A BAYESIAN PART-OF-SPEECH AND CASE TAGGER FOR MODERN GREEK

Manolis Maragoudakis	UNIVERSITY OF PATRAS
Katia Kermanidis	UNIVERSITY OF PATRAS
Nikos Fakotakis	UNIVERSITY OF PATRAS

Περίληψη

Το παρόν άρθρο παρουσιάζει και αξιολογεί ένα πρωτοποριακό στοχαστικό μοντέλο πρόβλεψης και επισημείωσης μερών του λόγου για κείμενα της Νεοελληνικής γλώσσας. Η στατιστική του υπόσταση πηγάζει από τη θεωρία των πιθανοτικών δικτύων πεποίθησης Bayes. Η μεθοδολογία αυτή επιτρέπει την αξιοποίηση της γειτνίασης μιας λέξης όχι μόνο με βάση τις προγενέστερες λέξεις, μια μέθοδο που υιοθετεί η πλειονότητα των υπαρχόντων στατιστικών αλγορίθμων, αλλά και τις επόμενες. Επιπρόσθετα, σε αντίθεση με άλλα μοντέλα επισημείωσης μερών του λόγου, η προτεινόμενη μέθοδος εκμεταλλεύεται ελάχιστους γλωσσολογικούς πόρους ήτοι ένα μικρό λεξικό με λέξεις που ανήκουν στα άκλιτα μέρη του λόγου και τις λέξεις που ανήκουν στα κλειστά μέρη του λόγου. Οι δυνατότητες του μοντέλου έχουν επαυξηθεί ώστε να προβλέπει και την πτώση της εκάστοτε λέξης. Ενδελεχή πειραματικά αποτελέσματα καταδεικνύουν ικανοποιητική απόδοση που αγγίζει τα ανώτατα επίπεδα του χώρου. Συγκεκριμένα, για την πρόβλεψη μερών του λόγου, η απόδοση του συστήματος αγγίζει το 96% ενώ για την πρόβλεψη της πτώσης φθάνει το 97%.

Keywords

POS tagging, bayesian theory, bayesian networks, natural language processing

Introduction

A plethora of natural language processing tasks utilize the important role of lexical corpora resources, particularly annotated corpora. Since manual creation of such corpora is a difficult process, the development of automatic tools that will assign accurate tags to previously unseen words is of principal significance.

The majority of the existing systems that have been presented for Part-Of-Speech (POS) labelling are either rule-based or stochastic. Rule-based ones (Brill 1992, 1994, Elenious 1990, Voutilainen *et al.* 1992) use handcrafted linguistic information of language or application dependent POS constraints. Significant tagging accuracy is reported when using a restricted POS set. Marcus *et al.* (1993) referred to 94%-98% accuracy on the Penn Treebank corpus. However, when applied to large POS sets or many training data, the number of learned rules increases dramatically, resulting in a highly costly rule definition. Researchers observed that the number of learned rules is linearly increasing with the corpus size.

On the other hand, stochastic taggers use morphological as well as contextual information and obtain their model parameters by applying statistical algorithms to labelled text (Cerf-Danon and El Beze 1991, Church 1988, Kupiec 1992, Wothke *et al.* 1993, Merialdo 1994, Dermatas and Kokkinakis 1995). When a plethora of available data is available, the performance is close to that of rule-based systems. Nevertheless, there are cases where the

theoretical background of such taggers imposes restrictions and assumptions that are not met in real case natural language problems. Dermatas and Kokkinakis (1995) describe a HMM POS tagger which is based on the assumption that each word is uncorrelated with neighbouring words and their tags, a claim which is not necessarily valid in natural language texts.

For the present paper, a novel, Bayesian POS tagger is presented and evaluated for Modern Greek (MG), a language with a high degree of POS ambiguity. The construction and evaluation resources consist of two different corpora of newspaper balanced genre articles consisting of approximately 120,000 and 250,000 words each. This material was assembled in the framework of the ILSP/ ELEYTHEROTYPIA (Hatzigeorgiu et al., 2000) and the ESPRIT-860 (Partners of ESPRIT-291/ 860, 1986) projects. The tagger uses minimal linguistic resources, namely a small lexicon of only 400 entries, containing the words that belong to non-declinable POS categories and closed-class words. It exploits both lexical and contextual information without performing morphological analysis. This results in an adjustable module that could be applied to new languages or new feature sets with trivial effort. Furthermore, an additional case tagging model has been constructed using BBN.

1. Probabilistic analysis of the POS tagging task

In order to approach the task of resolving POS disambiguation, we define a stochastic model over $H*T$, where H is the set of possible lexical and labelling contexts $\{h_1,\ldots,h_k\}$ or *"variables"* and T is the set of allowable POS labels $\{t_1,\ldots,t_n\}$. Using Bayes' rule, the probability of the optimal tag T_{opt} equals to:

$$T_{opt} = \arg\max_{T\in(t_1,\ldots,t_n)} p(T \mid H) = \arg\max_{T\in(t_1,\ldots,t_n)} \frac{p(H \mid T)\cdot p(T)}{p(H)} = \arg\max_{T\in(t_1,\ldots,t_n)} p(H \mid T)\cdot p(T) \quad (1)$$

For a given sequence of observations of variables $h_1,\ldots,h_k$, equation (1) becomes:

$$T_{opt} = \arg\max_{T\in(t_1,\ldots,t_n)} = p(t_i)\cdot p(h_1,\ldots,h_k \mid t_i) \quad (2)$$

The HMM-based taggers assume that each h_i is uncorrelated with the other variables and their corresponding labels, and each label t_i is probabilistically related to the K previous labels only. Therefore, equation (2) is altered to:

$$T_{opt} = \arg\max_{T\in(t_1,\ldots,t_n)} \left(p(t_1)\prod_{i=2}^{k} p(t_i \mid t_{i-1},\ldots,t_i)\prod_{i=k+1}^{N} p(t_i \mid t_{i-1},\ldots,t_{i-N})\prod_{i=1}^{N} p(h_i \mid t_i)\right) \quad (3)$$

Nevertheless, this assumption barely holds true in real natural language texts. Bayesian networks are capable of effectively coping with the non realistic HMM restriction, since they allow stating conditional independence assumptions that apply to variables or subsets of

variables.

A Bayesian network is consisted of a qualitative and quantitative portion, namely its structure and its conditional probability distributions respectively. Given a set of attributes $A=\{A_1,\ldots,A_k\}$, where each variable A_i could take values from a finite set, a Bayesian network describes the probability distribution over this set of variables. We use capital letters as X,Y to denote variables and lower case as x,y, to denote values taken by these variables. Formally, a Bayesian network is an annotated directed acyclic graph (DAG) that encodes a joint probability distribution. We denote a network B as a pair B=<S,P> (Pearl, 1988) where S is a DAG whose nodes correspond to the attributes of A. P refers to the set of probability distributions that quantifies the network. S embeds the following conditional independence assumption:
Each variable A_i is independent of its non-descendants given its parent nodes.

P includes information about the probability distribution of a value a_i of variable A_i, given the values of its immediate predecessors in the graph, which are also called "parents". This probability distribution is stored in a table, which is called conditional probability table. The unique joint probability distribution over A that a network B describes can be computed using:

$$p_B(A_1,\ldots A_n) = \prod_{i=1}^{n} p(A_i \mid parents(A_i)) \tag{4}$$

Given equation (4), equation (2) becomes:

$$T_{opt} = \arg\max_{T\in(t_1,\ldots,t_n)} p(t_i)\prod_{i=1}^{k} p(h_i \mid parents(h_i), t_i) \tag{5}$$

Note that in formula (3) the most probable tag sequence was computed based on the *N-gram* approach solely, meaning that it assumed that only the *N-1* words have an effect on the probabilities of the next word *N*. On the other hand, Bayesian networks allow taking under consideration words (or "features" in general) that could be situated after the word whose tag is to be inferred apart from the *N-1* previous ones.

2. Linguistic Ressources

Modern Greek has a complex inflectional system. There are eleven different POS categories: articles, nouns, adjectives, pronouns, verbs (a participle is considered a sub-category of a verb), numerals, adverbs, prepositions, conjunctions, interjections and particles. The first six (articles, nouns, adjectives, pronouns, verbs and numerals) are declinable; the remaining five (adverbs, prepositions, conjunctions, interjections and particles) are indeclinable. Moreover, all indeclinable words plus articles and pronouns form closed sets of words (meaning that they are limited to a few dozens and no new words are added to these classes) while nouns, adjectives,

and verbs form open sets (i.e. their number is practically unlimited, since new words are added to these classes as the language evolves over time).

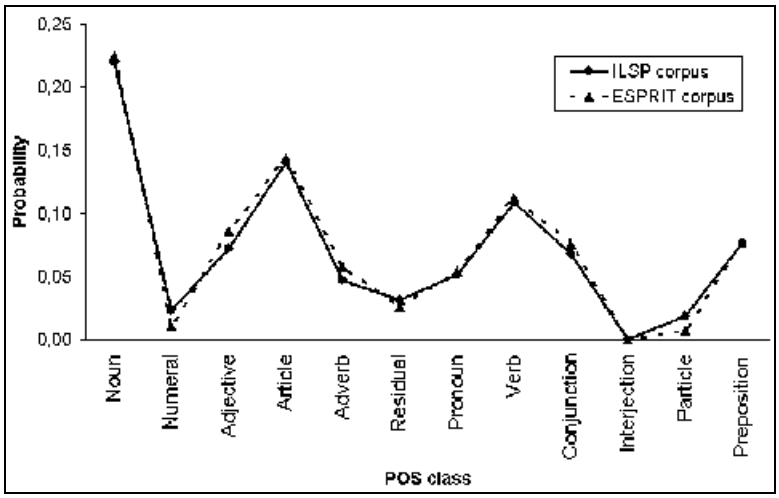
The case attribute characterizes nouns, adjectives, numerals and participles. Its possible values are: nominative, genitive, accusative and vocative. Although the dative case was extensively used in Ancient Greek, it appears in MG texts only within archaized expressions.

2.1 Corpora

The POS tagger was constructed and thoroughly evaluated using two different corpora of balanced genre newspaper articles. The former is the ILSP/ ELEFTHEROTYPIA corpus, consisting of 250.000 morphologically annotated texts by experienced linguists. The latter is the ESPRIT(291-860) Greek corpus of about 120.000 words, morphologically annotated using semi-automatic tagging tools.

An analysis on the distribution of POS tags in the two corpora revealed that despite the fact that they have been annotated using different methods and that they contain different texts, the POS categories of the containing words present approximately the same distribution (Figure 1). This observation contributes significantly to the process of choosing among the training and test data sets, since it indicates that there is no evidence that the model parameters of the trained model will not represent the test set.

Figure 1. Distribution of the grammatical classes of words for both types of Modern Greek corpora.



2.1.1 POS/Case ambiguity

Table 1 tabulates the POS label set which includes common categorization of the grammatical information for the two corpora. POS ambiguity relies mainly on the fact that certain adjectives and adverbs share the same orthographic form. The same holds for articles and some particular pronouns. As an example:

"Ήπιε πολύ" (He drank a lot) and
"Ήπιε πολύ κρασί" (He drank a lot of wine).

In the first case, the word *"πολύ"* (a lot) is an adverb, while in the second example *"πολύ"* is an adjective. The POS corpus ambiguity was calculated by the mean number of possible tags for each word for the whole set of grammatical features. Taking into account that the POS feature set of the training set and that of the test set would be identical, the ambiguity of a 10,000 words test set was computed for a 50,000 words training corpus (Table 2).

Table 1: The considered set of grammatical features

Pos category	POS-specific features	Common features
Adjective (ADJ)	Degree	Gender Number Case
Noun (N)	Common/proper	
Pronoun (PN)	Personal/relative, interrogative, person	
Participle (V)	Sub-category of verb	
Article (ART)	Definite/indefinite	
Numeral (NUM)	Ordinal/cardinal	
Verb (V)	Voice, mood, person, number	
Conjunction (CON)	Coordinating/subordinating	
Particle (PAR)	Of negation, of future, subjunctive	
Preposition(PRE)		
Adverb (ADV)		
Interjection (INT)		
Residuals (RES)	Acronym/abbreviation/foreign word	

In MG, case ambiguity relies on the fact that in many cases, declinable words have the same orthographic form in the nominative as well as in the accusative case. This applies for almost all nouns, adjectives, articles, pronouns and ordinal numerals, feminine and neutral, singular and plural. Consider the following example:
"Ακούει το παιδί" (*The child is listening*) and *"Ακούει το παιδί"* (*Someone is listening to the child*). In the first example the noun phrase *"το παιδί"* (*the child*) is in the nominative case (subject), while, in the second, it is in the accusative (object).

Table 2: Corpus ambiguity in terms of POS/Case tags for both corpora

Corpus Ambiguity	ILSP	ESPRIT
POS	1.789	1.855
Case	1.673	1.7

Taking into account the results of Table2, it is worth noting that the task of correctly identifying the POS and case label of a word is particularly difficult for MG due to the great number of duplicate tags each word might have. Furthermore, since we do not incorporate a large known words lexicon, the ambiguity is further increased.

2.2 Dealing with unseen words

We do not distinguish between known and unknown words in the corpus. For all methods adopting the above distinction, their model uses the set of known words for training and the unknown ones for testing. However, according to Dermatas and Kokkinakis (1995), the POS distribution of known words differs significantly from that of unknown words for seven European languages, including MG. Therefore, there is a great possibility that the known-word-based training model includes parameters that do not reflect the test set parameter distribution accurately. Moreover, using known words as training resources poses another, machine-learning complication. The trained model forces its feature expectations to match with those observed in the training data, resulting in a model that tends to perfectly classify the instances of the training set. While this seems like a reasonable strategy, it potentially leads to *"overfit"* the data, and is therefore not able to accurately classify a word that did not appear in the training set. This occurs when there is noise in the data or in case the features are somewhat insignificant to the target classification function. In the POS domain, where the POS ambiguity is particularly high, training a system using known words as features could lead to a very accurate classifier if these words appear in the test set, but to a very poor performance in case many unseen terms are found within the instances. In our approach, we consider as lexical resources only words that belong to non-declinable POS categories and closed-class words (like articles, pronouns, etc.) and a short, 150 entries list of suffixes of MG words, which do not expose such anomalies in the distribution of their grammatical properties. However, using BBN for training, poses a restriction: When the conditional probability table of the model is calculated from a limited amount of training data, a large number of tagging errors is observed, due to inaccurate estimation of conditional probabilities. Researchers that conducted POS labelling experiments on MG (Dermatas and Kokkinakis 1995 and Orphanos *et al.* 1999), report "baseline" results obtained from a 20,000 word training set.

3. Experimental Results

In order to conduct POS and case tagging experiments, each of the MG corpora has been partitioned using the 10-fold cross-validation method. The POS tagger accuracy has been calculated as the number of correctly classified POS labels by the number of words within the open test set. The case tagger accuracy has been measured by counting the number of correctly classified cases, again divided by the number of words found in the test set. We evaluated the results using our BBN implementation against a freely available, language-independent stochastic tagger, named *Tree-tagger*

During the evaluation process, the best window size in terms of efficiency versus computational complexity was found to be *{-3,+1}*. Figure 2 illustrates the progress of POS and case error rate using Bayesian networks for different sizes of training data. Figure 3 illustrates the progress of POS and case error rate using different sizes of training data against the POS tagging error rate of *Tree-tagger*.

Figure 2. Percentage of Bayesian networks POS and case error rate in the test set of 10,000 words for fluctuant training text.

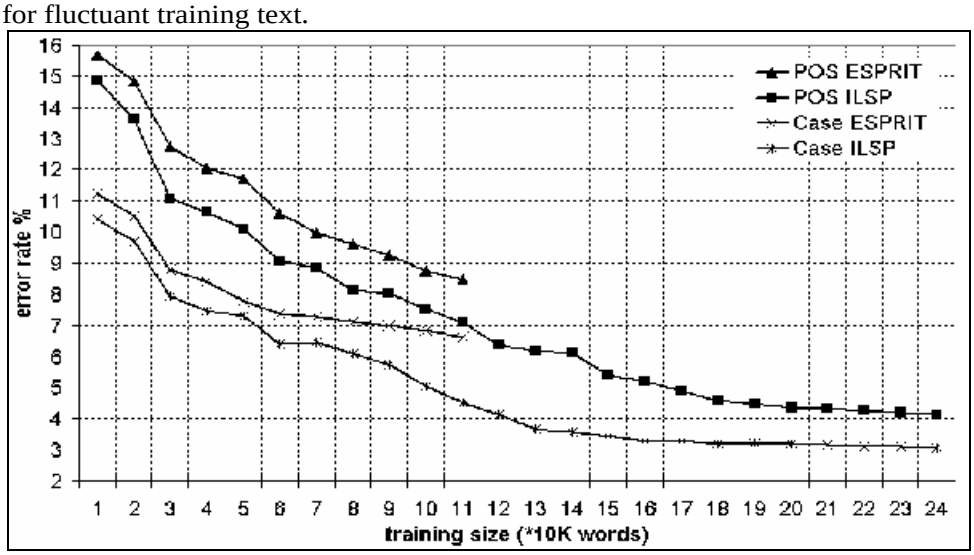
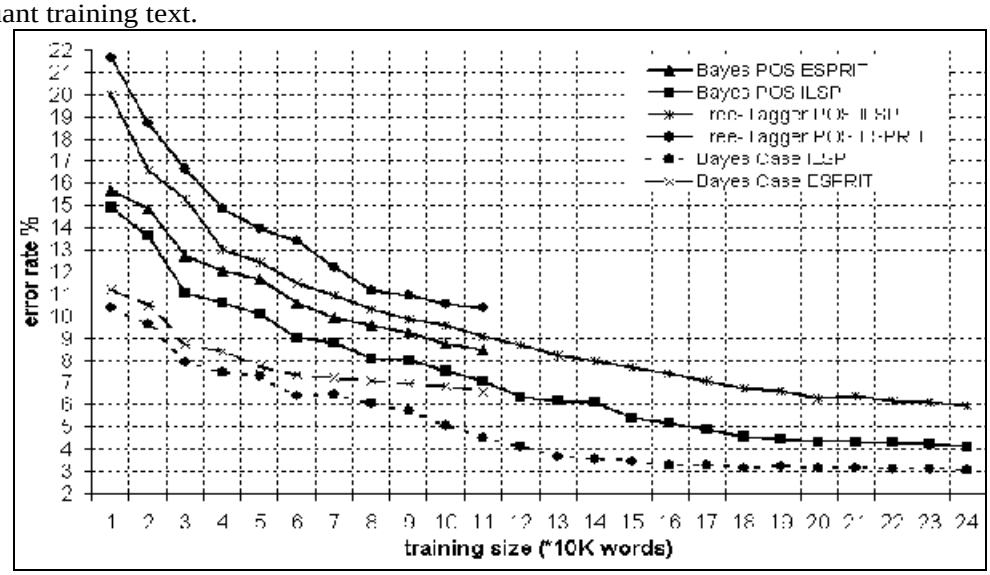


Figure 3. Evaluation of Bayesian networks versus Tree-Tagger in the test set of 10,000 words for fluctuant training text.



Conclusion

The proposed Bayesian network framework proves to be a flexible methodology for linguistic modelling, since it can exploit a rich linguistic feature set in the framework of a probability model. The strict assumption of the HMM stochastic models concerning the *N-1* previous words influence, which is not always encountered in real natural language problems, has been

successfully alleviated using the Bayesian model, by allowing taking *N+1 word* into account as well. In this paper, an implementation of this model resulted in a state-of-the-art POS and case tagger, as evidenced by the 96% and 97% accuracy respectively.

We have evaluated our systems using in two MG newspaper corpora, annotated with different methodologies, using a rich grammatical tag set (12 categories for POS and 6 for case tagging). Minimal linguistic knowledge has been incorporated in order to avoid using a large known words lexicon that could potentially produce a model that would overfit the data, thus performing poor when confronting with unknown terms. Furthermore, the model parameters are adjusted when new training data is entered, resulting in improvement of the accuracy, as expected.

References

Brill E. (1992) A *simple rule-based part of speech tagger*. In "Proceedings of the Third Conference on Applied Natural Language Processing", Trento, Italy, pp. 152-155.

Brill E. (1994) *Some advances in transformation-based part of speech tagging*. In "Proceedings of the Twelfth National Conference on Artificial Intelligence", volume 1, pp. 722-727.

Cerf-Danon H, and El-Beze M. (1991) *Three different probabilistic language models: Comparison and combination*. In "Proceedings of the International Conference on Acoustics, Speech and Signal Processing", pp. 297-300.

Church K. (1988) *A stochastic parts program and noun phrase parser for unrestricted text*. In "Proceedings of the Second Conference on Applied Natural Language Processing", Austin, Texas, pp. 136-143.

Dermatas E, and Kokkinakis G. (1995) *Automatic stochastic tagging of natural language texts*. Computational Linguistics, 21/2, pp. 137-163.

Elenius K. (1990) *Comparing connectionist and rule based model for assignment parts-of-speech*. In "Proceedings of the International Conference on Acoustics, Speech and Signal Processing", pp. 597-600.

Hatzigeorgiu N., M. Gavriilidou, S. Piperidis, G. Carayannis, A. Papakostopoulou, A. Spiliotopoulou, A. Vacalopoulou, P. Labropoulou, E. Mantzari, H. Papageorgiou and I. Demiros. (2000) *Design and Implementation of the online ILSP Greek Corpus*. Proceedings of LREC 2000, 1737-1742. Athens, Greece.

Kupiec J. (1992) *Robust part-of-speech tagging using a Hidden Markov Model*. Computer, Speech & Language, 6/3, pp. 225-242.

Lam, W. and F. Bacchus.(1994) *Using New Data to Refine a Bayesian Network*, In Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Mateo, California, 383-390.

Marcus M., Santorini B. and Marcinkiewicz M (1993) *Building a large annotated corpus of English: The Penn Treebank*. Computational Linguistics, Computational Linguistics, 192 pp. 315-330.

Merialdo B. (1994) *Tagging English text with a probabilistic model*. Computational Linguistics, 20/2, pp. 155-171.

Orphanos D., Kalles D, Papagelis A. and Christodoulakis (1999) *Decision trees and NLP: A Case Study in POS Tagging*. In "Proceedings of the ECCAI Advanced Course on Artificial Intelligence (ACAI), Chania, Greece, 1999".

Partners of ESPRIT-291/ 860. (1986) *Unification of the word classes of the ESPRIT Project 860*. BU-WKL-0376. Internal Report.

Pearl J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Suzuki J. (1993) *A construction of Bayesian networks from databases on a MDL scheme*. In Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence, 266-273. San Francisco, CA.

Voutilainen A., Heikkila J. and Antitila A. (1992) *Constraint grammar of English*. In "Publication 21, Department of General Linguistics, University of Helsinki", Finland.

Wothke K., Weck-Ulm I., Heinecke J., Mertineit O. and Pachunke T. (1993) S*tatistically based automatic tagging of German text corpora with parts-of-speech some experiments*. TR75.93.02-IBM. IBM Germany, Heidelberg Scientific Center.