for dialects of Arabic. We performed systematic changes on the corpus while maintaining consistency across dialects. A list of

[11] We also explored tripling the number of samples for the target dialect, but this was less effective.

[12] https://camel-guidelines.readthedocs.io/en/latest/

Dialects	GLF	LEV	EGY	MAG
our baseline	59.5	61.3	68.3	61.4
baseline upsampled	72.8	75.0	81.1	74.2
on new annotation	82.3	75.2	80.2	73.8
new annotation upsampled	73.6	73.8	80.7	73.8

Table 6: Summary of POS tagging accuracy per dialect before and after annotation changes.

the targeted annotations in shown in Table 5.

For the distribution of POS tags in each dialect before and after the corrections, see Figure 2, focusing on the four most frequent POS tags. The plots show that in the original annotations, the ratios per POS tag are similar across dialects; after the corrections, there are more differences, showing that we model differences between dialects better. Note that the size of the corpus has changed due to the annotation changes, resulting in differences in POS tag distributions within dialects: We reattached the verb suffixes (previously tagged as PRON), for example, the verb يظهر (V) and the suffix وَا (PRON) are reattached into a single word يظهروَا. We also reattached nominal suffixes (previously tagged as NSUFF), such as صيدلي (Noun) and ة (NSUFF) into the single word صيدلية. As a consequence, the number of words decreases (e.g., the word count for Levantine decreases by 6%). Reattaching verbal suffixes also causes a decrease in pronouns across all dialects but Levantine shown in Figure 2. In LEV, relative pronouns which were originally tagged as PART are now categorized as PRON.

We then perform experiments training a joint model on all dialects after annotation modification, and test on each dialect separately to check whether the annotation changes boost the tagging performance.

Results are reported in Table 6. When comparing the results after modifying the annotations, we notice a considerable improvement in results over the baseline for all dialects, with increases ranging from 11.9% (EGY) to 22.8% (GLF). For GLF and LEV, the results on the improved annotations without upsampling even increase over the upsampled baseline (i.e., from 72.8% to 82.32% for GLF). We attribute this improvement to a higher consistency in the annotations. A comparison of the results on the improved annotations with and without upsampling shows that given the improved annotations, upsampling is less relevant or even harmful: The accuracy for EGY increases from 80.2% to 80.7% while the accuracy for GLF and LEV decreases (GLF: from 82.3% to 73.6%), and the accuracy for MAG remains stable. One explanation is that some words became more ambiguous as a result of the annotation changes. The word مَا, for example, was annotated inconsistently across dialects. It is ambiguous between a pronoun and a particle reading. However, in the original annotations, it was mostly annotated as particle. Another example is the negation marker: This POS tag was originally used in all dialects but Gulf. Additionally, the dialects use different words for negation, but not all were annotated as such. Now they are annotated consistently across dialects, which has changed the majority reading from pronoun or particle to negation marker.

9 Using Pretrained Word Embeddings

Next we investigate whether word embeddings can be beneficial and have a positive impact on the quality of POS tagging. The assumption is that the pre-trained word embeddings derived from large corpora of dialectal Arabic can help mitigate problems with lexical coverage in the randomly initialized embeddings in the out-of-domain setting.

The choice of the pretrained embeddings is important. We use the word embeddings trained on a large corpus of dialectal Arabic (Erdmann et al., 2019).

To train the embeddings, Erdmann et al. (2018) collected data for four major dialects of Arabic: Gulf, Levantine, Egyptian, and Maghrebi, which cover the four dialects of our test data. The corpora are a mix of crawled data from a variety of forums and blogs, including comments on posts (Almeman and Lee, 2013; Khalifa et al., 2016; Zbib et al., 2012), MADAR (Bouamor et al., 2018), news commentary corpus (Zaidan and Callison-Burch, 2011), tweets from the corpus of Palestinian Arabic (Jarrar et al., 2014), with the number of sentences per dialect ranging between 1.1M and 1.7M. The model was trained using fastText (Bojanowski

	original	+ embeddings
vectors	1 998	2 134 625
dimension	300	400
window size	10	10
batch size	128	128

Table 7: Embedding layer parameters

	GLF	LEV	EGY	MAG
Without embeddings	82.3	75.2	80.2	73.8
With embeddings	83.2	84.3	87.9	78.9

Table 8: Accuracy of POS tagging with and without using pre-trained embeddings using improved annotations.

et al., 2017)[13]. Since MADAR is part of the training data for the embeddings, we can expect a higher lexical coverage for the test data.

Table 8 shows the results for POS tagging with and without using the pre-trained embedding and the improved annotations. The results show that the performance on all dialects increases, and for all but GLF the gains are considerable, LEV gains the most: For this dialect, the accuracy increases from 75.2% to 84.3%. For GLF, we see a moderate increase from 82.3% to 83.2%. This dialect had the highest accuracy before embeddings, as it has the highest lexical overlap with the training corpus.

We also had a look at the tagging errors for the model using the pre-trained embeddings. A heatmap of POS tag precision is provided in Figure 3. We see that numbers are still the most difficult POS tag, similar to the results in Figure 1. However, for all dialects but MAG, the accuracies are considerably higher. For MAG, most of the numbers were mistagged as NOUN. This seems to be due to inconsistencies in the training data. Since the spelling of numbers tends to differ between dialects, the POS tagger cannot learn from the other dialects. The same pattern of gains holds for particles, previously the second most difficult category, except for MAG. The current second most difficult POS tag are adjectives. Here we see a decrease over all dialects in comparison to Figure 1. This can be explained by the systematic ambiguity between nouns and adjectives. The POS tagger seems to favor annotating these ambiguous words as adjectives, which leads to a high precision for nouns,

[13]We do not use BERT embeddings since they cannot be easily integrated into the POS tagger architecture. See Table 7 for embedding parameters.

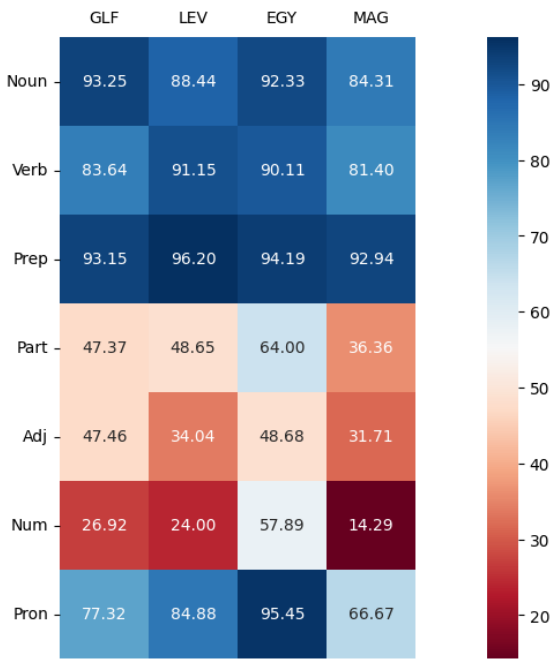

Figure 3: Results per dialect (precision) when using pre-trained embeddings.

and a low one for adjectives.

For instance, the adjective آسف (Eng.: sorry) is tagged as NOUN because of its alternative interpretation 'regret'.

10 Conclusion and Future Work

We have investigated POS tagging for Arabic dialects when the test set is out-of-domain. This setting has proven to be difficult, originally resulting in a low accuracy. Our work shows that we can improve the POS tagger's accuracy by upsampling the target dialect in the training data, by increasing consistency of annotations, and by using word embeddings pre-trained on a large corpus of dialectal Arabic. On average we have seen improvements of about 20 percent points.

Our overarching goal is the investigation of morpho-syntactic and syntactic differences between Arabic dialects. Our next step is to experiment with the granularity of POS tags. The current small POS tagset may not provide enough information for an investigation of syntactic differences. We also plan to develop a parsing model for Arabic dialects.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious

segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16, San Diego, CA.

Faisal Al-Shargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. Morphologically annotated corpora and morphological analyzers for Moroccan and Sanaani Yemeni Arabic. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Abdullah I. Alharbi and Mark Lee. 2020. BhamNLP at SemEval-2020 task 12: An ensemble of different word embeddings and emotion transfer learning for Arabic offensive language identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1532–1538, Barcelona (online). International Committee for Computational Linguistics.

Randah Alharbi, Walid Magdy, Kareem Darwish, Ahmed Abdelali, and Hamdy Mubarak. 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)*, Miyazaki, Japan.

Khalid Almeman and Mark Lee. 2013. Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. In *1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6, Sharjah, UAE.

Mohammed Attia and Ali Elkahky. 2019. Segmentation for domain adaptation in Arabic. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP)*, pages 119–129, Florence, Italy.

Shabnam Behzad and Amir Zeldes. 2020. A cross-genre ensemble approach to robust Reddit part of speech tagging. In *Proceedings of the 12th Web as Corpus Workshop*, pages 50–56, Marseille, France. European Language Resources Association.

Bernd Bohnet, Ryan McDonald, Goncalo Simoes, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2642–2652, Melbourne, Australia.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan.

Kareem Darwish, Mohammed Attia, Hamdy Mubarak, Younes Samih, Ahmed Abdelali, Lluís Màrquez, Mohamed Eldesouki, and Laura Kallmeyer. 2020. Effective multi-dialectal Arabic POS tagging. *Natural Language Engineering*, 26(6):677–690.

Kareem Darwish and Hamdy Mubarak. 2016. Farasa: A new fast and accurate Arabic word segmenter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 1070–1074, Portorož, Slovenia.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, and Mohamed Eldesouki. 2017. Arabic POS tagging: Don't abandon feature engineering just yet. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 130–137, Valencia, Spain.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-dialect Arabic POS tagging: A CRF approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan.

Mona Diab and Nizar Habash. 2007. Arabic dialect processing tutorial. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 5–6, New York, NY.

Kevin Duh and Katrin Kirchhoff. 2005. POS tagging of dialectal Arabic: A minimally supervised approach. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 55–62, Ann Arbor, MI.

Mohamed Eldesouki, Younes Samih, Ahmed Abdelali, Mohammed Attia, Hamdy Mubarak, Kareem Darwish, and Laura Kallmeyer. 2017. Arabic multi-dialect segmentation: bi-LSTM-CRF vs. SVM. *arXiv preprint arXiv:1708.05891*.

Ashraf Elnagar, Sane M Yagi, Ali Bou Nassif, Ismail Shahin, and Said A Salloum. 2021. Systematic literature review of dialectal Arabic: Identification and detection. *IEEE Access*, 9:31010–31042.

Alexander Erdmann, Salam Khalifa, Mai Oudah, Nizar Habash, and Houda Bouamor. 2019. A little linguistics goes a long way: Unsupervised segmentation with limited language specific guidance. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 113–124, Florence, Italy.

Alexander Erdmann, Nasser Zalmout, and Nizar Habash. 2018. Addressing noise in multidialectal word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 558–565, Melbourne, Australia.

Ali Farghaly and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):1–22.

Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb, editors. 2021. *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, Kyiv, Ukraine (Virtual).

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal Arabic. In *Proceedings of theConference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 426–432, Atlanta, GA.

Benjamin Heinzerling and Michael Strube. 2019. Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 273–291, Florence, Italy.

Tobias Horsmann and Torsten Zesch. 2015. Effectiveness of domain adaptation approaches for social media PoS tagging. In *Proceeding of the Second Italian Conference on Computational Linguistics*, pages 166–170, Trento, Italy.

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. Building a corpus for Palestinian Arabic: A preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 18–27, Doha, Qatar.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. *arXiv preprint arXiv:1609.02960*.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of Emirati Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Sandra Kübler and Eric Baucom. 2011. Fast domain adaptation for part of speech tagging for dialogues. In *Proceedings of the International Conference on Recent Advances in NLP (RANLP)*, Hissar, Bulgaria.

Prathyusha Kuncham, Chandu Khyathi Raghavi, Kovida Nelakuditi, and Dipti Misra Sharma. 2014. Domain adaptation in morphological analysis. *International Journal of Languages, Literature and Linguistics*, 1(2).

Atreyee Mukherjee, Sandra Kübler, and Matthias Scheutz. 2017. Creating POS tagging and dependency parsing experts via topic modeling. In *Proceedings of Fifteenth Conference of the European Chapter of the ACL (EACL)*, Valencia, Spain.

Melanie Neunerdt, Michael Reyer, and Rudolf Mathar. 2014. Efficient training data enrichment and unknown token handling for POS tagging of non-standardized texts. In *Conference on Natural Language Processing (KONVENS)*, pages 186–192, Hildesheim, Germany.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Younes Samih, Mohammed Attia, Mohamed Eldesouki, Ahmed Abdelali, Hamdy Mubarak, Laura Kallmeyer, and Kareem Darwish. 2017a. A neural architecture for dialectal Arabic segmentation. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 46–54, Valencia, Spain.

Younes Samih, Mohamed Eldesouki, Mohammed Attia, Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak, and Laura Kallmeyer. 2017b. Learning from relatives: Unified dialectal Arabic segmentation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 432–441, Vancouver, Canada.

Abdulhadi Shoufan and Sumaya Alameri. 2015. Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. *International Journal of Computational Linguistics & Chinese Language Processing: Special Issue on Invited Papers from ISCSLP 2006*, 12(3):303–324.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Automated concatenation of embeddings for structured prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2643–2660, Online.

Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 976–986, New Orleans, Louisiana. Association for Computational Linguistics.

Omar Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: An annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland OR.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada.

Imed Zitouni, Muhammad Abdul-Mageed, Houda Bouamor, Fethi Bougares, Mahmoud El-Haj, Nadi Tomeh, and Wajdi Zaghouani, editors. 2020. *Proceedings of the Fifth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, Barcelona, Spain (Online).

Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith, and Philippe Blache. 2017. Morphological disambiguation of Tunisian dialect. *Journal of King Saud University - Computer and Information Sciences*, 29(2):147–155.