Developing a Competitive HMM Arabic POS Tagger Using Small Training Corpora

Mohammed Albared, Nazlia Omar, and Mohd. Juzaiddin Ab Aziz

University Kebangsaan Malaysia, Faculty of Information Science and Technology, Department of Computer Science
mohammed_albared@yahoo.com, {no,din}@ftsm.ukm.my
http://www.ukm.my

Abstract. Part Of Speech (POS) tagging is the ability to computationally determine which POS of a word is activated by its use in a particular context. POS is one of the important processing steps for many natural language systems such as information extraction, question answering. This paper presents a study aiming to find out the appropriate strategy to develop a fast and accurate Arabic statistical POS tagger when only a limited amount of training material is available. This is an essential factor when dealing with languages like Arabic for which small annotated resources are scarce and not easily available. Different configurations of a HMM tagger are studied. Namely, bigram and trigram models are tested, as well as different smoothing techniques. In addition, new lexical model has been defined to handle unknown word POS guessing based on the linear interpolation of both word suffix probability and word prefix probability. Several experiments are carried out to determine the performance of the different configurations of HMM with two small training corpora. The first corpus includes about 29300 words from both Modern Standard Arabic and Classical Arabic. The second corpus is the Quranic Arabic Corpus which is consisting of 77,430 words of the Quranic Arabic.

Keywords: Arabic languages, Hidden Markov model, Unknown words.

1 Introduction

Part of speech disambiguation is the ability to computationally determine which POS of a word is activated by its use in a particular context. POS tagging is an important basis of many higher level NLP applications like speech processing and machine translation. Being one of the first processing steps in any such application, the performance of the POS tagger directly impacts the performance of any subsequent text processing steps. Language resources such as annotated corpora are fundamental for research and development in statistical computational linguistics and for the construction of NLP applications. The main challenge involved in constructing any Arabic NLP system for Arabic is amplified by the lack of these language resources [1]. In spite of recent progress, Arabic is still lacking such tools and annotated resources [2].

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 288–296, 2011.

The task of POS tagging is very difficult due to two main reasons. First, many words are ambiguous. For example, English word " past " can be an adjective (e.g. his past performance), an adverb (e.g. its ten past seven), or a noun (e.g. in the past). This ambiguity is exist in all languages. For example, Arabic word "علم" can be a noun (e.g. علم مفيد), or a verb (e.g. (علم الانسان). Moreover, words can be ambiguous to their grammatical properties. For example, Arabic word " كتب " can be a past verb (e.g.كتب الكتاب), or a passive verb (e.g.كتب الكتاب). Second, the existing of unknown words, words that appears in the test data and do not appear in the training data. The problem of unknown words is the main problem in POS tagging [3]. Actually, the size of this problem is proportional to many factors such as: size, genre and the quality of the training data. This Unknown words problem becomes more serious when the training data are small [4]. With small training data, it is very difficult to predict the distribution of the unknown words. The impact of this problem increases in languages which have huge vocabulary and rich morphological system like Arabic.

This work presents a study aiming to find out the appropriate way to develop a competitive Arabic statistical POS tagger when only a limited amount of training material is available. To do so, we compare different smoothing techniques and different order HMMs. In addition, we propose a new lexical model to better handling unknown word in Arabic POS tagging. This new lexical model is based on the linear interpolation of both word's suffix probability and word's prefix probability. Several experiments are conducted to determine the performance of the different configurations using two small training corpora.

The remainder of this paper is organized as follows: First, Section 2 reviews some related works. Then, the training corpora are described in Section 3. Section 4 presents a breif description of the investigated HMM models.We then describe our method for unknown word POS guessing in Sections 5. Section 6 shows the realized experiments and the obtained results. Finally, Section 7 states some conclusions and further work.

2 Related Work on POS Tagging

POS tagging has been largely studied. There are different approaches have been used for POS tagging. There are some machine-learning taggers [5,6,7,8] and some rule based taggers [9]. POS tagging is often considered to be a "solved task ", with published tagging accuracies around 97%. However, In the real-life scenario, Giesbrecht and Evert [10]showed that five state-of-the-art POS taggers are unsuitable for fully automatic processing. Among all these state-of-the-art taggers, HMM taggers are more robust and much faster than other advanced machine learning approaches in the real-life scenario. Previous work on POS tagging has utilized different kind of features to tackle unknown word POS tagging. These features are mainly based on word substring information, word context information and/or global information.

Padr et al. [11] and Ferrndez et al. [12] investigate different configurations of the HMM Spanish POS Tagger when minimum amount of training data is

available. In contrast with our work, they assume the none-existence of unknown words, which is the main problem when the training data is small [3,4]. However, they work with Spanish language which is very different from Arabic language.

Present day Arabic has two literary styles. One is called the Classic Arabic and the other is the Modern standard Arabic (MSA). Classical Arabic is the language of formal writing until nearly the first half of the 20th century and was also the spoken language before the medieval times [13]. MSA is the language of formal modern writing in all Arabic countries.

Recently, several works have been proposed to Arabic POS tagging such as [14,15,16,17,18,19].For more details about Arabic work and also about POS tagging techniques in general,see [20]. Almost all of these taggers are generally developed for Modern Standard Arabic (MSA) and few works are interested in Classical Arabic [21]. In contrast with other Arabic taggers, our POS tagger deals with both MSA and Classic Arabic together. Additionally up to our knowledge, this is the first work which aims to find the appropriate configuration of Arabic POS tagger when small amount of training data is available. Furthermore, our work has defined a new lexical model to handle unknown words in Arabic POS tagging. The proposed lexical model demonstrates to be effective and efficient in handling unknown word even with small training data. An unknown words POS tagging accuracy of 85.3% obtained using the introduced method is as of yet the highest reported in the Arabic POS tagging literature.

3 The Used Data

For our experiments, we used two different and small Arabic corpora: the FUS-HA corpus and The Quranic Arabic Corpus. The FUS-HA corpus is composed of two sub-corpora:

1. MSA corpus: The MSA corpus is composed of journalistic articles discussing general news topics. The news topics cover various subjects of politics, economics and culture. The corpus includes more than 12000 words forms. The This data are 2009-2010 newswire feeds collected from different online Arabic newspaper archives, such as Al-Jazeera and Alsharq Al-Awsat.
2. Classic Arabic corpus: The Classic Arabic corpus is composed of some texts extracted from ALJAHEZ's book entitled "Albayan-wa-tabyin" (255 Hijri). "Albayan-wa-tabyin" "The art of communication and demonstration" is one of the best and earliest Arabic books on Arabic literary theory and literary criticism. The book covers various subjects, such as rhetorical speeches, history and science. The Classic Arabic corpus includes more than 17000 word forms.

We annotated the FUS-HA corpus using two Arabic tag sets. The first one is the Arabic TreeBank tagset, which is consist of 23 tags, used by Diab et al [16]. The second one is quite similar to the first one. We only add some modifications to handle some linguistic limitation. We introduce a tag for the Broken Plural to distinguish between it and the singular noun. Broken Plurals

constitute 10% of any Arabic text and form 40% of the Arabic plurals [22]. The second modification, our tagset does not include NO_FUNC (no solution chosen) tag, which is used as a tag in the above mentioned Arabic TreeBank tagset. Finally, we distinguish between inflected and non inflected verbs. However, the modified tag set consists of 24 tags.

The Quranic Arabic Corpus[23] is an annotated linguistic resource which shows the Arabic grammar, syntax and morphology for each word in the Holy Quran, the religious book of Islam which is written in classical Quranic Arabic (c. 600 CE). The research project is organized at the University of Leeds, and is part of the Arabic language computing research group within the School of Computing. The Quranic Arabic Corpus is consisting of 77,430 words of Quranic Arabic.

4 Hidden Markov Models

In HMM, the POS problem can be defined as the finding the best tag sequence t^n given the word sequence w^n. The label sequence t^n generated by the model is the one which has highest probability among all the possible label sequences for the input word sequence. This is can be formally expressed as:

$$t_1^n = \arg\max_{t_1^n} \prod_{i=1}^{n} p\left(t_i | t_{i-1} \ldots . t_1\right) . p\left(w_i | t_i \ldots . t_1\right) \ . \tag{1}$$

The two models, the state transition probabilities and the emission probabilities, parameters are estimated from annotated corpus by Maximum Likelihood Estimation (MLE), which is derived from the relative frequencies. We will use Hidden Markov Models POS taggers of order two and three. Having computed the state transition probabilities and the emission probabilities and assigning all possible tag sequences to all words in a sentence, now we need an algorithm that can search the tagging sequences and find the most likely sequence of tags. For this purpose we use the Viterbi algorithm [24] which compute the maximized tag sequence with the best score. However, MLE is a bad estimator for statistical inference especially, in NLP application, because data tends to be sparse. In this work, two smoothing methods are used. With the Bigram version, we use the Modified Kneser Ney smoothing technique [25]. In the Trigram version, we use the linear interpolation of unigram, bigram and trigram maximum likelihood estimates [6] in order to estimate the trigram transition probability.

5 Unknown Words Handling

In POS tagging, we frequently encounter words that do not appear in training data. Such words are called unknown words or out-of-vocabulary (OOV) words. The existence of unknown words is the main problem for POS taggers, since the

statistical information of these words are unavailable. It is a non-negligible problem especially where only a limited amount of training material is available. Unknown words are usually handled by an exceptional processing. Accuracy of POS tagging for unknown words is usually much lower than that for known words.

In order to handle the POS tagging for unknown words in Arabic POS tagging, we have defined a new lexical model based on the linear interpolation of both word suffix probability and word prefix probability. It combines together both word suffix information and word prefix information. The main linguistic motivation behind combining affixes information is that in Arabic word sometimes an affix requires or forbids the existence of another affix [13]. Prefix and suffix are the first n and m letters of the word, and are not necessarily morphologically meaningful. In this model, the lexical probabilities are estimated as follows:

1. Given an unknown word w, the lexical probabilities $P(\text{suffix(w)}|t)$ are estimated using the suffix tries as in the following equation:

$$P\left(t|c_{n-i+1},...,c_n\right) = \frac{P(t,c_{n-i+1},...,c_n) + \theta P(t,c_{n-i+2},...,c_n)}{1+\theta} . \quad (2)$$

$$\theta = \frac{1}{S-1}\sum_{J=1}^{S}\left(P\left(t_j\right)-\overline{P}\right)^2, \overline{P} = \frac{1}{s}\sum_{J=0}^{S}P\left(t_j\right)$$

   where $c_{n-i+1},...,c_n$ represent the last n characters of the word, $S$is the number of tags in the tagsets.
2. Then, the lexical probabilities $P(\text{prefix(w)}|t)$ are estimated using the prefix tries as in Equation 2. But, we reverse the letters in the words before adding them to the new word trie in order to find the prefix probability. Here, the probability distribution for a unknown word prefix is generated from all words in the training set that have the same prefix up to some predefined maximum length.
3. Finally, we use the linear interpolation of both the lexical probabilities obtained from both word's suffix and word's prefix to calculate the lexical probability of the word w as in the following equation:

$$P\left(w|t\right) = \lambda P\left(suffix(w)|t\right) + (1-\lambda)P(prefix(w)|t) \quad (3)$$

   where λ is an interpolation factor. In addition, the experiments also utilize the following features: the presence of non-alphabetic characters and the existence of foreign characters.

6 Experiments and Results

The main objective of this work is to study the behavior of different configurations for a HMM POS tagger, in order to determine the appropriate way to develop competitive Arabic tagger. To do so, we have carried out several experiments when small amount of training data is available. The data used for

the empirical evaluation come from the above described corpora. The FUS-HA corpus is divided into 78% for training and 22 % for testing. The percentage of unknown words in this test set is 10.7% . Whereas the Quranic Arabic Corpus is divided into 90.1 % for training and 9.9 % for testing. The percentage of unknown words in the Quranic Arabic Corpus test set is 14.9%. This decision was taken because the test data has to be guaranteed as unseen during training [26]and also it is necessary for comparisons to be consistent with previous evaluation works[27].

Results obtained for each HMM tagger configuration are summarized in Table 1 . Results are given both for the FUS-HA corpus and the Quranic Arabic Corpus. We define the tagging accuracy as the ratio of the correctly tagged words to the total number of words. As we can see in Table1, the experiments show that the best results on all test sets were achieved by the trigram HMM with the linear interpolation of unigram, bigram and trigram smoothing technique.

Table 1. Obtained results for all HMM POS tagger configurations using both corpora

λ	Bigram HMM						Trigram HMM					
	FUSHA 23 tags		FUSHA 24 tags		Quranic		FUSHA 23 tags		FUSHA 24 tags		Quranic	
	Unknown	Overall	Unknown	Overall	Unknown	Overall	Unknown	Overall	Unknown	Overall	Unknown	Overall
0	66.2	94.4	61.5	93.6	72.6	93	64.4	94.6	49.7	92.9	72.2	92.9
0.1	68.6	94.7	62.4	93.7	75	93.5	67.1	94.9	53.3	93.3	76.1	93.5
0.2	68.9	94.7	62.3	93.8	76.7	93.7	69.6	95.2	57	93.7	77.2	93.7
0.3	69.6	94.8	64.8	94	79	94.1	71.4	95.4	60.5	94	80.8	94.3
0.4	71.3	94.9	62.8	93.8	80.7	94.3	70.2	95.2	61.3	94.1	81.9	94.4
0.5	71	94.9	63.1	93.8	80.8	94.3	70.2	95.2	62.6	94.3	84	94.8
0.6	68.9	94.7	61.6	93.6	82.6	94.6	69.3	95.1	62.7	94.3	**85.3**	95
0.7	66.3	94.4	59.9	93.4	82.3	94.6	66.9	94.8	63	94.3	84.6	94.9
0.8	64.4	94.1	58.1	93.3	81.6	94.5	62.8	94.4	60.9	94.1	82.9	94.6
0.9	63.3	94	56.1	93	80.8	94.4	60.8	94.2	59.3	93.9	82.2	94.5
1	61.7	93.8	55.2	92.9	79.3	94.2	56.4	93.7	58.1	93.8	79.4	94.1

Comparing the results for the different order models, we can draw the following conclusions:

- In general, taggers trained the Quranic Arabic Corpus using have higher precision than those taggers trained using the FUS-HA corpus. This demonstrates that increasing the size of the training set has a positive impact on the accuracy rate.
- It is clearly that working with a trigram HMM gives higher precision than working with a bigram one, for both training corpora.

– The most important conclusion is that the proposed lexical model, the linear interpolation of both the suffix probability and the prefix probability, improves the tagging of Arabic text. The linear interpolation model($0 \prec \lambda \prec 1$)improves the POS tagging of unknown words dramatically compared to the suffix model ($\lambda = 1$), which has been used in several previous studies and proofed to be effective for other languages [6],[28], and the prefix model ($\lambda = 0$).

Nevertheless, some important observations can be extracted from these results:

– A competitive HMM taggers may be built using relatively small train sets, which is interesting, especially, with the lack of language resources.
– The linear interpolation, which combine information from both word suffix and word prefix togather, is a good indicator of Arabic unknown word POS.

7 Conclusion

Part of speech tagging is an important tool in many NLP applications. In addition, it accelerates the development of large manually annotated corpora. In this paper, we have studied how competitive Arabic HMM-based POS taggers can be developed using relatively small training corpus. Different configurations of a HMM tagger are investigated. We conducted a series of experiments using two small Arabic training corpora. Results indicate that accurate Arabic taggers can be build provided appropriate lexical model to handle unknown words. Between all configurations studied here, in general the one that gives a higher precision is trigram HMM with the linear interpolation of unigram, bigram and trigram smoothing technique and the linear interpolation of both word's suffix probability and word's prefix probability unknown word guessing algorithm. In the future work, it might also be worthwhile to improve the tagging accuracy of unknown words. This improvement can be done through the integration of highly coverage Arabic morphological analyzer, increasing the size of our training corpus as well as by using specific features of Arabic words, which can be better predictor for Arabic words POS than words suffixes and prefixes. Also, we plan to study their influence on the taggers developed on small corpora.

References

1. Farghaly, A., Shaalan, K.: Arabic Natural Language Processing: Challenges and Solutions. ACM Transactions on Asian Language Information Processing (TALIP), 1–22 (2009), doi:http://doi.acm.org/10.1145/1644879.1644881
2. Maamouri, M., Bies, A., Kulick, S.: Enhanced Annotation and Parsing of the Arabic Treebank. In: INFOS (2008)
3. Fischl, W.: Part of Speech Tagging - A solved problem? Center for Integrative Bioinformatics Vienna, CIBIV (2009) (Unpublished report)
4. Nakagawa, T.: Multilingual word segmentation and part-of-speech tagging: a machine learning approach incorporating diverse features. PhD Thesis, Nara Institute of Science and Technology, Japan (2006)

5. Ratnaparkhi, A.: A maximum entropy part of speech tagger. In: Brill, E., Church, K. (eds.) Conference on Empirical Methods in Natural Language Processing. University of Pennsylvania, Philadelphia (1996)
6. Brants, T.: TnT: A statistical part-of-speech tagger. In: Proceedings of the 6th Conference on applied Natural Language Processing, Seattle, WA, USA (2000)
7. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning, MA, USA (2001)
8. Goldwater, S., Griffiths, T.: A fully Bayesian approach to unsupervised part-of-speech tagging. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (2007)
9. Brill, E.: A Corpus-based Approach to Language Learning. PhD thesis, Department of Computer and Information Science. University of Pennsylvania, Philadelphia (1993)
10. Giesbrecht, E., Stefan, E.: Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In: Proceedings of the 5th Web as Corpus Workshop (WAC5), Donostia (2009)
11. Padró, M., Padró, L.: Developing Competitive HMM PoS Taggers Using Small Training Corpora. In: Vicedo, J.L., Martínez-Barco, P., Muńoz, R., Saiz Noeda, M. (eds.) EsTAL 2004. LNCS (LNAI), vol. 3230, pp. 127–136. Springer, Heidelberg (2004)
12. Ferrández, S., Peral, J.: Investigating the Best Configuration of HMM Spanish PoS Tagger when Minimum Amount of Training Data Is Available. In: Montoyo, A., Muńoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 341–344. Springer, Heidelberg (2005)
13. Attia, M.: Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation. PhD thesis, School of Languages, Linguistics and Cultures, Univ. of Manchester, UK (2008)
14. AlGahtani, S., Black, W., McNaught, J.: Arabic Part-Of-Speech Tagging using Transformation-Based Learning. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt (2009)
15. Kulick, S.: Simultaneous Tokenization and Part-of-Speech Tagging for Arabic without a Morphological Analyzer. In: Proceedings of ACL 2010 (2010)
16. Diab, M., Kadri, H., Daniel, J.: Automatic tagging of Arabic text: from raw text to base phrase chunks. In: Proceedings of the 2004 Conference of the North American Chapter of the ACL (2004)
17. Habash, N., Rambow, O.: Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In: Proceedings of the 43rd Annual Meeting on ACL, Ann Arbor, Michigan (2005), doi:10.3115/1219840.1219911
18. Al Shamsi, F., Guessoum, A.: A hidden Markov model-based POS tagger for Arabic. In: Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France, pp. 31–42 (2006)
19. Albared, M., Omar, N., Ab Aziz, M., Ahmad Nazri, M.: Automatic Part of Speech Tagging for Arabic: An Experiment Using Bigram Hidden Markov Model. In: Yu, J., Greco, S., Lingras, P., Wang, G., Skowron, A. (eds.) RSKT 2010. LNCS, vol. 6401, pp. 361–370. Springer, Heidelberg (2010), doi:10.1007/978-3-642-16248-0_52
20. Albared, M., Omar, N., Ab Aziz, M.J.: Arabic Part Of Speech Disambiguation: A Survey. International Review on Computers and Software, 517–532 (2009)

21. El Hadj, Y., Al-Sughayeir, I., Al-Ansari, A.: Arabic Part-Of-Speech Tagging using the Sentence Structure. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt (2009)
22. Goweder, A., De Roeck, A.: Assessment of a Significant Arabic Corpus. In: Proc. of Arabic NLP Workshop at ACL/EACL (2001)
23. Dukes, K., Habash, N.: Morphological Annotation of Quranic Arabic. In: Language Resources and Evaluation Conference (LREC), Valletta, Malta (2010)
24. Viterbi, A.J.: Error bounds for convolution codes and an asymptotically optimal decoding algorithm. IEEE Transactions on Information, 260–266 (1967)
25. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge(1998)
26. Carrasco, R.M., Gelbukh, A.: Evaluation of TnT Tagger for Spanish. In: Proceedings of the 4th Mexican international Conference on Computer Science. IEEE Computer Society, Washington, DC (2003)
27. Mihalcea, R.: Performance analysis of a part of speech tagging task. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, pp. 158–167. Springer, Heidelberg (2003)
28. Samuelsson, C.: Handling sparse data by successive abstraction. In: COLING 1996, Copenhagen, Denmark (1996)