2008

Using heuristic rules to improve Persian part of speech tagging accuracy

Mitra Mohtarami
Research and Development Center, Safa Rayaneh Company

Hadi Amiri
University of Tehran

Farhad Oroumchian
University of Wollongong in Dubai, farhado@uow.edu.au

Masoud Rahgozar
University of Tehran

Recommended Citation

Mohtarami, Mitra; Amiri, Hadi; Oroumchian, Farhad; and Rahgozar, Masoud: Using heuristic rules to improve Persian part of speech tagging accuracy 2008.
https://ro.uow.edu.au/dubaipapers/52

International Conference on Information and Knowledge Engineering (IKE'08),Las Vegas, USA, July 14-17, 2008

Using Heuristic Rules to Improve Persian Part of Speech Tagging Accuracy

Mitra Mohtarami
Research and Development Center, Safa Rayaneh Company, Tehran, Iran
m.mohtarami@yahoo.com

Hadi Amiri
Database Research Group, School of ECE, University of Tehran, Iran
h.amiri@ece.ut.ac.ir

Farhad Oroumchian
Department of IT University of Wollongong, Dubai, UAE
FarhadO@uow.edu.au

Masoud Rahgozar
Database Research Group, School of ECE, University of Tehran, Iran
m.rahgozar.ut.ac.ir

Abstract

One of the major activities in Natural Language Processing is determining a word's part of speech (POS) tag. In this research we focus on improving the accuracy of Persian part of speech tagging by applying post processing heuristic rules. To evaluate the effects of those rules we use Bijankhan tagged corpus and for tagging, Maximum Likelihood Estimation (MLE) approach is selected because of its simplicity and the ease of implementation. Furthermore, we have studied the effect of size of training on the accuracy of the MLE method. The experimental results show that the heuristic rules improve the accuracy especially for the unknown words.

1. Introduction

Part of speech tagging (POS) is the task of annotating each word in a text with its most appropriate syntactic category. Having an accurate POS tagger is useful in many information related solutions such as information retrieval, information extraction, text to speech systems, linguistic analysis, etc.

A POS tagging solution has two major steps. First step is finding the possible set of tags for each word regardless of its role in the sentence and the second step is choosing the best tag among possible tags based on its context. There are several proposed approaches for generating a POS tagger. Hidden Markov Models are statistical methods which choose the tag sequence which maximizes the product of lexical probability and the contextual probability. This method is applied successfully on different languages such as German [2], English [2, 3], Slovene [4] and Spanish [5].

Another approach is rule-based which uses some rules and a lexicon to resolve the tag ambiguity. These rules can either be hand-crafted or learned [10]. Other machine learning models used for tagging include maximum entropy and other log-linear models, decision trees, memory-based learning, and transformation based learning [6, 7].

In this research, we focus on memory-based learning methods. Memory-based taggers are trained with a training set and they use learned information to tag a new text. The tagger that is used in this research is MLE tagger [8, 9] and the corpus is BijanKhan's tagged corpus [1, 9]. In addition to study the MLE accuracy with different test and training sets, this research provides some heuristic rules to improve the accuracy of the Persian POS taggers.

The previous experimental results show that the overall accuracy of the MLE method for unknown words (the word that was not seen before in the training set) is low. This value is 0.032% for the first model of MLE, MLE-DEFAULT, and is %54.78 for the second model of MLE, MLE-N_SING [9]. The new heuristic rules used in this study has improved the accuracy of both MLE methods for unknown words, 19.49% improvement for MLE-DEFAULT and 11.43% improvement for MLE-N_SING. Finally the overall accuracies have improved by 1.50% for MLE-DEFAULT and 0.86% for MLE-N_SING.

In section 2 we describe the MLE tagger. The new heuristic rules are exhibited in section 3. Section 4 explains the evaluation environment, the MLE accuracy and the effect of applying heuristic rules. Finally, section 5 is the conclusion and future work.

2. Maximum Likelihood Estimation

In this section we present the maximum likelihood estimation approach for POS tagging. For every word in the training set, MLE calculates the tag which is assigned to the word more often than the other tags [10]. For this purpose, MLE calculates the maximum likelihood probabilities for each tag assigned to any word in the training set. Then it picks the tag with the greater maximum likelihood probability for each word and makes it the only tag assignable to that word. This tag is called the *designated* tag for that word. For the purpose of tagging, this method analyzes the words in the test set and assigns the *designated* tags to the words in the test set [10, 9]. There are two different models of MLE method namely, MLE-N_SING and MLE-DEFAULT (N_SING and DEFAULT are the name of 2 tags in the tag set). MLE-DEFAULT assigns the “DEFAULT” tag to each unknown word while MLE-N_SING assigns the “N_SING” tag, which is the most frequent tag in the collection, to each unknown word. Hence, the designated tag for unknown words in MLE-DEFAULT model is “DEFAULT” and in MLE-N_SING model is N_SING.

We ran the MLE Estimation on five different test and training sets of reduced-tags Bijankhan corpus, the same five sets that is used in [8, 9]. These sets had generated by randomly dividing the corpus into two parts with an 85% to 15% ratio. Table 1 shows the test collections.

Table 1 Test and Training Sets Used in [8, 9]

Run	**Training Tokens/Percent**	**Test Tokens/Percent**	**Total**
1	2196166 / 84.52	402050 / 15.47	2598216
2	2235558 / 86.04	362658 / 13.96	2598216
3	2192411 / 84.38	405805 / 15.61	2598216
4	2178963 / 83.86	419253 / 16.13	2598216
5	2186811 / 84.16	411405 / 15.83	2598216
Avg.	2197982 / 84.59	400234.2/ 15.40	

As Table 1 shows, in [8, 9] the authors in each run have considered %85 of the collection as training set and the remaining, %15, as the test set and then evaluated the accuracy of MLE. However a good idea is to study the accuracy of MLE method with different distribution of words in test and training sets. So, for further investigation, especially unknown words, we divided the reduced-tag Bijankhan collection into different test and training sets. Our main purpose was to study the accuracy of MLE in a situation that the amount of training data is low and the effect of it on the MLE’s accuracy. The information of the new sets is shown in Table 2.

Table 2 Test and Training Sets with Different Distribution in Test and Training Sets

Run	**Training Tokens/Percent**	**Test Tokens/Percent**	**Total**
6	2208488 / 85.00	389728 / 15.00	2598216
7	1948695 / 75.00	649521 / 25.00	2598216
8	1688874 / 65.00	909342 / 35.00	2598216
9	1429038 / 55.00	1169178 / 45.00	2598216
10	1169236 / 45.00	1428980 / 55.00	2598216

Table 3 and 4 show the accuracy of two models of MLE tagging quoted from [9]. As it is shown in this table the accuracy of both two MLE models is the same for known words because these models behave differently only with unknown words. Both tables show that the accuracy of MLE is very low for unknown words from nearly 0.1% for MLE-DEFAULT to nearly 50% for MLE-N_SING.

Table 3 Accuracy of MLE-DEFAULT

Run	**Known words**	**Unknown words**	**Overall**
1	96.50%	0.12%	94.55%
2	96.78%	0.16%	94.91%
3	96.53%	0.18%	94.53%
4	96.53%	0.09%	94.51%
5	96.64%	0.23%	94.68%
Avg.	**96.60%**	**0.15%**	**94.63%**

Table 4 Accuracy of MLE-N_SING

Run	**Known words**	**Unknown words**	**Overall**
1	96.50%	52.60%	95.61%
2	96.78%	56.63%	96.00%
3	96.53%	51.49%	95.59%
4	96.53%	55.48%	95.67%
5	96.64%	54.34%	95.78%
Avg.	**96.60%**	**54.11%**	**95.73%**

To find out the problem we investigated the unknown words and their tags in the test collection. After examining the unknown words in the collection, we have come up with some heuristic rules to improve the accuracy of MLE models.

3. Heuristic Rules

In this section we present some heuristic rules to improve the accuracy of MLE for unknown words.

According to our investigation we realized some interesting points. Firstly, as it truly mentioned in [8, 9] the correct tag for most of the unknown words is "N_SING". That explains why the MLE- N_SING model which selects "N_SING" as designated tag has better accuracy. Second, some of the unknown words those were plural nouns ("N_PL") tagged as "DEFAULT" or "N_SING" by MLE models incorrectly. In Persian language, most of plural nouns end with suffixes like "ها", "گان", "ان","ات" etc. For example the word "کتاب" (book in English) is a singular noun ("N_SING") and "کتاب ها" (books) is its plural form ("N_PL"). So, it is possible to process the output of the MLE method (or any other POS tagging method) with this simple heuristic as: if an unknown word ends with any of the plural suffixes it should be tagged as “N_PL”. It should be mentioned that we just look at the head and tail of the words to detect their prefixes and suffixes. However, this solution doesn’t work for all such words. As an example consider the word “بینی ات” (your nose in English). This word has the substring “ات” at its end as suffix but it is a single noun. So based on this heuristic it will be tagged incorrectly as "N_PL". Similar heuristics could be formed for many of the part of speech tags. Table 5 lists part of speech tags with their most common suffixes and prefixes presented in [11]. In this paper we name this rule sets *First-Rule* set.

Table 5 Unknown Words Features- First-Rule Set

Real tag of the unknown word	Unknown word’s morphemes	Suffix/ Prefix
ADJ_CMPR (Comparative Adjective)	تر، تری	Suffix
ADJ_SUP (Superlative Adjective)	ترین	Suffix
N_PL (Plural Noun)	ها، های، هایی، ان، هایم، هایت، هایش، هایمان، هایتان، هایشان، ین، ات،ان،	Suffix
V_PA (Past Verb)	ام، ای، یم، ید، ند	Suffix
V_PRE (Attributive Verb)	ست	Suffix
V_SUB (Implicit Verb)	ب، ن	Prefix
V_PRS (Present Verb)	می، نمی	Prefix

In addition to Table 5 we add some other useful heuristic rules that are shown in Table 6. The combination of these rules with the rules depicted in table 5 is named *Second-Rule* set.

Table 6 Additional Unknown Words Features

Real tag of the unknown word	Unknown word’s morphemes	Suffix/ Prefix/ Word
N_PL (Plural Noun)	گان، جات، جاتی، اجات، هجات، یجات، وجات، ویان، یان، یون	Suffix
CON (Conjunction)	درهرحال، مادامي كه، گو آن كه، با وجود آنكه، ازآنجا كه، گذشته از اينكه، پيش ازاينكه، از طرفي ديگر، بالتبع، ازآن رو، به صورتي كه، بدين قرار، شگفت اينكه، به همين علت، صرفنظر از اين كه، معذالك، به اين معنا كه، به بيان ديگر، هر چه كه، از اين	Word

رو

Hence, based on these heuristics we will post process the output of taggers and for unknown words, we will modify their tags based on these suffixes or prefixes.

3. Experimental Results

In this section we show how simple heuristic rules can improve the accuracy of predicting the tags for the unknown words. In addition, we study the accuracy of MLE with different distribution of test and training sets.

At First, we study the effect of heuristic rules and the amount of improvement obtained by applying these rules. Table 7 shows the effect of heuristic rules on accuracy of MLE-DEFALT method for unknown words.

Table 7 MLE-DEFAULT Model- Unknown Words

Run	MLE-DEFAULT	First-Rule Improvement	Second-Rule Improvement
1	0.012%	17.99%(+17.98)	18.26% (+18.25)
2	0.016%	19.27%(+19.25)	19.90% (+19.88)
3	0.018%	20.25%(+20.23)	20.47% (+20.45)
4	0.090%	18.89%(+18.80)	19.07% (+18.98)
5	0.023%	21.01%(+20.99)	21.42% (+21.40)
6	0.020%	18.86%(+18.84)	18.95% (+18.93)
7	0.024%	19.56%(+19.54)	19.63% (+19.61)
8	0.031%	19.22%(+19.19)	19.36% (+19.33)
9	0.044%	19.09%(+19.05)	19.25% (+19.21)
10	0.037%	18.66%(+18.62)	18.84% (+18.80)
AVG	0.032%	19.28%(+19.25)	19.51%(+19.49)

Table 7 shows the amount of improvement, this value is shown by (+*improvement-value*) in the columns. This value shows the improvement over the MLE-DEFAULT without applying heuristic rules. The improvement achieved by using heuristic rules is very acceptable. After applying the *First-Rule* set the unknown word accuracy for MLE-DEFAULT is increased 19.25 percent in average. This value is 19.49 for the *Second-Rule* set of rules. Table 8 depicts the accuracy improvement achieved for MLE-N_SING of unknown words.

Table 8 MLE- N_SING Model- Unknown Words

Run	MLE-N_SING	First-Rule Improvement	Second-Rule Improvement
1	52.60%	63.55% (+10.95)	63.75% (+11.15)
2	56.63%	67.78% (+11.15)	68.34% (+11.71)
3	51.49%	64.20% (+12.71)	64.29% (+12.80)
4	55.48%	66.52% (+11.04)	66.64% (+11.16)

5	54.34%	66.72% (+12.38)	67.02% (+12.68)
6	53.42%	64.86% (+11.44)	64.87% (+11.45)
7	56.08%	67.63% (+11.55)	67.56% (+11.48)
8	56.07%	66.96% (+10.89)	66.96% (+10.89)
9	55.67%	66.38% (+10.71)	66.41% (+10.74)
10	56.10%	66.34% (+10.24)	66.40% (+10.30)
AVG	54.78%	66.09%(+11.30)	66.22%(+11.43)

The improvement of the *First-Rule* set for MLE-N_SING is 11.30 percent in average. This value is 11.43 percent for the *Second-Rule* set of rules.

Table 7 and 8 both show that we have achieved a reasonable amount of improvement by the heuristic rules. Both set of rules improved the accuracy of MLE for unknown words. This improvement for MLE-N_SING is less than MLE-DEFAULT. The reason is that the correct tag for most of the unknown words is "N_SING", So MLE-N_SING which selects "N_SING" as designated tag has better results for unknown words than MLE-DEFAULT which selects "DEFAULT" as designated tag. Hence weaker POS tagger benefits more from heuristic rules.

To have a better view Figure 1 illustrates the tags distribution in Bijankhan corpus [8]. The tags that are less frequent than others in the corpus are categorized into a new group called "ETC".

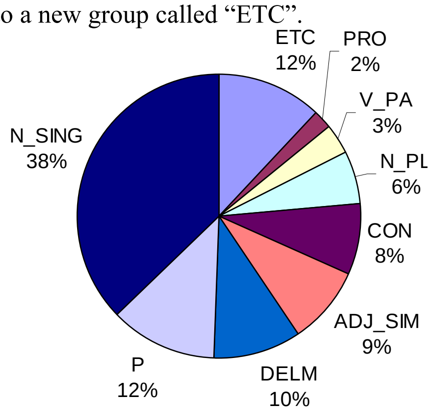

Figure 1 Tag Distribution for Collection

The tags which are selected for "ETC" group are the ones whose number of occurrences is below 5000 times in the corpus.

Table 9 and 10 show the accuracy of two models of MLE tagging after applying the heuristic rules. As it is shown in this table the accuracy of both two MLE models is the same for known words.

Table 9 Accuracy of MLE-DEFAULT- Second Rule Improvement

Run	Known words	Unknown words	Overall
6	96.24%	18.95%	93.17%
7	96.15%	19.63%	93.05%
8	96.09%	19.36%	92.83%
9	95.77%	19.25%	92.01%
10	95.38%	18.84%	91.54%

Table 10 Accuracy of MLE-N_SING- Second Rule Improvement

Run	Known words	Unknown words	Overall
6	96.24%	64.87%	94.65%
7	96.15%	67.56%	94.95%
8	96.09%	66.96%	94.85%
9	95.77%	66.41%	94.38%
10	95.38%	66.40%	93.85%

As depicted in Table 9 and 10 accuracy of the MLE models do not change dramatically when the amount of training data decreases. As an instance, the overall accuracy of MLE-DEFAULT is 93.17% for run 6 which uses 85% of the collection for training, but in spite of decreasing the amount of training data by 40% in run 10 the overall accuracy of MLE- DEFAULT does not change so much. Table 11 shows a good view of the overall accuracy of the two models with and without applying the heuristic rules.

Table 11 Overall Accuracy of MLE Models

Model	Without post processing	First-Rule Improvement	Second-Rule Improvement
DEFALT	92.23%	93.00%(+0.73)	93.77%(+1.50)
N_SING	94.43%	94.89%(+0.46)	95.29%(+0.86)

5. Conclusion and future work

This paper described experiments conducted with Maximum Likelihood approaches for POS tagging. The best MLE model, MLE-N_SING has produced an overall accurate tagging around 95.29% by using new heuristic rules which is about the best Persian POS tagger. We also experiment with different distributions of the training and test sets and investigate the effect of the size of the training on the effectiveness of the tagger. Furthermore, we have introduced a set of post processing heuristic rules that improves the performance of MLE taggers. The overall accuracy is improved by 1.5 for MLE-DEFAULT model and by 0.86 for MLE-N_SING.

In future we would like to continue these experiments with other types of Part of Speech tagging models.

6. References

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

[2] T. Brants. TnT – a Statistical Part-of-Speech Tagger. *In Proc. sixth conference on applied natural language processing ANLP-2000*, 2000.

[3] R. Mihalcea. Performance Analysis of a Part of Speech Tagging Task. *In Proc. Computational Linguistics and Intelligent Text Processing, Gelbukh A. Editor, Centro de Investigaciń en Computaciń IPN*, 2003.

[4] S. Dzeroski, T. Erjavec and J. Zavrel. Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. *In Proc. LREC 2000*. 2000.

[5] R. M. Carrasco, A. Gelbukh. Evaluation of TnT Tagger for Spanish. *In Proc. Fourth Mexican International Conference on Computer Science ENC'03,* 2003.

[6] J. Zavrel, W. Daelemans. Recent Advances in Memory-Based Part of Speech Tagging. *VI Simposio Internacionale de Comunicacion.* . 1999.

[7] Ratnaparkhi. A maximum entropy part-of-speech tagger. *In Proc. Of the Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania, 1996.

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

[9] H. Amiri, H. Hojjat and F. Oroumchian. Investigation on a Feasible Corpus for Persian POS Tagging. *12th International CSI Computer Conference (CSICC) 2007*. 2007.

[10] J. Allen. *Natural Language Understanding, Second Edition*. Benjain/Cummings Publishing Company, 1995.

[11] F. Raja, H. Amiri, S. Tasharofi, M. Sarmadi, H. Hojjat, F. Oroumchian. Evaluation of Part of Speech Tagging on Persian Text. *The Second Workshop on Computational Approaches to Arabic Script-based Languages, LSA 2007 Linguistic Institute, Stanford University, USA*, 2007.