A POS Tagger analysed in collaboration environments and literary texts

Dumitru-Clementin Cercel*
Ştefan Trăuşan-Matu**

*University *"Politehnica"* of Bucharest,
Department of Computer Science and Engineering
Splaiul Independenţei Bd., No. 313, Bucharest, Romania
clementin.cercel@gmail.com
**Romanian Academy Research Institute for Artificial Intelligence
13 Septembrie Street, No. 13, Bucharest, Romania
stefan.trausan@cs.pub.ro

Abstract. Part-of-speech (POS) tagging is often used in other modules of natural language processing and therefore the results of this process should be as precise as possible. Many different types of taggers have been developed to improve the accuracy of the results in the field of literature or newspapers. Nowadays when the internet is widespread, the environments for online collaboration as chats, forums, blogs, wikis have become important means of communication. The purpose of this research is to analyse the results of tagging the words obtained from the labelling of the words from the online collaboration environments and literary texts with the corresponding parts of speech. In the case of POS tagging, the ambiguities arise due to the fact that a word may have multiple morphological values depending on context.

1 Introduction

Part-of-speech (POS) tagging is the process of grammatical labelling of each word inside a text with its appropriate part of speech. Labelling may also contain extra information related to the morphological characteristics of the respective language like number, gender, person, tense and aspect of the verb.

Many different types of taggers have been developed to improve the accuracy of the results in the field of literature or newspapers. Nowadays when the internet is widespread, the environments for online collaboration as chats, forums, blogs, wikis have become important means of communication. The purpose of the research presented in this paper is performing a comparative analysis of POS tagging in collaborative corpora (specifically for chats, Wikipedia and Twitter as an example of microbloggins) and literature (the texts from Brown corpus). In this aim we implemented a trigram HMM tagger according to Jurafsky and Martin (2000) and Brants (2000).

The rest of this paper is structured as follows. In section 2 we briefly review: the state of the art approaches to POS tagging, the Markov Hidden Model, the implementation of the trigram HMM tagger and we will analyse the factors that influence the performance of the tagger, taking into account the differences between literary texts, wikis, microbloggings and chat copora. In Section 3 we analyse the result of our work. In section 4 we present the conclusions and identify the opportunities for a follow-up research.

2 POS tagging

Over time, the POS tagging has been an important topic of research and new methods have developed in order to improve the precision of the results. In Jurafsky and Martin (2000) is made a classification of the methods used in resolving POS-tagging in three categories: stochastic methods (probabilistic methods), rule-based methods and a combination of both (hybrid taggers). Usually, the building of a POS tagger follows certain phases (Voutilainen, 2005): tokenization, lexicon look-up and ambiguity resolution.

In one study released by the Association for Computational Linguistics, as regards the POS tagging problem is performed an analysis of the results of several implementations of automatically learning methods, by using for training and testing the Wall Street Journal corpus (Marcus et al., 1993). For the TnT Tagger proposed by Brants (2000) and based on HMM a 96.46% accuracy was obtained. The SVMTool Tagger introduced by Giménez and Márquez (2004) and based on Support Vector Machine obtained a 97.16% accuracy. The accuracy for Stanford Tagger 2.0 (Manning, 2011) using the maximum entropy model was 97.32%. The LTAG-spinal Tagger described by Shen et al. (2007), using an algorithm based on bidirectional perceptron learning has a 97.33% accuracy. The Morče / COMPOST Tagger presented in Spoustová et al. (2009), has a 97.44% accuracy, using a method based on the bidirectional perceptron learning. A result with a 97.50% accuracy was obtained by the SCCN Tagger proposed by Sogaard (2011), based on a model that uses condensed nearest neighbour. In terms of unknown words which are not in the training corpus, the highest accuracy of 91.29% was obtained by using the MElt Tagger described by (Denis and Sagot, 2009), and the most unsatisfactory result of 85.86% was obtained by the TnT Tagger.

The previous performances analysis were performed only (as we know) on texts which may be considered as being in the class of literature. As we mentioned, in this paper we will consider also the case of texts found in collaboration environments.

The Internet has been a major factor in developing the informal language through communication environments, which have brought changes to the literary language. The main thing that distinguishes tweets from other texts is that the Twitter message is up to 140 characters. Chats are based on sending multiple short messages between participants and therefore a phrase may not be entirely written in one instance message.

There are many similarities between the microblogging texts and the chat texts . One tendency is to use a lot of abbreviations and shortcuts (such as "BRB" - "be right back") or to drop the apostrophes or the full stops after abbreviations (e.g. "UE" instead of "U.E".). The participants also omit to use capital letters at the beginning of a sentence or name and they rarely use diacritics.

A common custom is to use emoticons in place of words in order to express emotions and feelings. Unlike the literary language, one of the consequence of a rapid message is to mistype words like reversing letters, missing letters, joining words and so on (e.g., "ill" instead of "I will"). Frequently appear misspellings as well as the use of emphasis through character repetition (e.g. "biiigggg" instead of "big"). Another particularity of the microbloggings and the chats consists in the ungrammatical inputs which are indeed more than in the literary works. The chatters often tend not to respect the order of words in a sentence or the punctuation marks although the order of words and these separators have an important role in understanding the meaning.

Wikipedia is an editable website which allows its users to add, change, or overwrite its content. In sharp contrast to the collective literary text, this characteristic of Wikipedia of being a collaborative creation can involve grammatical alterations of the text and thus, its meaning is successively modified. To prevent the acts of vandalism, the access to editing articles can be protected, even locked so that certain persons are able to make changes.

2.1 A trigram model for POS tagging

An HMM according to Manning and Schutze (1999) is specified by the parameters (N, K, A, B, π). For the POS tagging, the labels are represented as states in finite automaton HMM. Thus N is the number of labels used by the model, size K is the number of distinct words from the vocabulary of the model, π_i is the probability that the first word of the sequence is labeled with t_i, a_{ij} is the probability that t_j label is preceded by t_i label, b_{jk} is the probability that the w_j word has the t_j tag. In a HMM, the sequence of states generated by the process is not known (it is hidden) because is known only the sequence of observations (the words in our case).

Given a sequence of words $W = w_1, w_2, \ldots, w_n$, POS tagging process implies the determination of the most probably sequence of states that the model is going through, i.e. the most probably sequence of $\hat{T}$ labels that maximizes $P(T|W)$.It results that the trigram model for POS tagging as shown in Jurafsky and Martin (2000) is:

$$\hat{T} = argmax_{t_1 \ldots t_n} [\prod_{i=1}^{n} P(t_i|t_{i-1}, t_{i-2}) P(w_i|t_i)] P(t_{n+1}|t_n) \quad (1)$$

Due to insufficient data from the training corpus, it may happen that a sequence of trigrams appears in the test corpus, but not in the training corpus and thus we come to wrongly establish the probability of the sequence set as zero. Even when the sequence of labels would appear too few times in the training corpus, the probability calculated for the respective sequence would not be an exact estimation. In our implementation we used a solution proposed by Thede and Harper (2011) which gives weights to the sequences of trigram, bigram and unigram.

The probability $P(w_i|t_i)$ is obtained by performing an analysis of the suffixes using a letter-based n-gram model. Suffixes can provide a good indication of the associated part of speech of a word. The suffixes of the words with a frequency less than or equal to the threshold frequency are used to build a data structure namely suffix tree. Thus, the algorithm builds a suffix tree that contains the suffixes for the words that begin with lowercase, another tree for the words starting with a capital letter and also a tree for the words that start with digits. Applying an approach due to Samuelsson and Reichl (1999) and Brants (2000), we calculate by interpolation the probability of a certain label, where are known the last i letters from a word of L letters.

The most likely sequence of tags given the observed sequence of words can be solved by a brute force search evaluating the probability of each possible sequence of tags for the sequence of input words, but it requires a large execution time. In order to obtain better results we have used in our implementation the (Viterbi, 1967) algorithm which is based on dynaming programming.

3 Evaluation

For the evaluation of the tagger we used several corpora. We used the Brown corpus which is one of the most known corpus for the English language described in Francis and Kucera (1979). The NPS Chat Corpus v. 1.0 (Forsyth and Martell, 2007) was created in 2006 from various online chat rooms and contains records from a short period on a particular day and consists of 10,567 utterances. WikiCorpus (Reese et al., 2010) represents a lexical semantic resource available for the NLP community. The English portion of the corpus contains large portions of Wikipedia pages available in 2006 (around 600 million words). Also, we used Twitter corpus labelled with parts of speech by Gimpel et al. (1993).

Sections train set	Sections test set	The number of words from the training set	Unknown words	The number of words from the test set	Precision known words	Precision unknow words
ca-cg	ch-cr	608001	26245	553191	95.82	78.52
ca-cj	ck-cr	857752	12719	303440	96.13	78.68
ca-cl	cm-cr	985663	6398	175529	96.40	79.58
ca-cn	cp-cr	1069475	2956	91717	96.65	79.63
ca-cp	cr	1139497	934	21695	96.28	79.33

TAB. 1 – *The tagging precision for the Brown corpus.*

The used corpus	The number of words from the training set	Unknown words	The number of words from the test set	Precision known words	Precision unknow words
NPS Chat	12694	7540	32314	92.77	62.99
	27185	3270	17823	93.33	64.95
	32646	2213	12362	92.94	67.78
	39420	736	5588	94.22	63.45
	42226	379	2782	95.46	60.15
WikiCorpus	1663043	579254	6277115	97.39	92.52
	655532	153776	1358854	97.08	90.66
	413137	2014386	257750	96.94	90.07
	354488	466773	3594006	97.00	89.65
	153937	575257	3594006	96.69	87.46
Twitter	14619	1901	7152	92.00	60.44
	21771	1207	4823	91.48	64.70
	23682	723	2912	91.18	68.46

TAB. 2 – *The tagging precision for NPS Chat, WikiCorpus, Twitter.*

The tagger was trained for each corpus and tested on itself. In the tables 1 and 2 are presented the results obtained for the tagger when the suffixes number and the threshold frequency of the words from the training set, used in the building of the suffixes trees are 2 and 4 respectively, these values being experimentally determined. The results for the Brown corpus have a good enough precision both for the unknown words (79.69) and for the words found in the training set (96.65%). The testing of the tagger on the chat corpus had an precision of 92-95% for the known words and an precision of 60-67% for the unknown words. Using the WikiCorpus the precision was over 97% for the known words and between 89-92% for the unknown words.

4 Conclusions

Building a POS tagger with a great accuracy for certain online collaboration environments is more difficult because there are many syntactic and semantic differences in comparison with the texts of literature or newspapers. In this paper, we have identified various differences that we consider the most important POS tagging results for specific online collaboration environments such as wikipedia, chats and microblogging.

In a future research we will perform the analysis of the results of the POS tagger trained on old corpora and tested on a recent data set in order to identify the grammatical differences. Also, an interesting direction of research using a POS tagger is the analysis of the differences of grammatical labelling of the words in the text corpora between a native language and its dialects.

References

Brants, T. (2000). Tnt - a statistical part-of-sppech tagger. In *The 6th Applied NLP Conference*, Seatle, WA.

Denis, P. and B. Sagot (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *PACLIC*.

Forsyth, E. and C. Martell (2007). Lexical and discourse analysis of online chat dialog. In *The First IEEE International Conference on Semantic Computing (ICSC 2007)*, pp. 19–26.

Francis, W. and H. Kucera (1979). Brown corpus manual. Technical report, Brown University, Department of Linguistics, Providence, Rhode Island.

Giménez, J. and L. Márquez (2004). Svmtool: A general pos tagger generator based on support vector machines. In *The 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Gimpel, K., N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, and D. Yogatama (1993). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Jurafsky, D. and J. Martin (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition.* Upper Saddle River: Prentice Hall.

Manning, C. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *The 12th International Conference on Intelligent Text Processing and Computational Linguistics.*, pp. 171–189.

Manning, C. and H. Schutze (1999). *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press.

Marcus, M., B. Santorini, and M. Marcinkiewicz (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics 19(2).*

Reese, S., G. Boleda, M. Cuadros, L. Padró, and G. Rigau (2010). Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. In *The 7th Language Resources and Evaluation Conference (LREC'10)*, La Valleta, Malta.

Samuelsson, C. and W. Reichl (1999). A class-based language model for large vocabulary speech recognition extracted from part-of-speech statistics. In *IEEE ICASSP-99*, pp. 537–540.

Shen, L., G. Satta, and A. Joshi (2007). Guided learning for bidirectional sequence classification. In *The 45th Annual Meeting of the Association of Computational Linguistics (ACL).*

Sogaard, A. (2011). Semi-supervised condensed nearest neighbor for part-of-speech tagging. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT).*

Spoustová, D., J. Hajic, J. Raab, and M. Spousta (2009). Semi-supervised training for the averaged perceptron pos tagger. In *The 12 EACL.*

Thede, S. and M. Harper (2011). A second-order hidden markov model for part-of-speech tagging. In *The 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99)*, pp. 175–182.

Viterbi, A. J. (1967). Error bounds for convolutional codes and asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory 13*, 260–269.

Voutilainen, A. (2005). *The Oxford handbook of computational linguistics*, Chapter ch. 11: Part-of-Speech Tagging. Oxford University Press. pp. 219-232.

Résumé

L'étiquetage grammatical est souvent un composant des autres modules du traitement du langage naturel dont les résultats doivent etre aussi précis que possible. De nombreux types d'étiqueteurs grammaticaux ont été développés pour améliorer la précision des résultats dans le domaine de la littérature ou de la presse. De nos jours, quand l'Internet est tres répandu, les environnements de collaboration en ligne comme les clavardages, les forums, les blogs, les wikis sont devenus des moyens importants de communication. Le but de cette recherche est d'analyser les résultats obtenus dans l'étiquetage des parties du discours pour les corpora d'environnements de collaboration en ligne et le corpus de la littérature. Dans le cas de 'étiquetage grammatical, les ambiguités surviennent lorsqu'un mot peut avoir plusieurs valeurs morphologiques en fonction du contexte.