PART OF SPEECH TAGGING IN ROMANIAN TEXTS

Claudia Cîrcioroabă[1], Mihai Stancu[2], Daniel I. Morariu[3], Daniel Volovici[4]

[134]"Lucian Blaga" University of Sibiu, Engineering Faculty, Computer Science and Electrical and Electronics Engineering Department, România

[2]"Lucian Blaga" University of Sibiu, The Faculty of Economics, PhD Student, România

Abstract

Identifying Parts of Speech (PoS) represents the process by which grammar tags containing their corresponding PoS are attached automatic to every word within a sentence. Since no word acts as just one single PoS—their syntactic value depending on the context they are used in—identifying parts of speech is not a trivial matter. In this paper we have taken into account two tagging methods, based on Naïve Bayes' classifier probabilities and the occurring context of the word for which the PoS must be identified. We have called these methods Backward Naïve Bayes and Forward Naïve Bayes. For Romanian language, we have taken into account seven different PoS as: noun, verb, adjective, adverb, article, preposition plus the „and others" category. From conducted experiments, we have observed that identifying the PoS for a word based on the PoS for the previous word produces better results in all respects. We have studied each PoS separately and have concluded that there also are more easily identifiable PoS in Romanian as well: article, preposition, noun, verb; meanwhile the adjective and adverb are more problematic in identifying the PoS.

Keywords: Part of speech tagging, Document Classification, Naïve Bayes

1 INTRODUCTION

In an effort to build machines with enhanced performance, which would ease our workload, the current trend is to make man-machine communication as easy and as natural for human as possible. Ordinary language makes for the „handiest" method of human-machine interaction, having more benefits for humans, especially for people who are not familiar with computers. Some of these methods rely on vocabulary, immediately available to human and which does not suppose prior training [12]. Processing a natural language by means of a computer includes both understanding and generating messages (feedbacks) in that language. Since understanding automatically involves computer identifying of the syntactic structure in messages, it is to be concluded that processing ordinary language cannot be done without strong mechanisms of representation and processing [10].

Database systems and information recognition systems are currently working with different types of data, thus resulting some problems from database systems, usually

when information contained in them is not present in information recognition systems. There are also some issues in recognizing shared information, which does not usually appear in traditional database systems [8].

2 EXPERIMENTAL FRAMEWORK

Knowledge discovery in data collections means discovering various patterns—potentially useful and easily understandable—that can be found in data, such as: databases, collections of documents, collections of images, Web, etc. Within data collections, data extraction and analysis can be performed by applying algorithms in order to identify patterns aimed at obtaining new knowledge.

2.1 Corpus

In computational linguistics domain, a corpus is a structured collection of texts in electronic format, which are represented in a certain formalism so that they can be easily "read" by applications specializing in extracting information from these collections.

In this paper we used a corpus in Romanian taken from the web address [4], called „train_corpus_augusto_xml”.

The corpus contains a collection of sentences taken from Romanian literature within each sentence words are separated and annotated.

2.2 Part-of-speech

The lemma of a word is its basic dictionary form. Although the text is in Romanian, the part of speech for every word is specified in English. Within the *postag* attribute the first letter shows the syntactic value of the word, while the other ones contain information relating to gender, person, number etc. During data selection, it was noticed that in order to identify PoS it sufficed extracting only the first letter in the *postag* attribute, the remaining information being irrelevant. Within the entire data set we have identified ten main PoS, two parts of sentence (as defined by Romanian grammar: attribute and complement) and two PoS speech falling into the category „others” - negation and digits.

In this paper, the fourteen distinct values identified within the *postag* attribute have been reduced to seven. Although there are ten main PoS in Romanian grammar, we have hereafter kept only seven (six main PoS plus the category „others”), those we have deemed more relevant and repeated themselves sufficiently enough in the data set. Those PoS closely-related in meaning were united under one category, whereas those occurring occasionally were added to the category „others”.

Part-of-speech tagging is the process of grammatically labeling each word within a simple sentence, complex sentence or paragraph with the appropriate part of speech. Tags are generally PoS (e.g.: nouns, verbs, adjectives, prepositions, interjections), but may also contain additional information relating to the morphological characteristics of the language, such as number, gender, person, time or verb aspect.

3 LEARNING ALGORITHM

Starting from classifier Naïve Bayes' idea, according to whom unlabeled data in the test set is classified by means of estimates from labeled training data, the purpose of this paper is to identify PoS in Romanian texts using classifier Naïve Bayes' ideas [7].

In this paper we have decided to use just the syntactic value of a word and the context for its occurrence, i.e. the PoS of the word preceding and following to the current word. The reason we did not use the word itself being the difficulty of finding a set of words in the same context, and the little relevance this bears.

3.1 Basic Global Probability

By using *Basic Global Probability* classification our goal is to count for each lemma all distinct PoS that occur in the training set. Thus, by using the training set, extracted from the original data set we could count all syntactic values that occur for every lemma within the training data set.

For global probabilities, we have used the classical probability formula:

$$P = \frac{The\_number\_of\_favorable\_cases}{The\_number\_of\_all\_cases} \tag{3.1}$$

Lemma occurrences have been counted and a frequency vector for each individual lemma was created. In order to count lemma it was necessary to establish how many times every word occurs as: noun, adverb, adjective, preposition, verb, article and „others".

Results obtained from the global count of syntactic values for the PoS taking in consideration are:

```
"PartOfSpeechStatistics": {
    "adv": 3236,
    "subst": 5527,
    "adj": 1941,
    "prepozitie": 3385,
    "verb": 4044,
    "articol": 3651,
    "punct": 1200,
    "altele": 5
  }
```

Figure 3.1 – Overall Parts of Speech (PoS)in Training Set

3.2 Naïve Bayes Training

During Naïve Bayes training stage, the information from the training set have been prepared, so that it would be easier to calculate probabilities during testing stage. The Naïve Bayes training process for identifying PoS entails counting lemma, producing statistical data for each distinct part of speech and calculating probabilities based on Bayes' formula (conditional probabilities) for Basic Global Probability, Backward Naïve Bayes and Forward Naïve Bayes.

A simple Bayesian classifier departs from the „naive" assumption that attributes for a given class are independent of each other. This assumption is called class conditional

independence. Within this classifier we depart from the assumption mentioned previously in order to simplify calculations [5].

3.3 Backward Naïve Bayes

According to article [6], Backward Naïve Bayes method, based on Bayes' theorem identifies the part of speech for a current word by taking into account the global probability for their word resulting from the training set and the syntactical value of the previous word.

Statistical data for PoS shows that the most frequent part of speech is the *noun*. Therefore, we considered the first word in each sentence as having the syntactical value of noun. In processing sentences, we considered dots as being sentence delimiters, clearly marking the beginning of a new sentence.

To identify PoS by means of Backward Naïve Bayes, we applied equation (3.2) to calculate conditional probabilities, wherein x is the word and *POS* is one of the PoS assigned to the current word at the moment global probabilities were calculated on the training set.

$$P(x = POS / predecessor) = \frac{P(predecessor / x = POS) * P(x = POS)}{P(predecessor)} \quad (3.2)$$

We shall take into account the following example and intend to identify what part of speech is most likely to follow the word "*frumos"*, with reference to global probability and to the part of speech for the word following it.

El își petrece zilele frumoase de vară la bunici.

We notice that the word „*frumos*" is preceded by a noun. The word *frunos* is lemma part for the word *frumoase.* According to training data set statistics the word „frumos" can be either an adjective or a noun. Within this context we must identify what part of speech can the word „frumos" be, when a noun precedes it.

Results for counting lemma „*frumos"* and all PoS likely to precede it are shown in Fig. 3.2. For an easier interpretation of results, we have chosen to present data as follows: the first word is the lemma; then the part of speech preceding the lemma; and the number of occurrences for the syntactic value of that lemma, according to the part of speech preceding it.

```
"frumos": {
                "preposition":     {   "adj": 4,     "subst": 1   },
                "noun":  (*)  {   "adj": 3,     "subst": 1   },
                "adjective":       {   "adj": 2                 },
                "adverb":          {   "adj": 4                 },
                "verb":        {   "adj": 1                 },
                "others":          {   "adj": 2                 }
        }
```

Figure 3.2 Example from the Training Set for the Word *"frumos"*

From the above statistical data, what is of interest to us is when a noun precedes „*frumos"*—occurrence identified by (*). Based on the equation (3.2) we get:

$$P(frumos = adjective \mid pred = noun) = \frac{P(pred = noun \mid frumos = adj) * P(frumos = adj)}{P(pred = noun)}$$
$$= \frac{3/4 * 16/18}{3/4} = 88.89\%$$
$$P(frumos = noun \mid pred = noun) = \frac{P(pred = noun \mid frumos = noun) * P(frumos = noun)}{P(pred = noun)} \quad (3.3)$$
$$= \frac{1/4 * 2/18}{1/4} = 11.11\%$$

It is therefore more likely that word ”frumos” be an adjective than a noun when preceded by a noun. The algorithm will consequently choose „adjective”.

3.4 Forward Naïve Bayes

Forward Naïve Bayes method is like Backward Naïve Bayes, only that the identification being done in this case based on the global probability resulting from the training set and the syntactic value of the word that follow the current word.

To identify the part of speech for a word using Forward Naïve Bayes, we have applied the following formula to calculate conditional probabilities:

$$P(x = POS \mid successor) = \frac{P(successor / x = POS) * P(x = POS)}{P(successor)} \quad (3.4)$$

x is the word and POS one of the PoS assigned to the current word at the moment global probabilities were calculated using the classical formula for calculating probabilities, POS can be {noun, verb, adverb, adjective, article, preposition, others} and successor is the part of speech that follow after the current word, successor can be {noun, verb, adverb, adjective, article, preposition, others}.

4 EVALUATING CLASSIFIER PERFORMANCE

To evaluate performance indicators for pattern identification/recognition/classification systems we used external measurements, as they exist in literature: precision, recall, and accuracy. These metrics can be more easily computed if a contingency matrix is used. Such is the case for Table 4.1, where information about current and predicted classification - made by a classification system -, is shown [1].

Table 4.1 Contingency Matrix

		True condition	
		Positive class	Negative class(s)
Predicted condition	Positive class	True positive (TP)	False positive (FP)
	Negative class(s)	False negative (FN)	True negative (TN)

Based on the indicators mentioned above, we present some external measurements that were used in the application for evaluating the quality of PoS identification.

Accuracy is the ratio between PoS in category *i* labeled correctly and overall PoS in category *i*, where *i* can be {noun, verb, adverb, adjective, article, preposition}.

Precision (also known as *positive predictive value*) is the ratio of PoS in category *i* labeled correctly, from all PoS labeled by classifier as belonging to category *i.*

$$\Pr ecision = \frac{TP}{TP + FP} \tag{4.1}$$

Recall (also known as *sensitivity*) is the ratio of PoS in category *i* labeled correctly, from all PoS that can needed to be in category *i*.

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{4.2}$$

5 EXPERIMENTAL RESULTS

5.1 Naïve Bayes Testing

Algorithm evaluation has been performed on a testing set, different from the training set. Naïve Bayes identification algorithm is applied on the training set, while performance evaluation for Naïve Bayes classifier was undergone on the test set, which is seen as unknown data (data encountered for the first time) by the algorithm.

Backward Naïve Bayes method identifies, for test set the part of speech for every word by considering the following information: probability for the occurrence of certain PoS relating to the current word and conditional probability for PoS occurrence in a certain order - probability that the current word may or may be not preceded by the PoS that preceded it within the training set [5].

Forward Naïve Bayes method identifies the part of speech for every word by taking into account the following information: probability for the occurrence of certain PoS relating to the current word and conditional probability for PoS occurrence in a certain order – probability that the current word may or may be not followed by the PoS that followed it within the training set [6].

5.2 Accuracy

Accuracy results obtained are shown in Table 5.1. It is noted that the best accuracy is achieved with Backward Naïve Bayes, meaning there is a connection to PoS order within a sentence, in the sense that part of speech for the current word depends on the part of speech for the previous word.

Table 5.1 Accuracy values by using all classifiers

Accuracy	
Global Probability (GP)	90.7439%
Backward Naive Bayes(BNB)	90.8555%
Forward Naive Bayes(FNB)	90.8454%

5.3 Precision and recall

In Table 5.2 and Fig. 5.3 results values for precision and recall are shown. These values and results have been achieved on the test set to which we have applied Global Probability, Backward Naïve Bayes and Forward Naïve Bayes. Results for both metrics are displayed for each part of speech separately, while their average is to be found at the end.

By analyzing results shown in Table 5.2 we notice that precision achieved with Backward Naïve Bayes and Forward Naïve Bayes is generally better than its Global Probability counterpart: this happens in four out of seven instances.

Table 5.2 Values Achieved for Precision, Recall an F-measure measures

	Precision			**Recall**			**F-measure**		
POS	GP	BNB	FNB	GP	BNB	FNB	GP	BNB	FNB
noun	78.95%	79.56%	79.65%	90.44%	89.86%	89.83%	84.30%	84.40%	84.43%
verb	99.68%	99.57%	99.62%	67.11%	67.44%	67.40%	80.22%	80.42%	80.40%
adjective	87.28%	88.53%	88.51%	32.19%	32.54%	31.59%	47.04%	47.59%	46.56%
adverb	95.76%	94.65%	95.55%	95.76%	54.45%	54.69%	95.76%	69.13%	69.57%
Article	98.63%	98.33%	98.31%	98.63%	98.33%	59.28%	98.63%	98.33%	73.96%
preposition	99.86%	99.65%	98.67%	99.86%	99.65%	61.42%	99.86%	99.65%	75.71%
others	100%	100%	100%	0.0021%	0.0022%	0.0022%	0.00%	0.00%	0.00%
average	94.31%	94.33%	94.33%	69.17%	71.91%	52.06%	72.26%	68.50%	61.52%

A more exact measurement would be **F-measure**, a harmonic mean between precision and recall which dismisses algorithms that turn just one of these metrics with good results and the other one with worse results. Formula used for F-measure is presented in (5.1) and Results are shown in Table 5.2.

$$F\_measure = \frac{2 * precision * recall}{precision + recall} \tag{5.1}$$

The best results for this measurement are on average those achieved with Backward Naïve Bayes and *preposition*, *article*, *noun* and *verb* been the most correctly identified PoS.

Fig. 5.3 is a graphic comparison of results achieved with Backward Naïve Bayes on each individual part of speech. We can notice from it that the best identification results are those for article and preposition, close to 100%. Noun and verb also turn out good results, more than 80%. Adjective, adverb and the category „others" generate most problems in PoS identification.

Fig. 5.4 summarizes results achieved with Forward Naïve Bayes classifier. Results for this classifier are much lower than those for Backward Naïve Bayes, which means that PoS identification relies more on the syntactic value of the preceding word than on the syntactic value of the following word. Noun and verb are PoS that turned out an 80% result, whereas other PoS approach 60% in terms of F-measure.

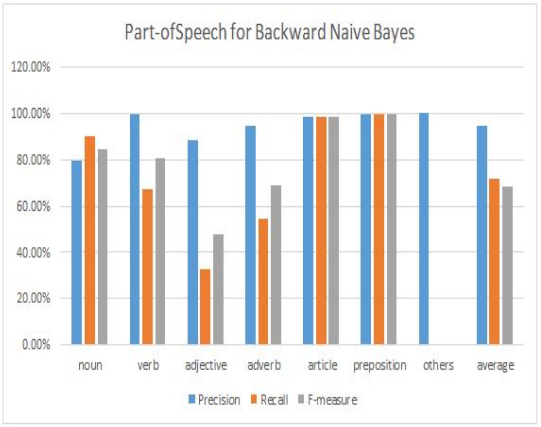

Figure 5.3 Results for Backward Naïve Bayes

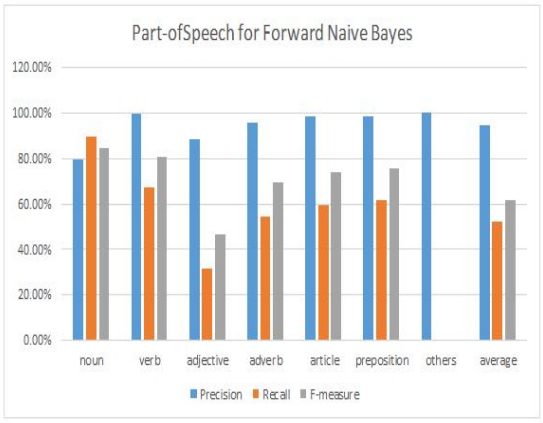

Figure 5.4 Results for Forward Naïve Bayes

5.4 Maximum Obtainable Limit for Naïve Bayes classifier

Backward Naïve Bayes and Forward Naïve Bayes methods' downside would be the fact that PoS identification for the *current* word relies on the PoS for the *preceding*, and respectively *following* word. Should part of speech for the *preceding* or *following* word be misidentified and used further, it may lead to error propagation and automatically, part of speech for the current word would be misidentified.

Table 5.5 Maximum Obtainable Limit for Classifier

	Accuracy		Precision		Recall		F-measure	
POS	Backwar dNB	Forward NB	Backward NB	Forward NB	Backward NB	Forward NB	Backward NB	Forward NB
noun	-	-	80.34%	81.96%	90.32%	90.45%	85.04%	86.00%
verb	-	-	99.62%	99.62%	68.51%	70.26%	81.19%	82.40%
adjective	-	-	89.81%	91.24%	34.25%	37.21%	49.59%	52.86%
adverb	-	-	95.31%	96.21%	55.64%	57.87%	70.26%	72.27%
article	-	-	98.34%	98.54%	60.84%	62.23%	75.17%	76.28%
prepositio	-	-	99.65%	98.67%	62.20%	64.40%	76.59%	77.93%
others	-	-	100%	100%	0.003%	0.002%	0.01%	0.00%
average	91.27%	91.89%	94.72%	95.18%	53.16%	54.67%	62.55%	63.97%

As a conclusion, by updating and correcting PoS values where needed, accuracy increased a little over 1%, precision increased by just under 1%, while recall registered the best increase, 2,5%.

6 CONCLUSIONS

The purpose of this paper is to analyze performance for part of speech identification algorithms, studied and implemented in [6], when are used in a different language. The first article deals with algorithms used to identify PoS in English texts, and only

four main PoS were considered. For the purpose of this paper we used documents in Romanian language and considered six main PoS.

As a conclusion, we notice that there are PoS more easily and less easily identifiable. Article and preposition are the most correctly identified PoS, the percentage being close to 100%. Noun and verb also turn out good results, more than 80%. Adjective, adverb and the „others" category generate most problems in problem of PoS tagging, their percentage being below 60% and even 0 in terms of the harmonic mean between precision and recall.

Results achieved with Forward Naïve Bayes classifier are lower than those achieved with Backward Naïve Bayes, which means that PoS tagging relies more on the syntactic value of the preceding word than on the syntactic value of the following word.

7 REFERENCES

[1] Agavriloaei Ioan, Modele şi Algoritmi Mining, PhD thesis, 2012,

[2] Dumitru-Clementin Cercel, *POS tagger bazat pe modelul HMM*, Rumanian journal of Human- Computer interaction, 2012

[3] Colhon M., *Procesarea Limbajului Natural*, 2012, https://www.google.ro/webhp?sourceid= chrome-instant&ion=1&espv=2&ie=UTF-8#q=Colhon+M.%2C+Procesarea +Limbajului+Natural , accessed in February 2016

[4] http://www.mcolhon.ro/patterns/index.html - accessed in February 2016

[5] Radu G. Cretulescu, Daniel I. Morariu, *Text Mining. Tehnici de clasificare si clustering al documentelor*, Published at Editura Albastra, Cluj Napoca, 2012, ISBN 978-973-650-289-7

[6] R. CRETULESCU, A. DAVID, D. MORARIU, L. VINŢAN - *Part of Speech Tagging with Naive Bayes Methods*, Proceedings of The 18-th International Conference on System Theory, Control and Computing, Sinaia (Romania), October 17 - 19, 2014

[7] Dan Jurafsky, James H. Martin, *Speech and Language Processing*, 2016, https://web.stanford.edu/~jurafsky/slp3/, accessed in February 2017

[8] Daniel I. Morariu, *Text Mining Methods based on Support Vector Machine*, MATRIX ROM Publishing house, Bucharest, ISBN 978-973-755-343-0, 168 pages, 2008.

[9] Robi Polikar, *Pattern recognition*, Wiley Encyclopedia of BioMedical Engineering, 2006

[10] *Data Mining From A to Z*, SAS Institute Inc., 2015, www.Sas.com

[11] Catalin Stoean, Ruxandra Stoean, *Support Vector Machines and Evolutionary Algorithms for Classification: Single or Together?*, Intelligent Systems Reference Library, Volume 69, Springer, 2014

[12] Dan Tufiş, *Promovarea limbii române în SI* – SC, www.racai.ro/media/Tufis-SISC2001.pdf, published in 2001