

# Fast and Robust POS tagger for Arabic Tweets Using Agreement-based Bootstrapping

Fahad Albogamy, Allan Ramsay

School of Computer Science  
The University of Manchester  
Manchester, M13 9PL, UK

{fahad.albogamy, allan.ramsay}@manchester.ac.uk

## Abstract

Part-of-Speech (POS) tagging is a key step in many NLP algorithms. However, tweets are difficult to POS tag because they 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. On the basis of this analysis, we combine normalisation and external knowledge to handle the domain noisiness and exploit bootstrapping to construct extra training data in order to improve POS tagging for Arabic tweets. Our results show significant improvements over the performance of a number of well-known taggers for Arabic.

**Keywords:** Arabic tweets, POS tagging, Bootstrapping

## 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. A study prepared and published by SemioCast in 2012 has revealed that Arabic was the fastest growing language on Twitter in 2011. People post about their lives, share opinions on a variety of topics and discuss current issues. There are millions of tweets daily, yielding a corpus which is noisy and informal, but which is sometimes informative. As a result, Twitter has become one of the most important social information mutual platforms. The nature of the text content of microblogs differs from traditional blogs. In Twitter, for example, a tweet is short and contains a maximum of 140 characters. Tweets also are not always written maintaining formal grammar and proper spelling. Slang and abbreviations are often used to overcome their restricted lengths (Java et al., 2007).

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 (Barbosa and Feng, 2010). 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; El-Fishawy et al., 2014). They do not use the standard linguistic pipeline tools such as POS tagging which might enable a richer linguistic analysis (Gimpel et al., 2011). The properties listed above of the microblogging domain make POS tagging on Twitter very different from its counterpart in more formal texts. It is an open question how well the features and techniques of NLP used on more well-formed data (e.g. in the newswire domain) will transfer to Twitter in order to understand and exploit tweets. Therefore, we experimentally evaluate the performance of state-of-the-art POS taggers for Modern Standard Arabic (MSA) on Arabic tweets. POS tagging accuracy drops from about 97% on MSA to 49-65% on Arabic tweets. We also analyse their

limitations and errors they made. Finally, we propose an approach to boost their performance.

Our contributions in this paper are as follows:

1. Evaluating how robust state-of-the-art POS taggers for MSA are on Arabic tweets (Section 4.2.) and identifying problem areas in tagging Arabic tweets and what caused the majority of errors (Section 4.3.).
2. Boosting the taggers' performance on Arabic tweets by using pre- and post-processing techniques to address Arabic tweets' noisiness (Section 5.1.).
3. Investigating 'agreement-based bootstrapping' on unlabelled Arabic tweets, to create a sufficient amount of tweets training data (Section 5.2.).

## 2. Related Work

POS tagging is a well-studied problem in computational linguistics and NLP over the past decades. This can be inferred from the 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 reaches 92.8%, 88.4% and 89.37% accuracy respectively (Derczynski et al., 2013).

People working on Arabic tweets have tended to concentrate on lexical relations because a tagger that can actually work on this domain with an acceptance degree of accuracy is yet to be developed (Elsahar and El-Beltagy, 2014). There has been relatively little work on building POS tools for Arabic tweets or similar text styles. (Al-Sabbagh and Girju, 2012; 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 such as information extraction and machine translation. We evaluate the existing state-of-the-art POS tagging tools on Arabic tweets, with an intention of developing a POS tagger for Arabic tweets by utilising 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; Abdul-Mageed et al., 2012). However, none of these resources are publicly available yet. They also do not contain all phenomena of tweets as they appear in their original forms in Twitter and they have been built to be used mainly in sentiment analysis. 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 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 tweets e.g. (Gimpel et al., 2011)).

### 4. Evaluating Existing POS Taggers

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

#### 4.1. Gold Standard

A set of correctly annotated tweets (gold standard) is required in order to be able to appraise the outputs of POS taggers. Once we have this, we can compare the outputs of the POS taggers with this gold standard. 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 and RET for replies, mentions, hashtags, links, usernames and retweets respectively) and we manually annotated our dataset. 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 is computed by comparing the output of each tagger against the manually corrected gold standard. We use standard precision, recall and F-score as evaluation measures. The results for AMIRA, MADA and Stanford, which were trained on newswire text, present poor success rates, for example, the precision for AMIRA, MADA and Stanford on Arabic tweets are 60.2%, 65.8% and 49.0% respectively (see Table 1). These figures are far below the performance of the same taggers on newswire genres, where accuracy is around 96% for AMIRA and Stanford whereas MADA achieves over 97% accuracy<sup>1</sup>. This huge drop in the accuracy of these taggers on Arabic tweets warrants some analysis of the problem and of mistagged cases.

Tagger	Newswire	Arabic Tweets
AMIRA	96.0%	<b>60.2%</b>
MADA	97.0%	<b>65.8%</b>
Stanford	96.5%	<b>49.0%</b>

Table 1: POS tagging accuracy for three off-the-shelf taggers

#### 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 (Foster et al., 2011). We identified the unknown words that were mistagged and classified them into three groups: Arabic words, Twitter-specific and non-Arabic tokens. Table 2 shows baseline taggers' performance on each category.

**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 73.2% of the total of mistagged items by AMIRA, MADA and Stanford respectively. We observed that words in

Tokens	AMIRA	MADA	Stanford
Arabic words	67.1%	73.8%	54.0%
Twitter-specific	0.0%	0.0%	0.0%
Non-Arabic	19.9%	9.1%	14.0%
<b>Overall</b>	<b>60.2%</b>	<b>65.8%</b>	<b>49.0%</b>

Table 2: POS tagging accuracy on Arabic Tweets for baseline taggers categorised by tokens

this category have different characteristics and most of them are twitter phenomena. So, we classify them into subcategories as follows:

**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%

<sup>1</sup>These taggers' performance on newswire are provided in the literature





difficult to detect all slangs in tweets domain. Therefore, we select the most frequent twenty slang words from 17k types in our corpus (10 million tokens) and map them to their IV equivalents as shown in Table 5.

**Handling Twitter-specific Items** We use 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 the way we deal with hashtags: all the remaining Twitter elements are tagged in similar ways. 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 Stanford. Finally, we inserted hashtag signs in their original place in tweets to indicate the beginning and the end of hashtags content as shown in Table 6.

Raw Tweet	حياتي فالبيت بقضيا جنب الشاحن !! بي جالاكسي لا تكلفني
MADA	... !,punc !,punc #,punc jAlAksy,noun →, noun <b>IA,verb</b> →,noun tklmny,verb
Preprocessing	حياتي فالبيت بقضيا جنب الشاحن !! جالاكسي لا تكلفني
MADA	... punc !,punc jAlAksy,noun <b>IA,part_neg</b> tklmny,verb
Postprocessing	... punc !,punc <hash> jAlAksy,noun IA,part_neg tklmny,verb </hash>

Table 6: Pre- and post-processing (tag hashtag’s words)

In fact, the taggers not just mistagged Twitter elements, but they also mistagged some MSA words in the same tweets because the text is noisy and the taggers rely on contextual clues. 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 words correctly as "IA" word in Table 6.

**Handling Named Entities** These can be recognised by using gazetteer lists. We use ANERGazet<sup>3</sup> which a collection of three Gazetteers, (i) Locations: it contains names of continents, countries, cities, etc.; (ii) People: it has names of people recollected manually from different Arabic websites; and finally (iii) Organizations: it contains names of organizations like companies, football teams, etc..

**Handling English Words** Our focus is on Arabic tweets, but some of them contain English words. These words may refer to events, locations, English hashtags or retweet of English tweets with comments written in Arabic and they are part of the syntactic structure of Arabic tweets. So, they need to be tagged correctly. In this case, we use Stanford for English (Toutanova et al., 2003) to tag English words as a post-processing step.

### Results for Pre- and Post-processing

In our experiments, the taggers were adapted to handle Twitter phenomena. The experiments were run using three off-the-shelf taggers trained on PATB and our augmented approach to address Arabic tweets noisiness as described in Section 5.. Table 7 shows the performance difference

on each of three categories of mistagged tokens and the overall performance compared with their baseline performance in Table 2. By combining normalisation and external knowledge, we are able to reduce unknown tokens in each category which boosts the taggers’ performance. The overall performance of the three taggers increases by absolute twelve percent accuracy for AMIRA and by absolute thirteen percent for MADA and Stanford. This improvement in accuracy will reduce the propagation of POS tagging errors to downstream applications on Arabic tweets such as information extraction.

Tokens	AMIRA	MADA	Stanford
Arabic words	70.4%	77.9%	62.1%
Twitter-specific	100%	100%	100%
Non-Arabic	68.3%	66.1%	66.1%
<b>Overall</b>	<b>72.6%</b>	<b>79.0%</b>	<b>65.2%</b>

Table 7: Impact of applying pre- and post-processing on POS tagging accuracy

We can infer from the above results that the augmented version of MADA is the most appropriate tagger for tagging Arabic tweets since it outperforms its counterparts of AMIRA and Stanford in accuracy. However, the accuracy is not only the factor that should be considered when tagging Arabic tweets. The tagging speed is another crucial factor to take into account since there are millions of tweets that need to be tagged. So, the most suitable tagger may be the one that performs well in terms of both speed and accuracy. So, we measure the taggers’ speed below.

**Tagging speeds** We measure speed performance of AMIRA v2.1, MADA v3.2 and Stanford v3.5.1 POS taggers on 250k tweet tokens. It is measured in words processed per second. The speed evaluation is conducted for Stand-alone (raw input) modes on a Dell XPS laptop computer with Intel(R) Core(TM)2 Duo CPU T8100 @2.10GHz and 4GB memory.

Tagger	Accuracy	Speed(words/sec.)
AMIRA	72.6%	289
MADA	<b>79.0%</b>	48
Stanford	65.2%	<b>8966</b>

Table 8: Accuracy and speed comparison for the taggers

As shown in Table 8, MADA tagger outperforms AMIRA and Stanford in tagging accuracy, but it is the slowest tagger whereas Stanford tagger is the fastest tagger (31-186x faster than the others), but it is the lowest tagger in tagging accuracy. So, neither one of them is ideal for tagging Arabic tweets at this stage and they need more improvements. Due to the fact that we have no access to their codes, the reasons for the variation in speed are not visible. Based on that, we decide to further improve the accuracy of Stanford tagger while preserving its speed by using Arabic tweets training data. However, there is no labelled training data available to do so. Therefore, we use bootstrapping on unlabelled data to create a sufficient amount of labelled training tweets.

<sup>3</sup><http://users.dsic.upv.es/grupos/nle/?file=kop4.php>

## 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 follow (Clark et al., 2003) in using agreement-based training method. We 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. Then, we retrain Stanford tagger on the selected labelled data.

Tag	Gloss
CC	coordinating conjunction
CD	cardinal number
DT	demonstrative pronoun
IN	subordinating conjunction or preposition
JJ	adjective
NN	common noun
NNP	proper noun
PRP	personal pronoun
PUNC	punctuation
RB	adverb
RP	particle
UH	interjection
VB	verb
WP	relative pronoun
WRB	wh-adverb

Table 9: Collapsed tagset

**Tagset Unification** The taggers use different tagsets. AMIRA uses RTS tagset which consists of 25 tags whereas Stanford uses Bies’s tagset which is similar to RTS tagset with extra tags to represent the determiner "AI". So, we reduce Stanford tagset to AMIRA tagset by omitting determiner tags. On the other hand, MADA has 34 different tags which make different types of distinctions with RTS\Bies tagsets. For example, RTS tagset uses one tag (RP) to cover a range of particles which are subdivided to nine subclasses by MADA (`part_det`, `part_focus`, `part_fut`, `part_interrog`, `part_neg`, `part_restrict`, `part_verb`, `part_voc`, `part`); and it uses several tags to distinguish between verbs tenses (VBD, VBG, VBN, VBP) where MADA just uses one tag (`verb`). Therefore, mapping between these tagsets is a prerequisite for using agreement-based method on the taggers’ outputs. We construct a unified tagset consisting of the main POS tags as shown in Table 9 to do the mapping.

**Agreement-based Bootstrapping Results** We used the augmented versions of AMIRA, MADA and Stanford taggers to tag 25K Arabic tweets which contain 203682 tokens. In these experiments, we focused on Arabic words so we omitted all non-Arabic tokens from the tagged data. This reduced the number of tokens used in agreement-based bootstrapping experiment to 166573 Arabic words. Then, we mapped the taggers’ outputs to our unified tagset. Next, we admitted all Arabic words that all taggers were agreed upon to the pool of training data. Taggers reached agreement on 60.4% of Arabic words which formed 100691 agreed bootstrapped tokens

as a training data. Finally, we trained Stanford tagger on the bootstrapped training data. Training on bootstrapped data improves the performance of Stanford from 58.5% to 66.5% on Arabic words and from 49.0% to 69.1% on overall tokens (Table 10). We tried adding 50K further Arabic words to the training pool by repeating the same experiment on different Arabic tweets, but they did not give a performance increase which suggests no potential benefit from more bootstrapping data.

We have noticed that 75% from our dataset in Section 4.2. is MSA words, so we argue that adding some amounts of labelled newswire tokens to the pool of bootstrapped data above and retrain Stanford on it will boost its accuracy. Table 10 shows the performance of Stanford improves from 66.5% to 72.1% on Arabic words and from 69.1% to 74% on overall tokens when PATB was added to the training pool. By using the above approaches, we are able to increase the accuracy of Stanford tagger while preserving its speed. It outperforms AMIRA in terms of speed and accuracy. It also outperforms MADA in terms of speed and it is not a long way behind MADA’s accuracy. So, it is the most suitable taggers for tagging Arabic tweets.

Training data	Overall	Arabic words	size
PATB	65.2%	62.1%	145K
Bootstrapped	69.1%	66.5%	100K
PATB+Bootstrapped	<b>74.0%</b>	<b>72.1%</b>	245K

Table 10: Stanford performance on Arabic Tweets using different training data

## 6. Conclusion

We have examined the consequences of applying MSA-trained POS tagging to Arabic tweets. Off-the-shelf taggers present poor success rates on Arabic tweets due to the domain noisiness. Furthermore, existing approaches suffer from insufficient labelled training data. We introduced approaches for avoiding the noisiness of domain by combining normalisation and external knowledge and for generating training data by applying agreement-based bootstrapping on heterogeneous taggers’ outputs. These combined to improve POS tagging for Arabic tweets. Our techniques yield a very fast and robust POS for Arabic tweets. By using a pool of bootstrapped data combined with PATB to train the augmented version of Stanford tagger, we are able to improve its accuracy from 49% to 74% on Arabic tweets while preserving its speed.

## 7. 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).

## 8. Bibliographical References

Abdul-Mageed, M., Kübler, S., and Diab, M. (2012). SAMAR: A system for subjectivity and sentiment analysis of Arabic social media. In *Proceedings of the 3rd*

- Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 19–28. Association for Computational Linguistics.
- Al-Sabbagh, R. and Girju, R. (2012). A supervised POS tagger for written Arabic social networking corpora. In *Proceedings of KONVENS*, pages 39–52.
- Albogamy, F. and Ramsay, A. (2015a). Towards POS tagging for Arabic tweets. In *Proceedings of ACL Workshop on Noisy User-generated Text*.
- Albogamy, F. and Ramsay, A. (2015b). POS tagging for Arabic tweets. In *Proceedings of Recent Advances in Natural Language Processing*, pages 1–8.
- Barbosa, L. and Feng, J. (2010). Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics.
- Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565.
- Clark, S., Curran, J. R., and Osborne, M. (2003). Bootstrapping POS taggers using unlabelled data. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4*, pages 49–55. Association for Computational Linguistics.
- Cucerzan, S. and Yarowsky, D. (2002). Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings of the 6th Conference on Natural Language Learning-Volume 20*, pages 1–7. Association for Computational Linguistics.
- Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*, pages 198–206.
- Diab, M. (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*.
- El-Fishawy, N., Hamouda, A., Attiya, G. M., and Atef, M. (2014). Arabic summarization in Twitter social network. *Ain Shams Engineering Journal*, 5(2):411–420.
- Elshahar, H. and El-Beltagy, S. R. (2014). A fully automated approach for Arabic slang lexicon extraction from microblogs. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8403, CICLing 2014*.
- Foster, J., Çetinoglu, Ö., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., and Van Genabith, J. (2011). #hardtoparse: POS tagging and parsing the twitterverse. In *AAAI 2011 Workshop on Analyzing Microtext*, pages 20–25.
- Gadde, P., Subramaniam, L., and Faruque, T. A. (2011). Adapting a WSJ trained part-of-speech tagger to noisy text: preliminary results. In *Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*, page 5. ACM.
- Gimpel, K., Schneider, N., O’Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.
- Habash, N., Rambow, O., and Roth, R. (2009). Mada+ token: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization.
- Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 368–378. Association for Computational Linguistics.
- Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we Twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM.
- Mourad, A. and Darwish, K. (2013). Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 55–64.
- Mustafa, S. H. (2012). Word stemming for Arabic information retrieval: The case for simple light stemming. *Abhath Al-Yarmouk: Science & Engineering Series*, 21(1):2012.
- Refaee, E. and Rieser, V. (2014). An Arabic Twitter corpus for subjectivity and sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC14), Reykjavik, Iceland, may. European Language Resources Association (ELRA)*.
- Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.
- Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.
- Zavrel, J. and Daelemans, W. (2000). Bootstrapping a tagged corpus through combination of existing heterogeneous taggers. In *Proceedings of the Second International Conference on Language Resources and evaluation LREC-2000*, pages 17–20.