Multi-Dialect Arabic POS Tagging: A CRF Approach

Multi-Dialect Arabic POS Tagging: A CRF Approach

Kareem Darwish*, Hamdy Mubarak*, Mohamed Eldesouki*, Ahmed Abdelali*,
Younes Samih†, Randah Alharbi‡, Mohammed Attia⋆, Walid Magdy‡, Laura Kallmeyer†
*QCRI, †University of Dusseldorf, ‡University of Edinburgh, ⋆Google Inc.
{kdarwish,hmubarak,mohamohamed,aabdelali}@hbku.edu.qa,
samih@phil.hhu.de, s1581951@sms.ed.ac.uk, attia@google.com,
wmagdy@inf.ed.ac.uk, kallmeyer@phil.hhu.de

Abstract

This paper introduces a new dataset of POS-tagged Arabic tweets in four major dialects along with tagging guidelines. The data, which we are releasing publicly, includes tweets in Egyptian, Levantine, Gulf, and Maghrebi, with 350 tweets for each dialect with appropriate train/test/development splits for 5-fold cross validation. We use a Conditional Random Fields (CRF) sequence labeler to train POS taggers for each dialect and examine the effect of cross and joint dialect training, and give benchmark results for the datasets. Using clitic n-grams, clitic metatypes, and stem templates as features, we were able to train a joint model that can correctly tag four different dialects with an average accuracy of 89.3%.

Keywords: Arabic dialects, POS tagging, CRF

1. Introduction

Part-of-speech (POS) tagging is important for a variety of applications such as parsing, information extraction, and machine translation. Though much work has focused on POS tagging of Modern Standard Arabic (MSA), work on Dialectal Arabic (DA) POS tagging is rather scant with POS tagged corpora for most dialects being nonexistent or of limited availability. Dialectal POS tagging is becoming increasingly important due to the ubiquity of social media, where users typically write in their dialects to match how they speak in their daily interactions. Dialectal text poses interesting challenges such as lack of spelling standards, pervasiveness of transformative morphological operations, such as word merging and letter substitution or deletion, in addition to lexical borrowing from foreign languages. Existing work on dialectal POS tagging focuses on building resources and tools for each dialect separately (Duh and Kirchhoff, 2005; Habash et al., 2013). The rationale for the separation is that different dialects have different affixes, make different lexical and word ordering choices, and are influenced by different foreign languages. However, performing reliable dialect identification to properly route text to the appropriate POS tagger may be problematic, because conventional dialectal identification may lead to results lower than 90% (Darwish et al., 2014). Thus, building a POS tagger that performs reliably across multiple dialects without the need for dialect identification is desirable.

In this paper, we present new POS-tagging annotations on a dialectal dataset that is composed of social media text from Twitter for four major Arabic dialects, namely Egyptian (EGY), Levantine (LEV), Gulf (GLF), and Maghrebi (MGR) (Eldesouki et al., 2017; Samih et al., 2017a). For each dialect, we tagged 350 tweets using an extended version of the Farasa tagset (Darwish et al., 2017). We extended the tagset to account for tweet-specific tokens, namely hashtags, user mentions, emoticons and emojis, and URL's. We created 5-fold partitions for cross-validation with 70/10/20 train/dev/test splits for each dialect. We used the new dataset to train dialectal POS taggers for each dialect separately to test the effectiveness of the taggers on test data from the same dialect or from different dialects. We also experimented with cross-dialect and joint training to see if POS tagging of one dialect can benefit from data from other dialects. We show that joint models can perform on average at par with dialect specific models. For all our experiments we used a Conditional Random Fields (CRF) sequence labeler.

The contributions of this paper are as follows:

1. We present new dialectal POS tagging annotations on a multi-dialectal tweet dataset.
2. We report benchmark results on DA POS tagging.
3. We show that we can develop an effective joint model for POS tagging of different dialects without the need for dialect identification or dialect specific models.

2. Background

The scarcity of dialectal resources have hampered research in the area of DA, despite efforts from large institutions (ex. LDC) and programs (ex. TIDES, GALE and BOLT). Limited resources were made available for researchers with limited size and coverage. CallHome Egyptian Colloquial Arabic (ECA)[1] was the first attempt for a corpus to address this shortage released in 1997. The corpus is a collection of transcripts that cover five to ten minute segments taken from 120 unscripted telephone conversations between native speakers of Egyptian Arabic. Levantine Colloquial Arabic[2] as well as the Iraqi Arabic Conversational Telephone Speech[3] are two additional resources to cover dialectal Arabic that were built between 2004 and 2006 (Maamouri et al., 2004). The most recent dataset in this

[1] https://catalog.ldc.upenn.edu/LDC97T19
[2] https://catalog.ldc.upenn.edu/LDC2005S14
[3] https://catalog.ldc.upenn.edu/LDC2006T16

series was BOLT Egyptian Arabic SMS/Chat and Transliteration dataset[4], which is considerably the largest resource for dialectal Arabic (over a million words) even though it covers only Egyptian. The availability of these resources is stinted given the license requirement. On the other hand, researchers used ad-hoc resources or small datasets that were curated locally and not widely available. Graja et al. (2010) created the Tunisian Dialect Corpus Interlocutor (TuDiCoI) which contains 893 utterances, 3,404 words from dialectal conversations between Tunisian railway staff. Bouamor et al. (2014) used a collection of 2,000 sentences in Egyptian dialect as a seed to build a multi-dialectal Arabic corpus. The seed sentences were translated by native speaker to their own dialects to create a parallel corpus of Standard Arabic, Egyptian, Tunisian, Jordanian, Palestinian and Syrian Arabic, in addition to English. Cotterell and Callison-Burch (2014) extended the work of Al-Sabbagh and Girju (2010) and Zaidan and Callison-Burch (2011) to build a larger collection of commentaries from five Arabic newspapers and tweets for automatic dialect identification. Duh and Kirchhoff (2005) used ECA to build a POS tagger for Egyptian with the support of the MSA ATB. An accuracy of 69.83% with a coverage of 74.10% was achieved. Habash et al. (2013) released a new adaptation for MADA (Roth et al., 2008), extending it to cover Egyptian as well.

The emergence of social media platforms and their support to languages other than Latin -bidirectional/left-to-right or right-to-left helped dialectal presence to be more apparent than ever. This created a new need for newer resource with wider coverage. Our work attempts to fill some of this gap by providing a collection of tweet data that covers 4 major Arabic dialects as well as POS annotation of the data, which is unique and a first to be open for researchers.

3. Data Description

We used the dialectal Arabic dataset described by Eldesouki et al. (2017) and Samih et al. (2017b), which includes a set of 350 tweets for four major Arabic dialects that were manually segmented. The size of the dataset is as follows:

Dialect	No of Tweets	No of Words
Egyptian (EGY)	350	7,481
Levantine (LEV)	350	7,221
Gulf (GLF)	350	6,767
Maghrebi (MGR)	350	6,400

The words in the dataset were segmented in place without any modification or standardization attempts (ex. CODA (Habash et al., 2012)), and the segmentation guidelines aimed to generate a number of segments that match the correct number of POS tags for a word.

We used the POS tagset described by Darwish et al. (2017) which has 18 tags for MSA POS tagging, and we added 2 dialect-specific tags (namely PROG_PART, and NEG_PART), and 4 tweet-specific tags (namely HASH, EMOT, MENTION, and URL). Table 1 contains description of the newly added tags[5].

[4]https://catalog.ldc.upenn.edu/LDC2017T07

[5]Buckwalter transliteration is used in the paper

POS	Description	Example
PROG_PART	Progressive Part.	بنكتب (**b**nktb) "we are writing"
NEG_PART	Negation Part.	ماكانش (mAkAn**$**) "he was not"
HASH	Hashtag	#يا_رب (#yA_rb) "#O_God"
EMOT	Emoticon/Emoji	:)
MENTION	Mention	@mohamedAli
URL	URL	http://t.co/EF5cW

Table 1: Dialect-specific and tweet-specific POS tags

Segmentation and POS tagging were applied on the original raw text without any correction as suggested by Eldesouki et al. (2017) to overcome the need for standardization of different dialectal writings proposed in CODA by Habash et al. (2012). For example the word ومبيقولوش (wmbyqwlw$) "and they are not saying" is segmented as و+م+ب+يقول+و+ش (w+m+b+yqwl+w+$) and tagged as: CONJ+PART+PROG_PART+V+PRON+NEG_PART.

Words are white-space and punctuation separated while hashtags, emotions, mentions and URL's are considered as single words without internal segmentation. Data is formatted in CoNLL format: Words are split into tokens (clitics), and POS is provided for each token. In our annotation scheme, tokens, words, and sentences are separated by token boundary tag (TB), word boundary tag (WB), and end of sentence tag (EOS) respectively as shown in Table 2. Tagging was performed by a native speaker for each dialect. Then, multiple rounds of quality control and revision were performed to obtain high accuracy and consistency across dialects.

Index	Token	POS
0	ب (b) (present cont. particle)	PROG_PART
0	TB	TB
0	حب (Hb) "I love"	V
0	WB	WB
1	اسمع (AsmE) "I listen"	V
1	WB	WB
..	..	..
n	EOS	EOS

Table 2: Data format for segmentation and POS tagging

Figure 1 compares the distribution of POS tags in the four dialects against a sample of 350 MSA sentences from ATB with similar number of words (7,385 words). From this figure, some interesting observations can be made. For example, the four dialects are generally similar to each other in their POS distribution, while MSA shows substantial divergence. For example, nouns, adjectives, prepositions, numbers, and definite articles appear more frequently in MSA than in dialects, while on the other hand dialects show higher frequency of verbs, pronouns and particles. Our justification for this noticeable disparity is that the POS distribution is affected by the genre. The MSA text is from the formal news domain with a special focus on facts and entities, while the dialects are informal expressions with a focus on events, attitudes, and conversations. Another observa-

tion is that MSA has more noun suffixes and grammatical case endings, while dialects have more progressive particles and negation suffixes. This variance is related more to the linguistic nature of the language rather than the genre.

4. Experiments and Evaluation

4.1. Experimental Setup

For the experiments that we conducted, we used the CRF++ implementation of a CRF sequence labeler with L2 regularization and default value of 10 for the generalization parameter "C". We conducted three sets of experiments.

In the **first**, we tested the effectiveness of different features including different size contexts, metatypes, and stem templates, which we describe later.

In the **second**, we used the training and dev parts for each split for every dialect for training and then we used the test parts for all dialects testing. In these experiments, we were interested in knowing the POS tagging effectiveness when training and testing data are from the same dialect or from different dialects. High results for cross dialect training and testing may indicate closeness between dialects.

In the **third** set of experiments, we trained on all the train and dev parts of all the dialects jointly and tested on test sets of all the dialects. We were interested in determining if POS tagging from one dialect can benefit from added training data from other dialects.

For our experiments, given a sequence of clitics $c_n...c_{-2}, c_{-1}, c_0, c_1, c_2...c_m$, where we assumed perfect segmentation, we used the following features for each clitic (c_0):

- **Clitic n-grams.** a combination of clitic unigram features $\{c_{-1};c_0;c_1\}$, bigram features $\{c_{-2}^{-1};c_{-1}^{0};c_0^1;c_1^2\}$, and alternatively trigram features $\{c_{-2}^{0};c_{-1}^{1};c_0^2\}$ and 4-gram features $\{c_{-3}^{0};c_{-2}^{1};c_{-1}^{2};c_0^3\}$.

- **Clitic metatypes.** We defined a set of 10 "metatypes" that we heuristically determined. They include: Hashtag (if clitic starts with "#"); Mention (if clitic starts with "@"); URL (if clitic starts with "http"); Emoticon/emoji (if it appears in a list of 2,730 emoticons/emojis that we constructed); Retweet (if the clitic is "RT"); Foreign (if it contains non-Arabic letters); Number (if it matches Arabic or Hindi numerals or a gazetteer of written out numbers that we obtained from Farasa (Abdelali et al., 2016)); Punctuation (if it matches punctuations in the UTF8 codepage); Arabic (if it contains Arabic letters only); and Other for all other clitics. Using metatypes was shown to be effective for MSA POS tagging (Darwish et al., 2017).

- **Clitic stem templates.** Arabic words are typically derived from a closed set of roots that are placed in so-called stem templates to generate stems. For example, the root *ktb* can be fit in the template *CCAC* to generate the stem *ktAb* (book). Stem templates may overwhelmingly have one POS tag (e.g., *yCCC* is overwhelmingly a V) or favor one tag over another (e.g., *CCAC* is more likely a NOUN than an ADJ). This was shown to be effective for MSA POS tagging (Darwish et al., 2017), and we were curious to see if this would be effective for dialects also, particularly given the overlap between MSA and dialectal Arabic. We used Farasa to determine stem templates (Abdelali et al., 2016).

For all the experiments, we trained on the training and dev parts and tested on the test part. As mentioned earlier, we also randomly selected 350 MSA sentences from Arabic Penn Treebank (ATB) and treated MSA as a language variety. Doing so would allow us to observe the divergence of dialects from MSA and the relative effectiveness of using a small dataset compared to much more data.

4.2. Evaluation

We conducted the the following sets of experiments:

Set 1: In this set, we examined the effectiveness of the different features by when training and testing on the same dialect. We evaluated the following feature sets:

- Baseline (BL): clitic n-grams only, where we used the aforementioned combination of clitic unigrams, bigrams, and trigrams.
- Baseline + stem template (+ST)
- Baseline + metatype (+MT)
- A combination of clitic unigrams and bigrams + stem template + metatype (+ST+MT (2g))
- A combination of clitic unigrams, bigrams, and trigrams + stem template + metatype (+ST+MT (3g))
- A combination of clitic unigrams, bigrams, trigrams, and 4-grams + stem template + metatype (+ST+MT (4g))

Table 3 reports on the word-level accuracy for the different experiments. To demonstrate the difference between word-level accuracy, which we report here, and clitic-level accuracy, consider the phrase هـ+يلعب ب+ال+كورة (h+ylEb b+Al+kwr+m – "he will play with the ball") with correct POS tags FUTURE_PART+V PREP+DET+NOUN+NSUFF. If tagger erroneously tagged the phrase as FUTURE_PART+V PREP+DET+ADJ+NSUFF, then word-level accuracy would be 1/2 while clitic level accuracy would be 5/6. Since we used 5-fold cross validation, we report on the average across all folds. As the results show, using clitic n-grams only yielded the lowest results. Using stem template and metatype features improved results over using clitic n-grams alone with the combination of both features leading to even greater gain. When combined with stem template and metatype features, a combination of clitic unigrams and bigrams (2g) yielded the best results edging the use of higher order n-grams.

Set 2: Next, we were interested in determining cross-dialect training results to see if dialects can learn from each other and whether models from one dialect can generalize to other dialects. For all experiments, we used a combination of clitic unigrams and bigrams with stem template and metatype features (+ST+MT (2g)). Table 4 reports on cross-dialect results. Not surprisingly, the best results for

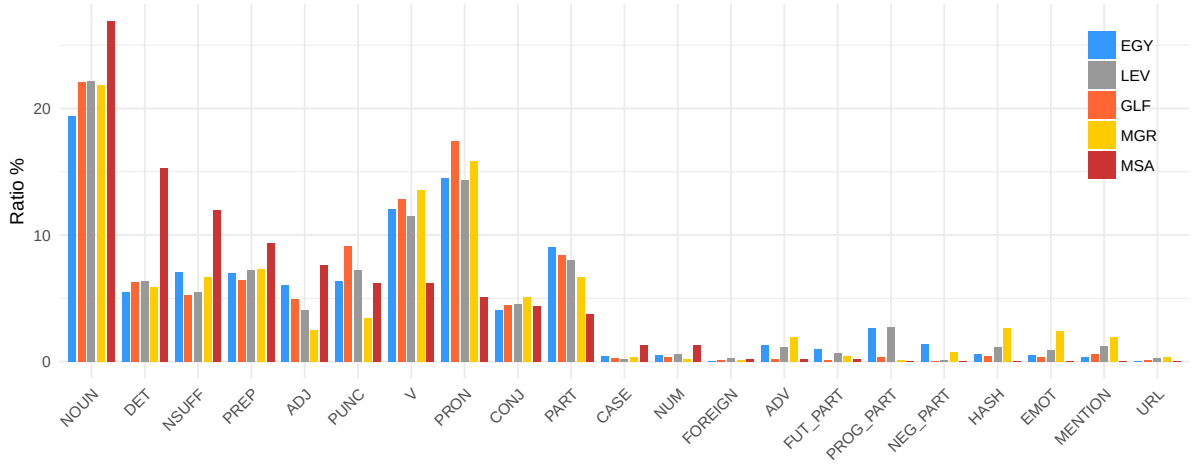

Figure 1: Distribution of POS tags per dialect and MSA. Tags are ordered according to average usage in MSA and DA

each dialect were obtained when the training and test sets were from the same dialect. Among the different dialects (excluding MSA), Maghrebi suffered the most when the training set was from another dialect, and conversely training on Maghrebi yielded the worst results for all the other dialects. This may indicate that Egyptian, Levantine, and Gulf were closer together and Maghrebi was most dissimilar to all of them. Further, training on MSA and testing on dialects yielded significantly lower results compared to training on dialects and testing on MSA. This may imply that the affixes that we observed in dialects are a superset of those observed in MSA.

Set 3: Lastly, we were curious to see if dialects can benefit from the addition of data from other dialects during training. Thus we combined the training and dev parts for all dialects and tested on the test part of each dialect. Table 5 reports on the results of joint learning with and without the inclusion of MSA. As the results show, Egyptian and Levantine benefited from the additional training data, while Maghrebi, Gulf, and MSA did not. The difference in accuracy (either positive or negative) ranged between 0.2% and 0.7%. On average across dialects only, the addition of MSA data marginally affected POS tagging effectiveness for different dialects. We suspect that if we use MSA tweets, instead of news sentences that we obtained from ATB, to match the genre of the dialect data would lead to greater improvement. Having a joint model that performs at par or better than dialect specific models across dialects is highly advantageous as it would avoid the need for dialect identification. The results show that a joint model may outperform dialect-specific models.

5. Error Analysis

To evaluate the strengths and weaknesses of our system, we analyzed the predicted results and the various types of errors made by the system. For such, we assessed the top 10 error types for each dialect and MSA for which correct POS of whole words is different than the guessed POS by the system. These errors represent 74%, 70%, 78%, 72% and 85% of all errors in EGY, LEV, GLF, MGR and MSA respectively. Next, we compiled together these errors and sorted them according to their average. Results are shown in Figure 1. On the average, 50% of all errors in DA are due

	MSA	EGY	LEV	GLF	MGR
BL	90.0	88.8	80.8	81.9	81.7
+ST	92.8	91.4	84.3	85.9	84.2
+MT	90.7	90.3	84.5	83.5	86.1
+ST+MT (2g)	**93.6**	**92.9**	**87.9**	**87.8**	**88.3**
+ST+MT (3g)	93.1	92.4	87.5	87.3	88.0
+ST+MT (4g)	92.5	91.6	86.8	86.6	87.2

Table 3: Per dialect training: baseline (BL), stem templates (+ST), and metatypes (+MT), and combined (+ST+MT) varying clitic n-grams (2g, 3g, and 4g).

	Training Set				
Test Set	MSA	EGY	LEV	GLF	MGR
MSA	**93.6**	76.1	76.1	76.9	72.7
EGY	54.5	**92.9**	74.3	78.1	72.7
LEV	52.0	74.7	**87.9**	73.5	69.8
GLF	58.7	78.8	76.7	**87.8**	74.4
MGR	50.8	71.1	73.1	70.1	**88.3**
Avg	61.9	**78.7**	77.6	77.3	75.6

Table 4: Cross dialect training using clitic bigrams (2g), stem templates, and metatypes as features.

to incorrect classification of nouns as verbs or adjectives and vice versa. This ratio increases to 65% in MSA.

Table 6 shows examples of the top error types across all dialects and MSA which represent 71% of all errors. Some of these errors are due to the fact that the words were not seen in training data. However, many of these words, such as منطقية (mnTqyp – "logical") and شوارع ($wArE – "streets"), are words that overlap between MSA and dialects and would exist in a large MSA corpus, such as ATB. Habash and Rambow (2006) and Mubarak (2017) reported an overlap of 60% for LEV verbs and 66% for EGY respectively. This suggests that domain adaptation with MSA data or using word embeddings that are trained on a large Arabic corpus would help overcome such errors.

Another source of errors is due to the lack of writing standards and handling words without correcting their spelling mistakes as in words وإسم (w<sm – "and name") and مأساه (m>sAh – "tragedy").

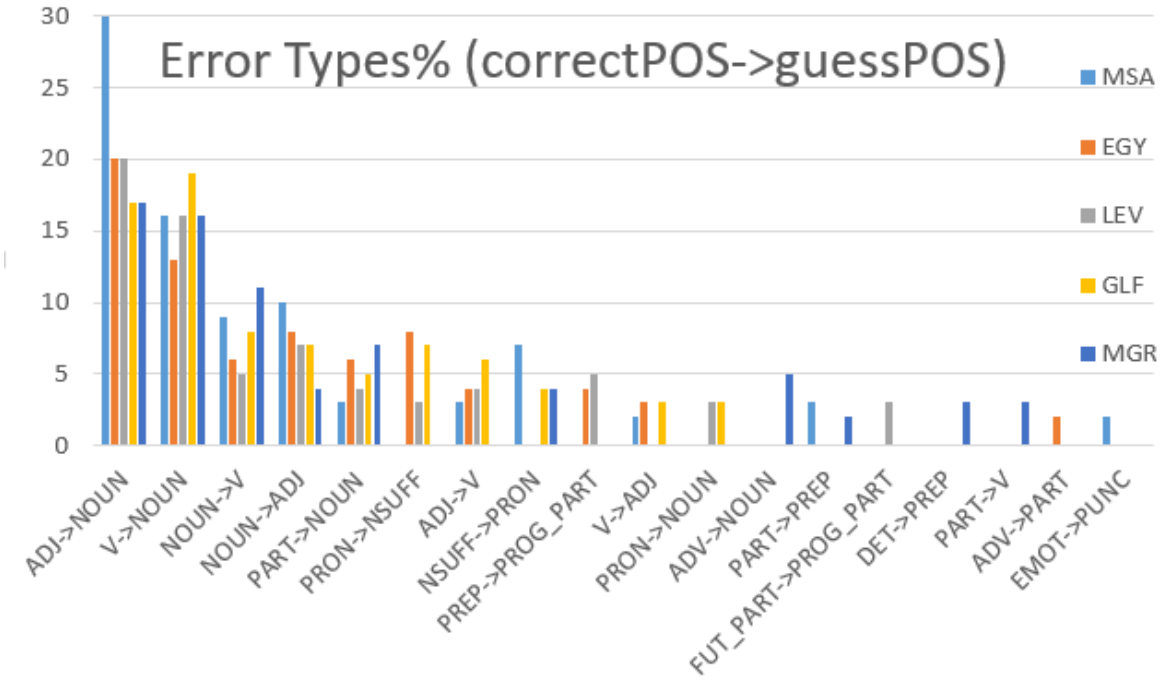

Figure 2: Distribution of Error types

	on self	joint (DA only)	joint (MSA&DA)
MSA	**93.6**	82.5	92.5
EGY	92.9	93.2	**93.4**
LEV	87.9	**88.6**	**88.6**
GLF	**87.8**	87.2	87.4
MGR	**88.3**	87.7	87.6
Avg (DA only)	89.2	89.2	**89.3**
Avg (MSA&DA)	**90.1**	87.8	89.9

Table 5: Results of joint learning

6. Conclusion and Future Work

In this paper, we introduced a new dataset for POS tagging of four major Arabic dialects that we constructed from tweets. We plan to provide the data freely to the community including our training, dev, and test splits. We also built POS taggers for the four dialects using a CRF sequence labeler using clitic n-gram features, stem templates, and clitic metatypes. Further, we show that we can train a joint model using data from all the dialects to train a POS tagger with comparable results to mono-dialectal training and testing, alleviating the need for dialect identification prior to POS tagging.

For future work, we plan to explore two distinct directions, namely:

- the use Brown clusters (Brown et al., 1992). Brown clustering is a hierarchical clustering of words based on their context and produces a kind of word embeddings that can be learned from large unlabeled texts. The rationale for using it here is that similar words, particularly those that share the same POS tag, tend to appear in similar contexts (Owoputi et al., 2013; Stratos and Collins, 2015).
- the use of Deep Neural Networks (DNN). DNNs have the advantage of alleviating the need for specific feature engineering including long distance relationships. Further, character-level models may be able to learning morphological patterns automatically.

7. Bibliographical References

Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H. (2016). Farasa: A fast and furious segmenter for arabic. In *HLT-NAACL Demos*, pages 11–16.

Al-Sabbagh, R. and Girju, R. (2010). Mining the web for the induction of a dialectical arabic lexicon. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 288–293.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A multidialectal parallel corpus of arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1240–1245.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Cotterell, R. and Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 241–245.

Darwish, K., Sajjad, H., and Mubarak, H. (2014). Verifiably effective arabic dialect identification. In *EMNLP*, pages 1465–1468.

Darwish, K., Mubarak, H., Abdelali, A., and Eldesouki, M. (2017). Arabic pos tagging: Dong't abandon feature en-

Error Type	Avg%	Example (Dialect)
ADJ->NOUN	21	غير منطقية (gyr mnTqyp) “not logical” (EGY)
V->NOUN	16	أعتقد ما في مجال (AEtqd mA fy mjAl) “I believe there is...” (LEV)
NOUN->V	8	لا شوارع (lA $wArE) “no streets” (GLF)
NOUN->ADJ	7	المغرب والجزائر خاوة (Almgrb wAljzAyr xAwp) “Morocco and Algeria are brothers” (MGR)
PART->NOUN	5	مثلما يستطيع الكثير (mvlmA ystTyE Alkvyr) “as many can do” (MSA)
PRON->NSUFF	4	عاجبه حالـه (Ejbh hAl+h) “likes his circumstances” (EGY)
ADJ->V	3	احنا فاهمين (AHnA fAhmyn) “we understand” (LEV)
NSUFF->PRON	3	عايشين مأساه (Ey$yn m>sA+h) “living in a tragedy” (EGY)
PREP->PROG-PART	2	عاملها بهايدك (EAmlhA b+Aydk) “you did it by your hand” (LEV)
V->ADJ	2	كي يصير (ky ySyr) “to become” (MSA)

Table 6: Error types examples from different DAs and MSA

gineering just yet. *WANLP 2017 (co-located with EACL 2017)*, page 130.

Duh, K. and Kirchhoff, K. (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. Association for Computational Linguistics.

Eldesouki, M., Samih, Y., Abdelali, A., Attia, M., Mubarak, H., Darwish, K., and Kallmeyer, L. (2017). Arabic multi-dialect segmentation: bi-lstm-crf vs. svm. *arXiv:1708.05891v1 [cs.CL]*.

Graja, M., Jaoua, M., and Hadrich Belguith, L. (2010). Lexical study of a spoken dialogue corpus in tunisian dialect. In *The international arab conference on information technology (ACIT), benghazi–libya*.

Habash, N. and Rambow, O. (2006). Magead: a morphological analyzer and generator for the arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 681–688. Association for Computational Linguistics.

Habash, N., Diab, M. T., and Rambow, O. (2012). Conventional orthography for dialectal arabic. In *LREC*, pages 711–718.

Habash, N., Roth, R., Rambow, O., Eskander, R., and Tomeh, N. (2013). Morphological analysis and disambiguation for dialectal arabic. In *Hlt-Naacl*, pages 426–432.

Maamouri, M., Buckwalter, T., and Cieri, C. (2004). Dialectal arabic telephone speech corpus: Principles, tool design, and transcription conventions. In *NEMLAR International Conference on Arabic Language Resources and Tools, Cairo*, pages 22–23.

Mubarak, H. (2017). Analysis and quantitative study of egyptian dialect on twitter. In *17TH INTERNATIONAL CONFERENCE ON WEB ENGINEERING*. ICWE 2017.

Owoputi, O., O’Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT 2013*, pages 380–390. Association for Computational Linguistics.

Roth, R., Rambow, O., Habash, N., Diab, M., and Rudin, C. (2008). Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. *Proceedings of the Conference of American Association for Computational Linguistics (ACL08)*.

Samih, Y., Attia, M., Eldesouki, M., Mubarak, H., Abdelali, A., Kallmeyer, L., and Darwish, K. (2017a). A neural architecture for dialectal arabic segmentation. *WANLP 2017 (co-located with EACL 2017)*, page 46.

Samih, Y., Eldesouki, M., Attia, M., Darwish, K., Abdelali, A., Mubarak, H., and Kallmeyer, L. (2017b). Learning from relatives: Unified dialectal arabic segmentation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 432–441.

Stratos, K. and Collins, M. (2015). Simple semi-supervised pos tagging. In *VS@ HLT-NAACL*, pages 79–87.

Zaidan, O. F. and Callison-Burch, C. (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: short papers-Volume 2*, pages 37–41. Association for Computational Linguistics.