JéTou 2011, Toulouse, April 7-8 2011

Statistical part of speech tagger for Persian words

Ali Asghar Behmanesh[1] Abdol Hamid Pilevar[2]

(1)NLP Lab, Computer Engineering Dept., Bu Ali Sina University
(2)NLP Lab, Computer Engineering Dept., Bu Ali Sina University
a.behmanesh@basu.ac.ir, pilevar@basu.ac.ir

Abstract: Corpora tagged with Part of speech (POS) information are often used as a prerequisite for more complex NLP applications such as information extraction, syntactic parsing, machine translation or semantic field annotation. They are also used to help train statistical models. This paper presents a Maximum Likelihood Estimation (MLE) method for evaluation of part of speech tagging on Persian texts. The MLE approach has been used for handling the unknown words in the proposed methods. Three pre-processing technique is implemented for improving the accuracy of the results. The experiments have been conducted on a manually part of speech tagged Persian corpus with over two millions of tagged words. The best accuracy that was achieved by the proposed MLE tagging methods was 96.07%, which it is satisfactory compare to some other known similar methods.

Résumé: Corpus étiqueté avec la partie du discours (POS) des informations sont souvent utilisés comme une condition préalable pour des applications plus complexes telles que la PNL extraction d'information, l'analyse syntaxique, la traduction automatique ou annotation champ sémantique. Ils sont également utilisés pour aider les modèles de formation statistique. Cet article présente un maximum de vraisemblance (MLE) méthode d'évaluation de la partie du discours de marquage sur les textes persans. L'approche MLE a été utilisé pour le traitement des mots inconnus dans les méthodes proposées. Trois technique de pré-traitement est mis en œuvre pour améliorer la précision des résultats. Les expériences ont été menées sur une partie du discours a ajouté manuellement corpus Persique avec plus de deux millions de mots associés. La meilleure précision qui a été réalisé par le MLE proposé méthodes de marquage a été 96,07%, ce qui est satisfaisant comparer à quelque autre connue des méthodes similaires.

Keywords: Natural Language Processing, Part of Speech Tagging, MLE tagger

Mots-clés: traitement du langage naturel, la partie du discours de marquage, tagger MLE

1 Introduction

Ensuring consistency of POS tagging plays an important role in constructing high-quality Persian corpora. It is the process in which each word is assigned to a corresponding POS tag that describes how this word be used in a sentence. Typically, the tags can be syntactic categories, such as noun, verb and so on. Firstly, Persian words do not have a strict one-to-one correspondence between their POS categories and functions in a sentence. Secondly, an ambiguous Persian word can act as different POS categories in different contexts without changing its form. Thirdly, there are many out-of-vocabulary (OOV) words in real Persian text whose POS categories are not defined in the dictionary used. All these factors make it much more difficult to achieve a high-performance POS tagger for Persian

In section 2 a brief about some POS tagging methods which are developed based on the BIJANKHAN corpus[i] is explained, this corpus is described in section 3. Section 4 describes about two implementations of Maximum Likelihood Estimation in POS tagging, which the proposed method is included in this section. Section 5 shows our experimental results and its comparison with some other similar MLE methods. Section 6 is the conclusion.

Previous work

Several part of speech tagging systems were developed on the well-known BIJANKHAN Farsi tagged corpus that contains 2,500,000+ tokens and each of them has different precision and accuracy. Forty tags are used In these taggers.

TnT provides overall tagging accuracy of 96.64%, specifically, 97.01% on known words and 77.77% on unknown words. TnT tagger shows 19.24% points accuracy difference between the words seen and those are not seen before (Tasharofi et al , 2006).

The overall part-of-speech tagging accuracy of Memory Based Tagger (MBT) is around 96.42%. MBT shows about 20% difference between accuracy of POS tagging for known and unknown words (Amiri et al, 2006).

The average overall accuracy for Hidden Markov Model (HMM) tagger is 95.11%. The accuracy of the known and unknown words are 96.136% and 60.25%, respectively (Azimizadeh et al, 2008). The best accuracy that was achieved by MLE taggers was 95.97% (Raja et al, 2007).

2 The Corpus

The corpus which was used in our experiments is a part of the BijanKhan's tagged Persian corpus (BijanKhan, 2004), which is maintained at the Linguistics laboratory of the University of Tehran. The training part contains about 2.2 Millions words, and the test data includes about 400,000 words. The corpus is gathered from daily news and common texts. Each document is assigned a subject such as political, cultural and so on. Totally, there are 4300 different subjects. This subject categorization provides an ideal experimental environment for clustering, filtering, categorization research. In this paper the subject categories of the documents is ignored and only the POS tags are considered. The corpus is tagged with a rich set of 550 tags. This vast amount of tags were used to achieve a fine grained part-of-speech tagging, i.e. a tagging that discriminates the subcategories in a general category. The large size of tags makes the automatic learning process impracticable. So, the number of tags was reduced (Oroumchian et al, 2006).

After the four stages, the number of tags is reduced to 40 tags. Table 1 shows the frequencies of the tags in the Bijankhan,s corpus.

In table 1, the tags and their corresponding frequencies in the corpus is depicted. Looking carefully in the table 1 reveals that the tag "N_SING" (singular noun) is the most frequent tag in the corpus. On the other hand, the "NN" tag with only two occurrences is the least one.

3 Implementation

The Maximum Likelihood Estimation (MLE) method is selected for its simplicity and ease of implementation. For constructing the training set, the words with greater tags are selected. In this regard, the maximum likelihood probability is used to calculate the assigned values for each word in the training set. The tags with the greater maximum likelihood probability is picked and assigned as a tag for the words. In order to evaluate this method, the words in the test set are analyzed and the designated tags are assigned to the words in the test set.

Tag name	Frequency in	Tag name	Frequency in corpus
ADJ	22	MQUA	361
ADJ_CMPR	7443	MS	261
ADJ_INO	27195	NN	2
ADJ_ORD	6592	NP	52
ADJ_SIM	231151	N_PL	160419

Table 1: Distribution of Tags in the Bejankhan corpus

Three methods for tagging the unknown words are proposed which are explained in this section:

DEFAULT method

In this method, the unknown words (the words those are not seen in the training set) are tagged as "default". The accuracy of this method is relatively very low (0.06%). Table 1 show that few words (only 192 words) are tagged as "default". Implementation results of this method are shown in Table 4.

3.1 N-SING method

As we can see in Table 1, 967545 words are tagged as "N_SING" (Singular Noun), which is noticeable compare to the total number of the words in the corpus; therefore we suggested that unknown words could be most probably of this kind and labeled the unknown words with "N_SING" tag. Implementation results of this method are shown in Table 5. This approach improves the overall accuracy to 95.37% and boosts the accuracy of the right placement of the unknown words to 54.79%.

3.2 PRE-POST method

In this method due to structure of the words by looking at the prefixes and suffixes the words are tagged. As shown in Table 2, words like "بهنوش" and "بهنام" are tagged 3130 times as N_SING and 69084 times with P tag. As we can see in the tables 2, all of the words which are started or ended with this kind of prefixes or suffixes are tagged with N_SING, therefore: in this method the following decision are made for tagging the unknown words:

Prefix	Tag name	frequencies	Example
به	N_SING	3130	بهنوش ، بهنام
	p	69084	
با	N_SING	11642	باهوش ، با هنر

Table2: two of mainly used prefixes, assigned tag names, and their frequencies in the corpus

1- All of the unknown words which are started with this type of prefixes are tagged with N_SING.

2- All of the unknown words which are ended with this type of suffixes are tagged as N_SING.
3- All of the unknown words which are not ended or started with this type of suffixes or prefixes are tagged as DEFAULT.

The results of implementation this method are depicted in Tables 6 and 7. The unknown words which are tagged as DEFAUL, in average have been 43.32% correct, and in average 69.35% of the words which are tagged with N_SING were correctly recognized.

run	Known words	Unknown Words	Overall
1	96.22%	0.01%	93.98%
2	96.58%	0.03%	94.69%
3	96.45%	0.12%	94.67%
4	96.90%	0.06%	95.25%
5	96.88%	0.09%	95.05%
Average	96.61%	0.06%	94.73%

Table4: Accuracies results of implementing DEFAULT method

#run	Known words	Unknown Words	Overall
1	96.22%	52.09%	95.19%
2	96.58%	53.49%	95.73%
3	96.45%	58.09%	95.74%
4	96.90%	56.69%	96.22%
5	96.88%	54.48%	96.08%
Average	96.61%	54.79%	95.79%

Table5: Accuracies results of implementing N-SING method

#run	Knownwords	Unknown words	overall
1	96.22%	40.59%	94.22%
2	96.58%	42.07%	95.51%
3	96.45%	42.03%	95.44%
4	96.90%	47.40%	95.06%
5	96.88%	44.49%	95.89%
average	96.88%	43.32%	95.57%

Table6: Accuracies results of implementing PRE-POST method, while tagging unknown words with DEFAULT

#run	Known words	Unknown words	overall
1	96.22%	67.77%	95.49%
2	96.58%	68.70%	96.03%
3	96.45%	70.93%	95.98%
4	96.90%	72.43%	96.48%
5	96.88%	69.90%	96.37%
average	96.61%	69.35%	96.07%

Table7: Accuracies results of implementing PRE-POST method, while tagging unknown words with N-SING

4 Results and Discussion

A software system is developed using the Perl programming language and three presented methods are tested with this system. The experimental results show that the DEFAULT method is not efficient; the N_SING method is acceptable, while the PRE-POST method is efficient.

We have compared the experimental results of our presented methods with some other similar POS tagging systems which are developed based on well-known Bijankhan Farsi tagged corpus and the comparison result is shown in Table 8.

The comparison of the accuracy rates of eight different unknown words tagging methods are shown in Table 8. As we can see in this table; the worst one is the DEFAULT method which improves the results only 0.06%. PRE-POST method while the N_SING tag is selected, shows 69.35% accuracy which is the best result between the presented methods. Therefore the accuracy of PRE-POST method (69.35%) is better than HMM (60.25%) and MLE (65.75%) methods. Its accuracy (69.35%) and overall (96.07%) results are very close to same in MBT (75.15%, 96.42%) and TnT (77.77%, 96.64%) methods.

Method	known Words	Unknown Words	overall
DEFAULT	96.61%	0.06%	94.73%
N_SING	96.61%	54.79%	95.79%
PRE-POST (selecting DEFAULT tag)	96.88%	43.32%	95.57%
PRE-POST (selecting N_SING tag)	96.61%	69.35%	96.07%
MBT	96.86%	75.15%	96.42%
TnT	97.01%	77.77%	96.64%
HMM	96.136%	60.25%	95.11%
MLE (Raja)	96.60%	65.75%	95.97%

Table8: Overall comparison of presented methods with four other similar methods

5 Conclusion

In this paper, the Maximum Likelihood Estimation method is used for POS tagging Persian texts. Three methods are presented for tagging the unknown words. The experimental results show that these simple heuristics methods have significant impact on improving the tagging results of the unknown words.

The comparison of the experimental results of some of the known MLE methods for Persian texts with our presented methods show that; the overall accuracy of one of the presented methods are close to the best taggers (TnT and MBT). Considering the simplicity of this method, and its efficiency, it can be used as a base for tagging the Persian texts.

References

Amiri H, Raja F, Sarmadi M, Tasharofi S, Hojjat H, Oroumchian F. A Survey of Part of Speech Tagging in Persian. *Submitted to Journal of Computer Speech & Language, Elsevier Ltd 2007.*

BijanKhan M. (2004). The Role of the Corpus in Writing a Grammar: An Introduction to a Software. *Iranian Journal of Linguistics*, 19(2).

Oroumchian F., Tasharofi S., Amiri H., Hojjat H., Raja F. (2006). Creating a Feasible Corpus for Persian POS Tagging. *Technical Report, no. TR3/06, University of Wollongong (Dubai Campus*).

Tasharofi S., R Aja F., O Roumchian F. & R Ahgozar M. (2007). Evaluation of Statistical Part of Speech Tagging of Persian Text*. In International Symposium on Signal Processing and its Applications, Sharjah, E.A.U.*

Azimizadeh, A., Mehdi M., Rahati S. (2008). Persian partof speech tagger based on Hidden Markov Model. *9th International Conference on the Statistical Analysis of Textual Data*.

Dermatas E., Kokkinakis G. (1995). Automatic Stochastic Tagging of Natural Language Texts. *Computational Linguistics*, 21(2), 137-163.

Mikheev A. (1996). Learning Part-of-Speech Guessing Rules from Lexicon: Extension to Non- Concatenative Operations. *In Proceedings of COLING*.

[i] Available at http://ece.ut.ac.ir/dbrg/bijankhan/