A Statistical Part-of-Speech Tagger for Persian

Mojgan Seraji
Department of Linguistics and Philology
Uppsala University, Sweden
mojgan.seraji@lingfil.uu.se

Abstract

This paper presents the statistical part-of-speech tagger HunPoS trained on a Persian corpus. The result of the experiments shows that HunPoS provides an overall accuracy of 96.9%, which is the best result reported for Persian part-of-speech tagging.

1 Introduction

Data driven (machine learning) techniques for word sense disambiguation have always been a very active field and have attracted great attention from many researchers in the computational linguistics community. One of the usages of these methods is in the task of automatic part-of-speech tagging and that has resulted in some successful data driven part-of-speech taggers such as MXPOST (Ratnaparkhi, 1996) based on the maximum entropy framework, the memory-based tagger (MBT) (Daelemans et al., 1997), Brill's tagger based on transformation-based learning (TBL) (Brill, 1995) and Trigram 'n' Tags (TnT) based on Hidden Markov models (Brants, 2000). More recent work on data-driven taggers include conditional random fields and support vector machines (Kumar and Gurpreet Singh, 2010) (Gimenez and Marquez, 2004).

HunPoS (Halacsy et al., 2007) is an open source part-of-speech tagger that was released as a reimplementation of TnT. The user can tune the tagger by using different feature settings depending on the language type. Hitherto, a lot of models and implementations have been designed and are already available for the task of tagging and most of them have been tested for English and other languages but not many have been tested on Persian texts. However, some statistical tagging methods; namely a memory-based tagging (MBT) approach and Maximum Likelihood Estimation (MLE), as well as TnT have been tried out, but comparing to other languages like English it is not sufficient. Therefore, the evaluation of other part-of-speech taggers like HunPoS would be of great interest to discover how the tagger performs when applied to Persian compared to other data-driven taggers. This paper describes an evaluation of the performance of the part-of-speech tagger HunPoS on Persian. We apply the tagger on BijanKhan corpus (Bijankhan, 2004) and vary the features used for tagging seen and unseen tokens. This paper contains the following sections. Section 2 presents the open source tagger HunPoS. Section 3 describes briefly the classification, the properties and the script of Persian, prior studies of some statistical tagging methods and also introduces BijanKhan's corpus. In section 4 the design of this experiment follows and it introduces the experimental set-up. Section 5 describes the results of the evaluation. Finally, section 6 concludes this study.

2 HunPoS

HunPoS is an open source reimplementation of TnT that is based on Hidden Markov Models (HMM) with trigram language models, allowing the user to tune the tagger by using different feature settings. The tagger is similar to TnT with the difference that it estimates emission/lexical probabilities based on current and previous tags. One additional difference compared to TnT stands in the fact that the tagger is open source whereas TnT is not. The strong side of TnT, namely its suffix-based guessing algorithm that is used for handling unseen words is also implemented in HunPoS. Moreover, HunPoS inserts a morphological analyzer to narrow down the list of alternatives (possible tags) that the algorithm needs to deal with, which not only speeds up search but also very significantly improves precision. In other words, the morphological analyzer generates the possible tags, to which the weights are assigned by suffix-

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 340–343

based gussing algorithm (Halacsy et al., 2007).

3 Persian

3.1 The Persian Language

Persian, also known as Parsi or Farsi belongs to the Indo-Iranian languages, a subfamily of the Indo-European languages. Persian is spoken in Iran (Farsi), Afghanistan (referred to as Dari) and Tajikistan (referred to as Tajiki). The language has been greatly influenced by Arabic vocabulary and has the same alphabet including four additional letters; پ ، چ ، ژ ، گ, which are the sounds of [p], [ʧ], [ʒ], [g], and texts are written from right to left. Although Persian is classified as a SOV language, colloquial speech does not usually follow this order. Assi and Abdolhosseini (2000) notes that the existence of a direct object marker enables the speakers of Persian to use subjects and objects in a free word order. In addition, there are no gender distinctions in Persian as there are for example in English (she/he). Possessiveness is indicated by the genitive morpheme -e (ezafeh) in a conversation but it is invisible in writing. Adverbs can appear virtually everywhere in a sentence and adjectives can follow or precede nouns. In Persian there are several plural markers; "-hâ" and "-ân", Arabic plural suffixes such as "-ât", "-in" and "un" (used only for words of Arabic origin). There is also a plural form in Persian that follows the Arabic template morphology and is called "broken plural".

3.2 Prior Studies of Some Statistical Tagging Methods

The lack of a perspicuous morphology in Persian for marking boundaries in an SOV system makes it difficult to determine where the subject ends and where the object begins. With respect to all the factors existing in Persian such as the complex verbal paradigm as well as the highly ambiguous structure of the noun phrase and so forth, quite good results have been reported on the performance of several part-of-speech tagging methods such as TnT, memory-based tagger (MBT) and Maximum Likelihood Estimation (MLE) (Raja et al., 2007). The utilized corpus in these experiments is the BijanKhan corpus, consisting of nearly 2.6 million words. Training and test set were created by randomly dividing the corpus into two parts with an 85% to 15% ratio and each experiment repeated five times in order to avoid accidental results. The overall accuracies reported for the three taggers in due order are 96.6%, 96.6%, and 95.9% (Raja et al., 2007).

3.3 Corpus

BijanKhan corpus was introduced in 2004 as the first manually tagged Persian (Farsi) corpus in Iran. The corpus is basically gathered from daily news and common texts, and consists of syntactic and semantic annotation of nearly 2.6 million words, done by Prof. M. BijanKhan (and several linguistics students following a particular instruction) prepared at the Research Center of Intelligent Signal Processing (RCISP) in Tehran. The corpus comes with statistical software for the calculation and extraction of language features such as: conditional distribution probability, word frequency, and recognition of homonyms, synonyms, concordances and lexical order with report functionality. In addition, the corpus original tag set contains 550 tags and are organized in a tree structure. The tag name starts with the name of the most general tag and continues with the names of the subcategories until it reaches the name of the leaf tag. An example of a hierarchical tag in third level of depth can be "N_PL_LOC"; where "N" represents noun, "PL" shows the tag plurality, and "LOC" defines the tag as location. This enormous number of tags are used to attain a fine grained part-of-speech tagging that discriminates the subcategories in a general category but since this vast amount of tags makes any machine learning process impracticable Oroumchian et al. (Oroumchian et al., 2006), decided to reduce the number of tags to 40. All tags with three or more levels in hierarchy were accordingly reduced to two-level tags; in other words, the above example reduced to "N_PL". Some two-level tags that were unnecessarily too specific were also reduced to one-level tags. More specifically, these tags are conjunctiones, morphemes, prepositions, pronouns, prepositional phrase, noun phrase, conditional prepositions, objective adjectives and wishes, quantifiers and mathematical signatures (Oroumchian et al., 2006). The corpus was processed in 2007 in order to be more suitable for NLP tasks. This version of BijanKhans corpus is in Unicode text format.

4 Experimental Set-up

This experiment has two phases, model selection and model assessment. The goal of choosing these two phases was to use model selection for estimating the performance of different models in order to choose the best one, and model assessment for having chosen a final model and estimating its generalization error on new data. The corpus was split into a training set for learning or fitting the models, a validation set (development test set) for validating and estimating prediction error for model selection, and a test set preserved for testing and evaluating the generalization error for the final chosen model. The size of each set was 80%, 10% and 10%, respectively, while in the model assessment the sample data was divided into 90% for training and 10% for testing. Prior to tagging we need to train the tagger on a suitable tagged corpus (the Bijankhan corpus) in order to build a model. The tagging process requires two files containing the model built by the training process and an untagged (raw) corpus. The untagged corpus, as its name indicates, contains no part-of-speech tags and it has only one column consisting of one token per line. Since the tagger has several training options we tried to make use of this flexibility by setting several parameters for training. Therefore, we ran several experiments to train the tagger with different feature settings and combining these as well. We experimented with the order of the tag transition probability by setting the option -t to either bigram tagging or the default trigram tagging in order to estimate the probability of a tag based on the previous tags. We also examined the order of the emission probability -e for estimating the probability of a token based on the tag of the token itself as well as the previous tags. For tag distributions of unseen words based on tag distributions of rare words (words seen less than N times in the training corpus) we used the option -f with the default value 10. Finally, we tested the -s parameter that sets the length of the longest suffix to be considered by the algorithm when it estimates an unseen words tag distribution with the default value 10. It is noteworthy that the most desirable possible value of this parameter (-s) may depend on the morphology and orthography of the language involved (Halacsy et al., 2007). Thus, we tested suffixes of length 10 (the default value), 8 and 4.

Tag Transitions	Word Emissions	Accuracy
bigram	unigram	95.8%
bigram	bigram	95.8%
trigram	unigram	96.0%
trigram	bigram	96.0%

Table 1: Comparison of different models for tag transitions and word emissions.

Max Suffix Length	Max Frequency	Accuracy
10	10	96.0%
8	10	96.0%
4	10	95.9%

Table 2: Comparison of different models for unseen words.

5 Results of the Evaluation

5.1 Model Selection

For the purpose of evaluating the results, the tagged file by HunPoS was compared to the gold standard (the original manually tagged validated file) and the differences were registered. We have evaluated the performance of HunPoS from different aspects: the accuracy of the assigned tags, precision, recall and F score (harmonic mean of the precision and recall) for different part-of-speech tags, training the tagger with different feature settings for the tagging lexical probabilities as well as for the treatment of unseen words. The results of training the tagger with a combination of different feature settings showed that by applying the trigram models, as could be predicted, we achieved a higher accuracy than with the bigram models (Table 1). In order to examine the tagger performance for unseen words we had the possibility to vary the length of the suffixes. Therefore, since the optimal value of this parameter can be dependent on the morphology and orthography of the language, we tested suffixes of length 10 (the default value), 8, and 4. Looking at the results appearing in Table 2, we can infer a decrease in accuracy when reducing the length of the suffixes. Thus, for Persian, suffix length set to 10 yields the best results. The accuracy of the model selection as it is depicted in Table 3 is 96.0%.

5.2 Model Assessment

Finally, in the model assessment, we augmented the size of the training data from 80% to 90% by adding the validation set (the development test set)

Total Tokens	268424
Tokens Correctly Tagged	257794
Tokens Incorrectly Tagged	10630
Accuracy	96.0%

Table 3: Tagger performance in the model selection

to the training set and using the 10% test set that we had preserved from the beginning of this study for evaluation. In order to evaluate the results of the model assessment, the file tagged by HunPoS was compared to the gold standard (the original manually tagged test file) and the differences were recorded. Results in Table 4 shows the accuracy achieved in the model assessment. However, we can also conclude that the tagger performance was probably influenced by the size of the training set, since the accuracy increased with the extension of the training data.

Total Tokens	268008
Tokens Correctly Tagged	259618
Tokens Incorrectly Tagged	8390
Accuracy	96.9%

Table 4: Tagger performance in the model Assessment

6 Conclusion

An evaluation of the open source part-of-speech tagger HunPoS on Persian was presented here. We applied the tagger to a Persian corpus and trained it with different feature settings. The experimental results revealed an overall accuracy of 96.9% for the Persian language. By training the tagger with different feature settings in this study we can deduce that applying the default settings of the tagger can yield the best results for Persian. Moreover, the size increment of the training data in the model assessment (90%) led the system to achieve higher accuracy. Finally, to conclude this paper, we can state that with respect to the performance of other data-driven part-of-speech taggers, such as TnT, memory-based tagger, and Maximum Likelihood Estimation, HunPoS is a good alternative for part-of-speech tagging of Persian. The results reported in this paper are the best published results so far, although the scores may not be directly comparable to those of Raja et al. (2007) because we do not know whether the two studies used the same training-test split.

References

Mostafa S. Assi and Haji M. Abdolhosseini. 2000. *Grammatical Tagging of a Persian Corpus*. International Jouranal of Corpus Linguistics 5(1):69-82.

Mahmood Bijankhan. 2004. *The Role of the Corpus in writing a Grammar: An Introduction to a Software*. Iranian Journal of Linguistics, 19.

Thorsten Brants. 2000. *TnT a Statistical Part-of-Speech Tagger*. In Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-00), Seattle, Washington, USA.

Erik Brill. 1995. *Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging*. Computational Linguistics, 21:543-566.

Walter Daelemans, Jakob Zavrel, Peter Berck, and Steven Gillis. 1997. *A memory-based part-of-speech tagger generator*. In Eva Ejerhed and Ido Dagan, editors, Proceedings of the Fourth Workshop on Very Large Corpora.

Jesus Gimenez and Lluis Marquez. 2004. *SVMTool: A general POS tagger generator based on Support Vector Machines*. In LREC, Lisbon, Portugal.

Peter Halacsy, Andras Kornai, and Csaba Oravecz. 2007. *Hunpos an open source trigram tagger*. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume Proceedings of the Demo and Poster Sessions. Association for Computational Linguistics, Prague, Czech Republic, pages 209-212.

Dinesh Kumar and Josan Gurpreet Singh. 2010. *Part of Speech Tagger for Morphologically Rich Indian Languages: A Survey*.

Farhad Oroumchian, Samira Tasharofi, Hadi Amiri, Hossein Hojjat, and Fahimeh Raja. 2006. *Creating a Feasible Corpus for Persian POS Tagging*. Technical report, no.TR3/06, University of Wollongong in Dubai.

Fahimeh Raja, Hadi Amiri, Samira Tasharofi, Hossein Hojjat, and Farhad Oroumchian. 2007. *Evaluation of part-of-speech tagging on Persian text*. The Second Workshop on Computational approaches to Arabic Script-based Languages, Linguistic Institute Stanford University.

Adwait Ratnaparkhi. 1996. *A maximum entropy model for part-of-speech tagging*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955