

Creating a Feasible Corpus for Persian POS Tagging

Farhad Oroumchian

FarhadO@uowdubai.ac.ae

College of IT, University of Wollongong in Dubai

Samira Tasharofi

s.tasharofi@ece.ut.ac.ir

Hadi Amiri

h.amiri@ece.ut.ac.ir

Hossein Hojjat

h.hojjat@ece.ut.ac.ir

Fahimeh Raja

f.raja@ece.ut.ac.ir

Electrical and Computer Engineering Department, University of Tehran

Abstract

This paper describes creation of a test collection for Persian Part of Speech Tagging experiments. This collection was created by modifying a manually Part of Speech (POS) tagged Persian corpus with over two million tagged words. The original collection had a tag set of 550 tags that are more than what any machine learning algorithm can handle. The number of tags for these experiments was reduced to only 40 tags. Since the main motivation for creation of this collection is building and testing efficient Persian Part of Speech taggers, a benchmark tagger was created for that used Maximum Likelihood Estimation (MLE) for guessing the correct tags in Persian. Two different versions this tagger were produced that differed on handling "Unseen words". One of them tagged the unseen words with a "DEFAULT" tag and in the other one with "N-SING" which was the most frequent tag used in the corpus. The best accuracy that was achieved by MLE tagging was (95.43%). The results of the experiments are encouraging as they are comparable to other languages such as English German and Spanish. This test corpus will be distributed among other researchers to be used as a base for Persian natural language processing.

Keywords: *Natural Language Processing, Part of Speech Tagging, Persian Natural Language Processing, Persian Information Retrieval, Test Collection Building, Persian Test Part of Speech Collection*

ABOUT THE AUTHORS

Farhad Oroumchian is an Associate Professor in College of Information Technology at the University of Wollongong Dubai. His areas of research are in Information Retrieval, Applied Natural Language Processing and Reasoning under Uncertainty in Artificial Intelligence.

Samira Tasharofi, Hadi Amiri, Hossein Hojjat and **Fahima Raja** are graduate students in the Electrical and Computer Engineering Department, University of Tehran, Iran.

1. Introduction

Part of speech tagging 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 tag set of 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 (e.g. Church, 1988; DeRose, 1988; Cutting et al. 1992; Merialdo, 1994, etc.) are statistical methods which choose the tag sequence which maximizes the product of word likelihood and tag sequence probability. Another approach is rule-based which uses some rules and a lexicon to resolve the tag ambiguity. These rules can either be hand-crafted (Garside et al., 1987; Klein & Simmons, 1963; Green & Rubin, 1971), or learned, as in Hindle (1989) or the transformation-based error-driven approach of Brill (1992) [1]. Other machine learning models used for tagging include maximum entropy and other log-linear models, decision trees, memory-based learning, and transformation based learning.

In section 2 we describe our efforts for creating an appropriate corpus with training and test sets for POS tagging experiments. In these experiments, we used Maximum Likelihood Estimation as a benchmark for our method. Section 3 describes the benchmark approach (MLE). Section 4 compares the MLE approach with memory based and other results from other experiments. Section 5 is the conclusion.

2. Corpus Study

The corpus which was used in this research is a part of the BijanKhan's tagged Persian corpus [1], 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 400000 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 research, we simply ignored the subject categories of the documents and concentrated on POS tags.

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. Therefore, we decided to reduce the number of tags. The next section describes the elimination process used to reduce the number of tags

2.1 Selecting Suitable Tags

As mentioned earlier the corpus is tagged with 550 fine grained tags. The large number of tags reduces the frequency of occurrence of each tag in the corpus. Therefore it poses considerable challenge for machine learning approaches that need to learn the tags by using context and tag frequencies and their usage context. Most of the tools for part-of-speech tagging require much smaller tag set in the learning process. In order to create a corpus feasible for POS experiments, we decide to reduce the size of the tag set. First a statistical analysis of the corpus was conducted and frequencies of each tags

was gathered. Then by using the steps described below many of the tags were grouped together and a smaller tag set was produced.. Prof. BijanKhan's corpus uses a rich hierarchical representation for tags. Each tag in the tag set is placed in a hierarchical structure. The hierarchy is reflected in the naming convention of the tags themselves also. The name of the POS tags in the middle nodes of the hierarchy and the leaves includes the names of their parent nodes (tags). Each tag name starts with the tag name that describes the word in a more general manner, and then it follows by other tag names that define the word in more detail. As an example, consider the tag "N_PL_LOC". This tag represents a POS tag in third level of depth in the hierarchy. The "N" at the beginning of the name stands for a noun. The second part, "PL" describes the plurality of the tag, and the last part of the tag name defines the tag as about locations. For another example, the tag "N_PL_DAY" could be assigned to a word that is a noun, a plural and describes a date. The reduction in the tag set size is achieved by following the steps below:

1. In the first step, we consider those tags that have three or longer level in hierarchy and decrease them to two-level ones. Hence both of the above examples will reduce to a two-level tag, namely "N_PL". The new tag shows that they are plural nouns, but nothing about locations or dates. After applying this step to the corpus, the tag set contained only 81 tags. After this step all of tags in the tag set hierarchy only contain one or two-levels.
2. There were some numerical tags in the remaining tags. After referring to the corpus, we noticed that these tags are not correct and produced from the mistakes in the tagging process of the corpus. To prevent decrease the accuracy of part-of-speech taggers that used this corpus, we renamed them to "DEFAULT" tag. So, the number of tags in the tag set reduced to 72 tags in this step.
3. In the third step, some of the unnecessary two-level tags are reduced to one-level tags. It is important to mention that the reduction process is based on the parent-child relation between the tags, e.g. those tags that only starts with ADV can reduce to ADV tags. These unnecessary tags are those that the occurrence of them in the corpus was too rare. These tags are specifically conjunctions, morphemes, prepositions, pronouns, prepositional phrases, noun phrases, conditional prepositions, objective adjectives, adverbs that describe locations, repetitions and wishes, quantifiers and mathematical signatures. By this modification, the number of tags reduced to 42.
4. In this step we reduced the tags that appeared rarely in the corpus. These are noun (N) and short infinitive verbs (V_SNFL). We consider the semantic relationship between these two tags and the other tags in the tag set. For example, since the meaning of the tag "N" is near to the single noun tag we replace it with "N_SING". Also as the meaning of the "V_SNFL" tag is not similar to any other tags in the corpus; we simply remove it from the corpus.

After the fourth stage, the tag set contained only 40 tags. Table 1 shows most and least frequent tags.

Table 1: Distribution of Tags

Tag Name	Frequency in Training Set	Percentage in Training Set	Frequency in Test Set	Percentage in Test Set
ADJ	21	0.001	1	0.000
ADJ_CMPR	5968	0.270	1475	0.377
ADJ_INO	22503	1.020	4693	1.199
ADJ_ORD	5743	0.260	849	0.217
ADJ_SIM	192171	8.709	38980	9.961
ADJ_SUP	6342	0.287	1001	0.256
ADV	1291	0.059	224	0.057
ADV_EXM	2398	0.109	793	0.203
ADV_I	1917	0.087	177	0.045
ADV_NEGG	1495	0.068	173	0.044
ADV_NI	18635	0.845	3265	0.834
ADV_TIME	7564	0.343	863	0.221
AR	2318	0.105	1175	0.300
CON	177769	8.056	32523	8.311
DEFAULT	48	0.002	32	0.008
DELM	217533	9.858	39062	9.982
DET	39783	1.803	6115	1.563
IF	2575	0.117	547	0.140
INT	111	0.005	2	0.001
MORP	2823	0.128	204	0.052
MQUA	250	0.011	111	0.028
MS	8	0.000	253	0.065
N_PL	135474	6.139	24945	6.375
N_SING	826571	37.459	140975	36.026
NN	2	0.000	0	0.000
NP	42	0.002	10	0.003
OH	271	0.012	12	0.003
OHH	15	0.001	5	0.001
P	270894	12.276	48964	12.513
PP	755	0.034	125	0.032
PRO	53792	2.438	8067	2.062
PS	296	0.013	37	0.009
QUA	12745	0.578	2673	0.683
SPEC	3281	0.149	528	0.135
V_AUX	13484	0.611	2386	0.610
V_IMP	1044	0.047	113	0.029
V_PA	73591	3.335	7003	1.790
V_PRE	35286	1.599	7209	1.842
V_PRS	41226	1.868	10512	2.686
V_SUB	28592	1.296	5228	1.336

2.2 Corpus and Tag Set Statistics

After reducing the number of the tags in the tag set to 40, the corpus was tagged by these new tags. Then five different sets of training and test sets were created by randomly dividing the corpus into two parts with a 85% to 15% ratio. Each experiment

was conducted 5 times, once on each pair of training-testing sets. Then the results were averaged and used for drawing conclusions.

Table 1 shows the most and the least frequent tag names and their frequencies and percentages both in the training and the test sets. As appears in the table, the tag “N_SING” (Noun-Singular) is the most common tag in both training and test sets. This tag is occurred 826571 times in the training set and 140975 times in the test set. On the other hand, the “NN” tag has appeared only twice in the training set and never in the test set.

Interestingly only the tags “ADJ_SIM”, “CON”, “DELM”, “N_PL”, “N_SING”, “P” has appeared more that 6% in both training and test sets. Generally, the percentages of the occurrence are very similar in both test and training sets. This means that the test sets and training sets have similar characteristics and the result obtained through experiments are reliable indicators for behavior of the algorithms. Table 2 shows the minimum, maximum, average and total number of POS tags in the corpus.

Table 2: The Minimum, Maximum, Average and Total Number of Tags

	Frequency in Training Set	Frequency in Test Set
Max	826571	140975
Min	2	0
Average	55165.7	3524.4
Sum	2206627	391310

2.3 Number of Different Tags for Each Word

Another piece of useful information about the corpus is the ambiguity of the words. That is how many words in the corpus have been assigned more than one POS tags. For instance, the word “ASEMAN” which means “the sky” in Persian is always tagged with N_SING in the whole corpus; but a word like “BALA” which means “high or above” has been tagged by several tags (ADJ_SIM, ADV, ADV_NI, N_SING, P, and PRO). Table 3 reports the number of words with various numbers of tags. It seems that majority of the words in the corpus have only a single POS tag (91%) and only a small fraction of the words have been assigned more than 3 different POS tags.

Table 3: Number of Different Tags

Number of words	Number of different tags
62568	1
4838	2
493	3
118	4
49	5
14	6
7	7
2	8
2	9
1	10

2.4 Classifying the Rare Words

Another interesting issue is the treatment of less frequent tags. Table 1 shows that there are many tags that occur lesser than others in the corpus. It seems reasonable to categorize them into a new group called “ETC”. The tags which are picked for “ETC” group are the ones whose number of occurrences is below 5000 times in the corpus. With this new tag the distribution of the words changes to the chart depicted in Figure 1.

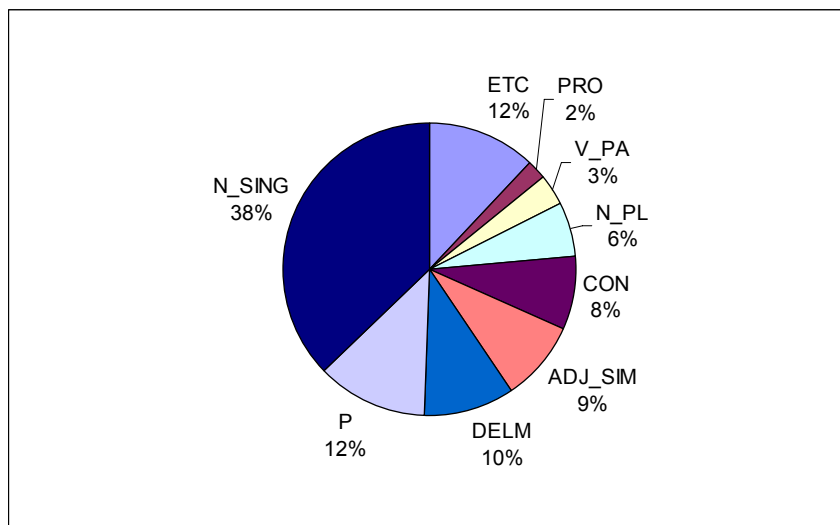


Figure 1: Tag Distribution for Collection (The tags which are picked for “ETC” group are the ones whose number of occurrences is below 5000 times in the corpus)

3. Maximum Likelihood Estimation

As a bench mark of POS tagging accuracy, we chose Maximum Likelihood Estimation (MLE) approach for its simplicity and ease of implementation. In this section we present the maximum likelihood estimation for the part of speech tagging of the corpus. In this approach, for every word in the training set we calculated the tag which is assigned to the word more than the other tags. For this purpose, we calculated the maximum likelihood probabilities for each tag assigned to any word in the training set. Then we pick the tag with the greater maximum likelihood probability for each word and make it the only tag assignable to that word. We call this tag the *designated* tag for that word. In order to evaluate this method we analyze the words in the test set and assign the *designated* tags to the words in the test set.

We ran the MLE Estimation on five different test collections as described in section 2. Table 4 shows the result of MLE Estimation for known words (the words that has seen before in the training set), Table 5 shows the result of MLE Estimation for unseen words (the word that has not seen before in the training set) and table 6 shows the overall result of MLE Estimation for both seen and unseen words. In table 5, for each unseen word we assigned the “DEFAULT” tag. In other words we considered the “DEFAULT” tag as the *designated* tag in these runs.

Table 4: MLE Estimation Results for Seen Words

Run	Percent	Tokens	Correct	Accuracy (%)
1	98.07	394290	380514	96.51
2	98.16	345913	334474	96.69
3	98.04	397849	384069	96.54
4	98.02	410970	396727	96.53
5	98.07	403460	393567	97.55
Average	98.072	390496.4	377870.2	96.76

Table 5: MLE Estimation Results with "DEFAULT" for Unseen Words

Run	Percent	Tokens	Correct	Accuracy (%)
1	1.93	7760	10	0.13
2	1.84	6689	11	0.16
3	1.96	7956	15	0.19
4	1.98	8283	8	0.10
5	1.93	7945	19	0.24
Average	1.928	7726.6	12.6	0.16

Table 6: The Overall Results for MLE Estimation

Run	Tokens	Correct	Accuracy (%)
1	402050	380524	94.65
2	362658	334485	92.23
3	405805	384084	94.65
4	419253	396735	94.63
5	411405	393586	95.67
Average	400234.2	377882.8	94.37

In the above experiments, we assign the "DEFAULT" tag to the unknown words. From all the "DEFAULT" tags assigned, the assignment was correct only at most 19 times. For the other occasions we were wrong. Therefore this approach to assignment of tags to unknown words reduced the overall accuracy to about 94.37%.

In a different approach, in order to achieve a better accuracy we assign the most frequent tag, "N_SING", to the Unknown words. Table 7 shows the result of MLE Estimation with "N_SING" tag for Unknown words. Table 8 shows that this approach improves the overall accuracy to 95.43%.

Table 7: MLE Estimation Results with "N_SING" for Unknown Words

Run	Percent	Tokens	Correct	Accuracy (%)
1	1.93	7760	4293	55.32
2	1.84	6689	3991	59.67
3	1.96	7956	4351	54.69
4	1.98	8283	4869	58.78
5	1.93	7945	3930	49.47
Average	1.928	7726.6	4286.8	55.59

Table 8: Overall Results for MLE Estimation

Run	Tokens	Correct	Accuracy (%)
1	402050	384807	95.71
2	362658	338465	93.33
3	405805	388420	95.72
4	419253	401596	95.79
5	411405	397497	96.62
Average	400234.2	382157	95.43

4. A Comparison of the Different Approaches

Table 9 compares the overall result obtained in our experiments with results reported in the literature for other languages such as German, English and Spanish.

Table 12: Overall Comparison of Results for MLE-Default, MLE-N-SING and MBT

Accuracy/Approach	Known Words	Unknown Words	Overall
MLE-DEFAULT	96.76%	0.16%	94.37%
MLE-N-SING	95.43%	55.59%	95.43%
ENGLISH	97.0%	85.5%	96.7%
GERMAN	97.7%	89.0%	96.7%
SPANISH	96.5%	79.8%	94.15%

As it is seen in the table the MLE approach with assigning DEFAULT tags to unknown words produces the least accurate results. MLE approach with assigning "N-SING" tag to unknown word which is the most frequent tag also, produces better results. By comparing the MLE results with the result obtained in other languages it seems MLE is a good benchmark and start for automatic part of speech tagging efforts for Persian language.

5. Conclusion

This paper describes the process of creation of a feasible corpus for Part of Speech Tagging experiments for Persian language from a finer grained manually created POS tagged corpus. This paper also described experiments conducted with Maximum Likelihood approaches for POS tagging. The taggers were trained on 85% of the corpus and were tested on the remaining 15%. As a result, MLE produced an overall accuracy of tagging around 95%. By comparing result obtained for MLE with the results obtained for other languages such as English, German and Spanish, it seems the MLE approach provides a very high benchmark for Persian POS tagging.

In future we would like to continue these experiments with other types of Part of Speech tagging models such as Hidden Markov models, Memory based models etc. We like to experiment also with the size of training and test collection and investigate the effect of the size of the training on the effectiveness of the tagger

6. Acknowledgements

We would like to thank Dr. Faili for his help in gathering and preparing the tagged corpus. We also highly appreciate Dr. BijanKhan for his valuable work in tagging the Persian language.

7. References

- Daelemans, W., J. Zavrel, P. Berck & S. Gillis. (1996) 'MBT: A Memory-Based Part of Speech Tagger Generator'. In *Proceedings of the Fourth Workshop on Very Large Corpora*, Copenhagen: ACL SIGDAT, pp. 14-27.
- Daelemans, W. & J. Zavrel (1999) 'Recent Advances in Memory-Based Part of Speech Tagging'. in: *Actas del VI Simposio Internacional de Comunicacion Social*, Santiago de Cuba, pp. 590-597.
- Daelemans, W., Van den Bosch, A., Weijters (1997) 'IGTree: Using Trees for Compression and Classification in Lazy Learning Algorithms', *Artificial Intelligence Review*, vol. 11, 1997, pp. 407-423.
- Zavrel, J. & W. Daelemans (1997) 'Memory-based learning: Using similarity for smoothing. In *Proceedings of the 35th Annual Meeting of the ACL, ACL97*, Madrid, Spain, 436-443.
- Ratnaparkhi, A. (1996) 'A maximum entropy part-of-speech tagger'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania. Pennsylvania, USA, pp. 133-142.
- BijanKhan, M., (2004) 'The role of the corpus in writing a grammar: an introduction to a software', *Iranian Journal of Linguistics*, Vol. 19, No. 2.
- Brants, (2000), T.:TnT – a Statistical Part-of-Speech Tagger. *Proceedings of the Sixth Conference On Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.
- Mihalcea, R. (2003) 'Performance Analysis of a Part of Speech Tagging Task'. *Proceedings of Computational Linguistics and Intelligent Text Processing*, Centro de Investigaci3n en Computaci3n IPN, M3xico.
- Carrasco, R. M. & Gelbukh, A. (2003) 'Evaluation of TnT Tagger for Spanish', *Proceedings of the Fourth Mexican International Conference on Computer Science (ENC'03)*. Tlaxcala, Mexico.