# Towards POS Tagging for Arabic Tweets

**Fahad Albogamy**
School of Computer Science,
University of Manchester,
Manchester, M13 9PL, UK
albogamf@cs.man.ac.uk

**Allan Ramasy**
School of Computer Science,
University of Manchester,
Manchester, M13 9PL, UK
allan.ramsay@cs.man.ac.uk

## Abstract

Part-of-Speech (POS) tagging is a key step in many NLP algorithms. However, tweets are difficult to POS tag because there are many phenomena that frequently appear in Twitter that are not as common, or are entirely absent, in other domains: tweets are short, are not always written maintaining formal grammar and proper spelling, and abbreviations are often used to overcome their restricted lengths. Arabic tweets also show a further range of linguistic phenomena such as usage of different dialects, romanised Arabic and borrowing foreign words. In this paper, we present an evaluation and a detailed error analysis of state-of-the-art POS taggers for Arabic when applied to Arabic tweets. The accuracy of standard Arabic taggers is typically excellent (96-97%) on Modern Standard Arabic (MSA) text ; however,their accuracy declines to 49-65% on Arabic tweets. Further, we present our initial approach to improve the taggers' performance. By making improvements based on observed errors, we are able to reach 74% tagging accuracy.

## 1 Introduction

The last few years have seen an enormous growth in the use of social networking platforms such as Twitter in the Arab World[1]. There are millions of tweets daily, yielding a corpus which is noisy and informal, but which is sometimes informative. Tweets are short and contain a maximum of 140 characters. Tweets also are not always written maintaining formal grammar and proper spelling. They are ambiguous and rich in acronyms. Slang and abbreviations are often used to overcome their restricted lengths. POS tagging is an essential processing step in a wide range of high level text processing applications such as information extraction, machine translation and sentiment analysis. However, people working on Arabic tweets have tended to concentrate on low level lexical relations which were used for shallow parsing and sentiment analysis such as Mourad and Darwish (2013) and El-Fishawy et al. (2014). The properties listed above of the microblogging domain make POS tagging on Twitter very different from their counterparts in more formal texts. It is an open question how well the features and techniques of NLP used on more well-formed data will transfer to Twitter in order to understand and exploit tweets. Our contributions in this paper are as follows: 1) Evaluating how robust state-of-the-art POS taggers for MSA are on Arabic tweets, 2) Identifying problem areas in tagging Arabic tweets and what caused the majority of errors and 3) Boosting the taggers' performance on Arabic tweets by making improvements based on observed errors.

## 2 Related Work

POS tagging is a well-studied problem in computational linguistics and NLP over the past decades. This can be inferred from high accuracy of state-of-the-art POS tagging not only for English, but also most other languages such as Arabic, which reaches 97% for Arabic and English being at 97.32% (Gadde et al., 2011). However, the performance of standard POS taggers for English is severely degraded on Tweets due to their noisiness and sparseness (Ritter et al., 2011). Therefore, POS taggers for English tweets have been developed such as ARK, T-Pos and GATE TwitIE which reach 92.8%, 88.4% and 89.37% accuracy respectively (Derczynski et al., 2013).

There has been relatively little work on building POS tools for Arabic tweets or similar text styles.

[1] Arabic was the fastest growing language on Twitter in 2011 (source:Semiocast)



*Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, pages 167–171,
Beijing, China, July 31, 2015. 

Al-Sabbagh and Girju (2012) and Abdul-Mageed et al. (2012) are strictly supervised approaches for tagging Arabic social media and they have assumed labelled training data. Their weakness is that they need a high quantity and quality of training data and this labelled data quickly becomes unrepresentative of what people post on Twitter. They also have been built specifically for dialectal Arabic and subjectivity and sentiment analysis.

Our work is, to best of our knowledge, the first step towards developing a POS tagger for Arabic tweets which can benefit a wide range of downstream NLP applications. We utilise the existing standard POS taggers for MSA instead of building a separate tagger. We use pre- and post-processing modules to improve their accuracy. Then, we use agreement-based bootstrapping on unlabelled data to create a sufficient amount of labelled training tweets that we can train our proposed tagger on it.

## 3 Data Collection

There is a growing interest within the NLP community to build Arabic social media corpora by harvesting the web such as Refaee and Rieser (2014) and Abdul-Mageed et al. (2012). However, none of these resources are publicly available yet. Hence, we built our own corpus which preserves all phenomena of Arabic tweets. We used Twitter Stream API to crawl Twitter by setting a query to retrieve tweets from the Arabian Peninsula and Egypt by using latitude and longitude coordinates of these regions since Arabic dialects in these regions share similar characteristics and they are the closest Arabic dialects to MSA. We did not restrict tweets language to "Arabic" in the query since users may use other character sets such as English to write their Arabic tweets (Romanisation) or they may mix Arabic script with another language in the same tweets. Next, we excluded all tweets which were written completely in English. Then, we sampled 390 tweets (5454 words) from the collected set to be used in our experiments (similar studies for English tweets also use a few hundred of tweets e.g. (Gimpel et al., 2011)).

## 4 Evaluating Existing POS Taggers

We evaluate three state-of-the-art publicly available POS taggers for Arabic, namely AMIRA (Diab, 2009), MADA (Habash et al., 2009) and Stanford Log-linear (Toutanova et al., 2003).

### 4.1 Gold Standard

A set of correctly annotated tweets (gold standard) is required in order to compare the outputs of the POS taggers with it. Since there is no publicly available annotated corpus for Arabic tweets, we have created POS tags for Twitter phenomena (i.e. REP, MEN, HASH, LINK, USERN, RET, EMOT and EMOJ for replies, mentions, hashtags, links, usernames, retweets, emoticons and emoji respectively). To speed up manual annotation, we tagged tweets by using the taggers, and then we corrected the output of the taggers to construct a gold standard.

### 4.2 POS Tagging Performance Comparison

We compare three taggers on 390 tweets (5454 words) from our corpus. The performance of these taggers are computed by comparing the output of each tagger against the manually corrected gold standard. The results for the AMIRA, MADA and Stanford which were trained on newswire text present poor success rates (see Table 1). This huge drop in the accuracy of these taggers when applied to Arabic tweets warrants some analysis of the problem and of mistagged cases.

| Tagger | Newswire | Arabic Tweets |
|---|---|---|
| AMIRA | 96.0% | **60.2%** |
| MADA | 97.0% | **65.8%** |
| Stanford | 96.5% | **49.0%** |

Table 1: POS tagging performance comparison

### 4.3 Error Analysis

We noticed that most of the mistagged tokens are unknown words. In this case, the taggers rely on contextual clues such as the word's morphology and its sentential context to assign them the most appropriate POS tags. We identified the unknown words that were mistagged and classified them into three groups: Arabic words, non-Arabic tokens and Twitter-specific (see Table 2).

**Arabic words** These are words which are written in Arabic, but which were assigned incorrect POS tags by the taggers. This category represents 73.5%, 68.1% and 79.2% of the total of mistagged items by AMIRA, MADA and Stanford respectively. We observed that words in this category have different characteristics and most of them are twitter phenomena. So, we classify them into subcategories as follows:

| Tagger | Types of mistagged items | Arabic Words | | | | | | | | Non-Arabic Tokens | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MSA words | Concatenation | Repeated letters | Named Entities | Spelling mistakes | Slang | Characters deletion | Transliteration | Romanisation | Emoticons | Emoji | Foreign words | Twitter specific |
| AMIRA | % of Errors | 53.3% | 1.8% | 0.8% | 8.7% | 0.6% | 6.2% | 0.9% | 1.2% | 1.0% | 0.5% | 2.8% | 2.6% | 19.6% |
| | Accuracy | 71.8% | 0.0% | 40.0% | 49.2% | 35.0% | 30.4% | 16.7% | 61.8% | 21.4% | 0.0% | 0.0% | 35.6% | 0.0% |
| MADA | % of Errors | 45.5% | 2.1% | 0.8% | 8.5% | 0.6% | 7.1% | 1.0% | 2.4% | 1.4% | 0.5% | 3.3% | 3.9% | 22.8% |
| | Accuracy | 79.3% | 0.0% | 50.0% | 57.0% | 40.0% | 32.0% | 20.8% | 35.3% | 7.1% | 0.0% | 0.0% | 17.2% | 0.0% |
| Stanford | % of Errors | 65.5% | 1.4% | 0.9% | 3.2% | 0.6% | 6.4% | 0.5% | 0.8% | 0.7% | 0.4% | 2.2% | 2.4% | 15.1% |
| | Accuracy | 55.0% | 0.0% | 20.0% | 75.7% | 20.0% | 7.2% | 45.8% | 67.6% | 25.0% | 0.0% | 0.0% | 21.8% | 0.0% |

Table 2: Errors percentage of each mistagged class and its accuracy

**MSA words** These are proper words which are used in well-formed text and part of MSA vocabulary, but which were assigned incorrect POS tags by the taggers. We observed that the accuracy of MSA words which are not noisy dropped from 96% for AMIRA, 97% for MADA and 96.5% for Stanford on newswire domain to 71.8%, 79.3% and 55% respectively on Arabic tweets.

**Concatenation** In this classification, two or more words were connected to each other to form one token. So, the taggers struggled to label them. Users may connect words deliberately to overcome tweets restricted length or accidentally. In this experiment, the taggers mistagged all connected words in the subset.

**Repeated letters** Words in this classification have one or more letters repeated. Users repeat letters deliberately to express subjectivity and sentiment.

**Named entities** All of these words should be labelled proper noun by the taggers because they refer to person, place or organization, but they mistagged them since these words were not part of their training data.

**Spelling mistakes** It is not easy to know the intent of the user, but some words seem likely to have been accidentally misspelled. Most words belonging to this category were mistagged by the taggers.

**Slang** The words in this category are regarded as informal and are typically restricted to a particular context or group of people. They are often mistagged by the taggers.

**Characters deletion** Arabic users delete letters from words deliberately to overcome tweets restricted length or because they do not have enough time to write complete words.

**Transliteration** Arabic users borrow some words and multiwords abbreviations from English. They use their Arabic transliteration in Arabic tweets. For example, LOL in English (Laugh Out Loud) is written in Arabic as "لول".

**Twitter-specific** They are elements that are unique to Twitter such as reply, mention, retweet, hashtag and url. They represent 19.6%, 22.8% and 15.1% of the total of mistagged items by AMIRA, MADA and Stanford respectively. In fact, taggers mistagged all Twitter-specific elements in the experiment and they tokenised them in different ways (see Table 3).

| | **AMIRA** | | **MADA/Stanford** | |
|---|---|---|---|---|
| Twitter element | Token | Tag | Token | Tag |
| @Moh_Ali | @ | PUNC | @Moh_Ali | noun |
| | Moh | NN | | |
| | _ | PUNC | | |
| | Ali | NN | | |

Table 3: Twitter element tokenised and tagged by taggers

**Non-Arabic tokens** This group contains the remaining twitter phenomena which are appear in Arabic tweets, but which are not written by using the Arabic alphabet. They represent 6.9%, 9.1% and 5.7% of the total of mistagged items by AMIRA, MADA and Stanford respectively.We classify them into subcategories based on their shared characteristics as follows:

**Romanisation** Arabic users sometimes use Latin letters and Arabic numerals to write Arabic tweets because the actual Arabic alphabet is unavailable for technical reasons, difficult to use or they speak Arabic but they cannot write Arabic script. For example, the word 3ala which is the Romanised form of the Arabic word "على".

**Emoticons** They are constructed by using traditional alphabetics or punctuation, usually a face expression. They are used by users to express their feelings or emotions in tweets. AMIRA and

MADA break emoticons into parts during tokenisation processes and they deal with each part as punctuation so all emoticons lost their meaning.

**Untagged emoji** Emoji means symbols provided in software as small pictures in line with the text which are used by users to express their feelings or emotions in tweets. AMIRA and MADA omitted these symbols in the tokenisation stage and they did not tag them.

**Foreign words** Some Arabic tweets contain foreign words especially from English. These words may refer to events, locations, English hashtags or retweet of English tweets with comments written in Arabic.

## 5 Improving POS Tagging Performance

Our experiments show that the taggers present poor success rates since they were trained on newswire text and designed to deal with MSA text. They fail to deal with Twitter phenomena. As a result, their outcomes are not useful to be used in linguistics downstream processing applications such as information extraction and machine translation in microblogging domain. Therefore, there is a need for a POS tagger which should take into consideration the characteristics of Arabic tweets and yield acceptable results.

Our goal is not to build a new POS tagger for Arabic tweets. The goal is to make existing POS taggers for MSA robust towards noise. There are two ways to do so, one is to retrain POS taggers on Arabic tweets and alter their implementation if needed, the other is to overcome noise through pre- and post-processing to the tagging. Our approach is based on both approaches. We will combine normalisation and external knowledge to boost the taggers' performance. Then, we will retrain our augmented version of Stanford tagger on Arabic tweets since its speed is ideal for tweets domain and it is only the retrainable tagger. However, we do not have suitable labelled training data to do so. Therefore, we will use bootstrapping on unlabelled data to create a sufficient amount of labelled training tweets.

### 5.1 Pre- and Post-processing

As seen in error analysis, unknown words (out-of-vocabulary tokens or OOV) represent a large proportion of mistagged tokens. We argue that normalisation will reduce this proportion which will improve the performance of the proposed tagger. Normalisation is the process of providing in-vocabulary (IV) versions of OOV words (Han and Baldwin, 2011). We will create a mapping from OOV tokens to their IV equivalents by using suitable dictionaries and the original token is replaced with its equivalent IV token. External sources of knowledge such as regular expression rules, gazetteer lists and an output of English tagger will also be used. The combination of normalisation and external knowledge will be applied to text as pre- and post-processing steps.

### 5.2 Agreement-based Bootstrapping

Bootstrapping is used to create a labelled training data from large amounts of unlabelled data (Cucerzan and Yarowsky, 2002; Zavrel and Daelemans, 2000). There are different ways to select the labelled data from the taggers' outputs. We will follow (Clark et al., 2003) in using agreement-based training method. We will use the augmented versions of AMIRA, MADA and Stanford taggers to tag a large amount of Arabic tweets and add the tokens which they agreed upon to the pool of training data. The taggers use different tagsets. Therefore, we will map these tagsets to a unified tagset consisting of main POS tags. Finally, we will retrain our augmented version of Stanford tagger on the selected labelled data.

## 6 Tagging Twitter-specific Items

We took the first step towards improving the accuracy of MSA taggers on Arabic tweets by tagging Twitter-specific elements. In these experiments, we used regular expression rules to detect and tag Twitter-specific elements such as mentions, hashtags, urls and etc. by doing some pre-processing and then tagging and finally doing post-processing. Due to the space limit, we present our effort to tag hashtags and all the remaining Twitter elements are tagged in similar way. First, we detected hashtags by using regular expression rules. Then, we removed the hashtag signs and underscores from raw tweets. Next, we tagged them by using AMIRA, MADA and Satnford. Finally, we inserted hashtag signs in their original place in tweets to indicate the beginning and the end of hashtags content. In fact, the taggers not just mistagged Twitter elements, but they also mistagged some MSA words in the same tweets because the text being noisy and the taggers rely on contextual clues.

## 7 Results

By using the above approach, we are not just able to tag Twitter elements correctly but we also make the context less noisy so the taggers are more likely to tag MSA word correctly. This approach boosts AMIRA, MADA and Stanford performance to 68.5%, 74.7% and 62.2% respectively as shown in Table 4.

| Tokens | AMIRA | MADA | Stanford |
|---|---|---|---|
| MSA words | 72.5% | 80.7% | 62.1% |
| Twitter-specific | 100% | 100% | 100% |
| **Overall** | **68.5%** | **74.7%** | **62.2%** |

Table 4: Taggers performance after tagging Twitter-specifics

## 8 Conclusion and Future Work

We have examined the consequences of applying MSA-trained POS tagging to Arabic tweets. Encouragingly, some comparatively simple pre- and post-processing steps go some of the way towards improving the taggers' accuracy over the MSA baseline. However, much work remains to be done to reach acceptable results. So, our next steps are to implement all the proposed steps in our approach to improve taggers' performance. Then, we will use bootstrapping and taggers agreement on unlabelled data to create a sufficient amount of labelled training tweets to retrain our augmented version of Stanford on it.

## Acknowledgments

The authors would like to thank the anonymous reviewers for their encouraging feedback and insights. Fahad would also like to thank King Saud University for their financial support. Allan Ramsay's contribution to this work was partially supported by Qatar National Research Foundation (grant NPRP-7-1334-6 -039).

## References

Muhammad Abdul-Mageed, Sandra Kübler, and Mona Diab. 2012. SAMAR: A system for subjectivity and sentiment analysis of arabic social media. In *Proceedings of WASSA*.

Rania Al-Sabbagh and Roxana Girju. 2012. A supervised POS tagger for written Arabic social networking corpora. In *Proceedings of KONVENS*.

Stephen Clark, James R. Curran, and Miles Osborne. 2003. Bootstrapping pos taggers using unlabelled data. In *Proceedings of NAACL*. ACL.

Silviu Cucerzan and David Yarowsky. 2002. Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings of NLL*. ACL.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of RANLP*.

Mona Diab. 2009. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools*.

Nawal El-Fishawy, Alaa Hamouda, Gamal M. Attiya, and Mohammed Atef. 2014. Arabic summarization in twitter social network. *Ain Shams Engineering Journal*.

Phani Gadde, L. V. Subramaniam, and Tanveer A. Faruquie. 2011. Adapting a WSJ trained part-of-speech tagger to noisy text: Preliminary results. In *Proceedings of MOCR*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of ACL: HLT*.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. Mada+ tokan: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of ACL: HLT*.

Ahmed Mourad and Kareem Darwish. 2013. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In *Proceedings of WASSA*. ACL.

Eshrag Refaee and Verena Rieser. 2014. An Arabic Twitter corpus for subjectivity and sentiment analysis. In *Proceedings of LREC*.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of EMNLP*.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL*.

Jakub Zavrel and Walter Daelemans. 2000. Bootstrapping a tagged corpus through combination of existing heterogeneous taggers. In *Proceedings of LREC*.