Towards a Bayesian Stochastic Part-Of-Speech and Case Tagger of Natural Language Corpora

Manolis Maragoudakis, Katia Kermanidis and Nikos Fakotakis
Speech and Language Technology Group
Department of Electrical and Computer Engineering
University of Patras
26500 Rion
Patras, Greece, 26500
mmarag,kerman,fakotaki@wcl.ee.upatras.gr

Abstract

This paper introduces and evaluates a Bayesian Network probabilistic model for automatic Part-Of-Speech tagging of Modern Greek natural language texts. The Bayesian model for the task of POS tagging is mathematically formed and is compared to that of Hidden Markov, a broadly applied methodology. Our model is trained from annotated corpora, using lexical as well as contextual information. Unlike the majority of existing taggers, it uses minimal linguistic resources, namely a small lexicon which contains the words that belong to non-declinable POS categories and closed-class words. Furthermore, the model is augmented to infer on the case of an unseen word as well. Experimental results depict accuracy in the range of 91%-96% for POS and a range of 93%-97% for case tagging.

Introduction

A significant number of natural language processing tasks exploit the key role of lexical corpora resources, particularly annotated corpora. Due to the fact that manual construction of such corpora is a laborious procedure, developing automated tools that will assign accurate tags to previously unseen words is of paramount importance. Under the perspective of real world applications, the acceptance of such taggers is originated on the ability to cope with unknown words, on self-improvement by training and on being able to minimize the tagging error rate.

The majority of the existing systems that have been presented for Part-Of-Speech (POS) labelling utilize either rules or stochastic approaches. Rule-based ones (Brill 1992, 1994, Elenious 1990, Vontilainen *et al.* 1992) use handcrafted linguistic knowledge of language or application-dependent POS constraints. Significant tagging accuracy is reported when using a restricted POS set. Marcus *et al.* (1993) referred to a value of 94%-98% in accuracy on the Penn Treebank corpus. However, when applied to large POS sets or voluminous training data, the number of learned rules increases dramatically, resulting in a highly costly rule definition. For example, Petasis *et al.* (1999) observed that the number of learned rules is linearly increasing with the corpus size.

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 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. In recent years, there has been a shift from treating POS taggers as a model of the sequence structure of sentences to consider them as classifiers. Ratnaparkhi (1996) presented a maximum entropy model for POS tagging. The context can be represented in terms of a rich set of features (e.g. surrounding words, tags and word-form features such as suffixes or prefixes). The construal of the POS tagging task as such a classification problem allows one to use a plethora of machine learning algorithms.

For the present paper, a novel, Bayesian Belief Network (BBN) 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[1] and the ESPRIT(291-860) 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. Experimental results indicate a POS tagging accuracy in the range of 91%-96% and a range of 93%-97% in case tagging. The inference of the case is performed given the POS tagger's predicted POS rather than having it extracted from the test data set.

The structure of this paper is as follows: In Section 1, the Bayesian stochastic tagging model is presented and theoretically evaluated against that of HMM. In Section 2 the morphological properties of MG and the target language resources are discussed, plus a detailed analysis on the influence of known and unknown words in the process of training the model. In Section 3 a short presentation of the implementation is mustered, followed by a detailed description of the experimental results in Section 4. The concluding remarks are discussed in Section 5.

1. The probability model

Researchers that focus on the stochastic modelling of POS disambiguation, define the 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:

$$\mathrm{T_{opt}} = \underset{\mathrm{T}\in(\mathrm{t_1},\ldots,\mathrm{t_n})}{\arg\max}\, p(\mathrm{T}\mid\mathrm{H}) = \underset{\mathrm{T}\in(\mathrm{t_1},\ldots,\mathrm{t_n})}{\operatorname{argmax}} \frac{p(\mathrm{H}\mid\mathrm{T})p(\mathrm{T})}{p(\mathrm{H})} = \underset{\mathrm{T}\in(\mathrm{t_1},\ldots,\mathrm{t_n})}{\operatorname{argmax}}\, p(\mathrm{H}\mid\mathrm{T})p(\mathrm{T}) \qquad (1)$$

Approximations of the probability distributions of (1) deal with the trade-off between computational complexity and efficiency. For a given sequence of observations of variables $h_1,\ldots,h_k$, equation (1) becomes:

$$T_{opt} = \underset{T\in(t_1,\ldots,t_n)}{\operatorname{argmax}}\, p(t_i)p(h_1,\ldots,h_k\mid t_i) \qquad (2)$$

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:

$$\mathrm{T_{opt}} = \underset{\mathrm{T}\in(\mathrm{t_1},\ldots,\mathrm{t_n})}{\operatorname{argmax}}\left[p(\mathrm{t_1})\prod_{\mathrm{i=2}}^{\mathrm{k}} p(\mathrm{t_i}\mid \mathrm{t_{i-1}},\ldots,\mathrm{t_1})\prod_{\mathrm{i=k+1}}^{\mathrm{n}} p(\mathrm{t_i}\mid \mathrm{t_{i-1}},\ldots,\mathrm{t_{i-N}})\prod_{\mathrm{i=1}}^{\mathrm{n}} p(\mathrm{h_i}\mid \mathrm{t_i}) \right] \qquad (3)$$

However, as mentioned before, 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. In general, a BBN represents this knowledge as a directed acyclic graph with nodes corresponding to variables and arcs depicting their interconnection. A BBN also encodes the degree of the dependencies among variables in terms of probabilities in an embodied conditional probability table. The joint probability of any assignment of values $<a_1,\ldots,a_n>$ to the tuple of network variables $<A_1,\ldots A_n>$ is given by the formula:

$$\mathrm{p(a_1,\ldots,a_n)}=\prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{p(a_i|Parents(A_i))} \qquad (4)$$

[1] http://www.elda.fr/cata/text/W0022.html

where $Parents(A_i)$ denotes the set of immediate predecessors of the variable A_i in the network. This set is comprised of all the variables that directly affect the conditional distribution of A_i. The values of probability $p(a_i \mid Parents(A_i))$ are precisely kept in the conditional probability table. Given equation (4), equation (2) becomes:

$$T_{opt} = \underset{T \in (t_1,\ldots,t_n)}{\operatorname{argmax}} \, p(t_i) \prod_{j=1}^{k} p(h_j | Parents(h_j), t_i) \qquad (5)$$

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 besides the *n-1* previous ones.

The estimation of probabilities in formula (5) is straightforward, provided that we know the interconnection of our data features i.e. the structure of the Bayesian network. In our case, provided that the tagger should be fully automated and free of linguistic handcrafted knowledge, this network structure was not known in advance. Thus, a search algorithm that would learn it from the observable data should be applied. Since the problem of learning networks from data is known to be NP-hard (Mitchell 1996) (i.e. there are $2^{\frac{n(n-1)}{2}}$ possible networks that could describe *n* different variables), we used a modification of the Cooper and Herskovits (1992) heuristic search algorithm *K2*. Instead of comparing all candidate networks, we consider investigating among the set of those that resemble the current best network most. A detailed description of the search process is included in Section 3, where the implementation issues are presented.

2. Resources and task complexity

The Greek language 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.

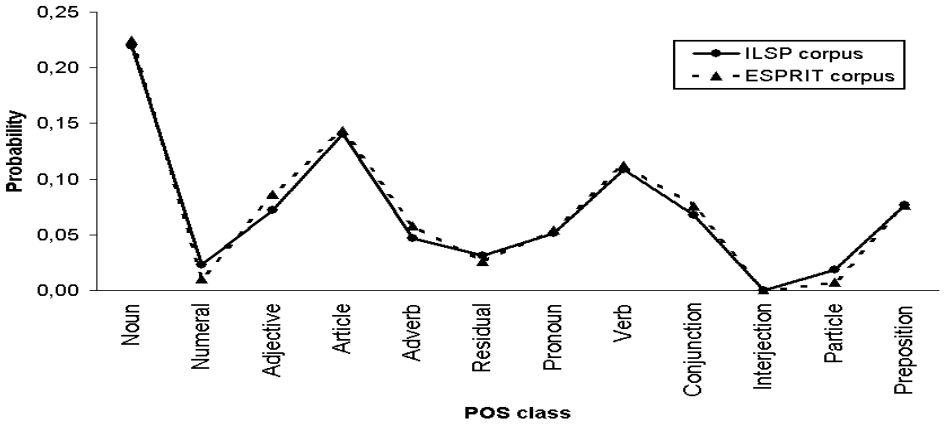

Figure 1: Distribution of the grammatical classes of words for both types of Modern Greek corpora.

2.1.1 POS/Case corpus ambiguity

Table 1 tabulates the POS label set which includes common categorization of the grammatical information for the two corpora.

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	

Table 1: The considered set of grammatical features.

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:
"*Ηπιε πολύ*" (He drank a lot) and
"*Ηπιε 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).

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).

Corpus Ambiguity	ILSP	ESPRIT
POS	1.789	1.855
Case	1.673	1.7

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

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 Coping 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, Petasis *et al.* 1999, Orphanos *et al.* 1999), report "baseline" results obtained from a 20,000-word training set. The BBN POS tagger needed about 30,000 words in order to reach the same level of performance. Nevertheless, this is essentially disadvantageous for languages with limited available annotated texts. MG and the majority of European Languages have a plethora of existing annotated large corpora, so the threshold of 30,000 words could be put across.

3. Implementation

As described in Section 1, the Bayesian model hypothesis space is defined over *H*T*, where *H* contains the complete set of lexical and contextual features that were considered, while *T* denotes the target values for the POS or case classification. For the task of POS tagging, *T* contains the 12 different categories found in Table 1. Concerning the case tagger, 6 different values were taken into account: accusative, nominative, genitive, dative, vocative and a *null* value to denote that the word does not contain any case. As regards to the feature selection, contextual and lexical information was integrated. A small lexicon of 400 words belonging to the group of non-declinable POS categories and closed-class words was maintained, plus a list of all suffixes of MG language

(around 150 different unambiguous suffixes). Furthermore, morphological features were included such as the gender, number and case of a word.

Unlike most previous approaches that attempted to estimate the optimal POS tag of a word by inducing inference based on the *N-1* previous ones, a flexible-sized window of *{N-1,N+1}* words was utilized. The *N-1* words incorporated all the lexical and morphological features while the *N+1* words and the current one were only tagged according to their suffixes and their presence in the 400 items lexicon. The parsing of the training corpus resulted in a bi-dimensional table *D=I*J* where each row corresponds to the pair of word-selected window and each column indicates each feature of that pair plus the POS and case category of the word whose POS and case should be learned.

In the process of determining the most probable network structure from data, prior knowledge of the impact each feature poses to the POS class was intentionally not given to the model. The aim of the proposed tagger is to automatically determine the inter-relation between feature nodes and the class node, hence not be a language dependent model. The parameter needed for evaluating two candidate networks is their probability ratio over the set of training examples. If we denote *r* as the relation of two networks B_1 and B_2 respectively, then using Bayes' theorem:

$$r = \frac{p(D \mid B_1)}{p(D \mid B_2)} \tag{6}$$

$$p(D \mid B) = \frac{p(B \mid D)p(D)}{p(B)} \tag{7}$$

$$r = \frac{p(B_1 \mid D)\,p(B_2)}{p(B_2 \mid D)\,p(B1)} \tag{8}$$

Having not seen the data, no prior knowledge is obtainable and thus no straightforward method for computing probabilities $P(B_1)$ and $P(B_2)$ of equation (8) is feasible. A common way to deal with this, is to assume that all networks has the same probability, so equation (8) becomes:

$$r = \frac{p(B_1 \mid D)}{p(B_2 \mid D)} \tag{9}$$

The columns of *D* correspond with the nodes of the BBN, and the degree of each node's influence to the other was determined using data from the rows of *D*. The probability of each candidate network *B* given table *D* was computed using the formula of Glymour and Cooper (1992):

$$p(B \mid D) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma\left(\frac{\Xi}{q_i}\right)}{\Gamma\left(\frac{\Xi}{q_i} + N_{ij}\right)} \prod_{k=1}^{r_i} \frac{\Gamma\left(\frac{\Xi}{r_i q_i} + N_{ijk}\right)}{\Gamma\left(\frac{\Xi}{r_i q_i}\right)} \tag{10}$$

where:

- Γ is the gamma function.
- N equals to the number of variables.
- r_i denotes the number of values in i:th variable and q_i denotes the number of possible different data value combinations the parent variables can take.
- N_{ij} depicts the number of rows in data that have j:th data value combinations for parents of i:th variable. N_{ijk} corresponds to the number of rows that have k:th value for the i:th variable and which also have j:th data value combinations for parents of i:th variable.
- Ξ is the equivalent sample size, equals to the average number of values variables have, divided by 2.

Since, as the number of nodes increases (the window size is expanded), the population of possible networks becomes cumbersome, a search strategy had to be followed: Initially, the most probable forest-structured network is constructed (i.e. a network in which every node has at most one parent). A greedy search is performed by adding, deleting or reversing the arcs randomly. In case that a change results in a more probable network, it is accepted, otherwise cancelled. Throughout this process, a repository of networks with high probability is maintained. When the search reaches a local maximum, a network is randomly selected from the repository and the search process is activated again. It should be noted that in order to avoid the convergence to the previous local maximum the network is slightly modified, meaning that we delete some arcs. Since the training data set is large we also sub-sample the data to speed the network evaluation process up. During the search, the size of the sub-samples is increased. A restriction on the network complexity is also applied during the search, so that a limited number of arcs is allowed in the beginning and, as the process progresses, more and more arcs are approved. These two annealing schemes (sub-sampling and complexity restrictions) have proven to have the effect of avoiding many bad local maxima.

4. Experimental Results

In order to conduct POS and case tagging experiments, each of the MG corpora has been partitioned into equal parts of 10,000 words. The 90% of those parts have been used for training material while the remaining 10% have comprised the testing set. This process has been repeated 10 times, using different parts of the text for training and evaluation each time. The performance has been measured by the average accuracy over the 10 repetitions. 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. It is important to stress that the case tagger infers about the case of a word taking the POS tagger estimation on the POS of that word into account, in order for the tagger to be applicable on raw corpora. Nevertheless, the module is parametric concerning this factor, so the user could decide whether the POS information will be available or it should be inferred from the POS tagger.

During the evaluation process, the best window size in terms of efficiency versus computational complexity was found to be *{-3,+1}*. The computational time of the Bayesian network learning phase is $O(N^2)$, where N is the number of linguistic features for a given pair of word-window. The confusion matrices, obtained using the best window, of both POS and case tagging for each corpus are tabulated in Tables 3 and 4. An additional metric of % Per Class Accuracy *(%PCA*) is presented in order to provide a clearer view of each tagger's potential classification weaknesses. *%PCA* is the percentage of the correctly labelled tags divided by the number of instances of that class found in the testing words collection. Figure 2 illustrates the progress of POS and case error rate using different sizes of training data. The immediate improvement of the model using 30,000 training words could be observed. This supports the theoretical claim that more training data is needed by the BBN approach than by the other stochastic or rule-based approaches.

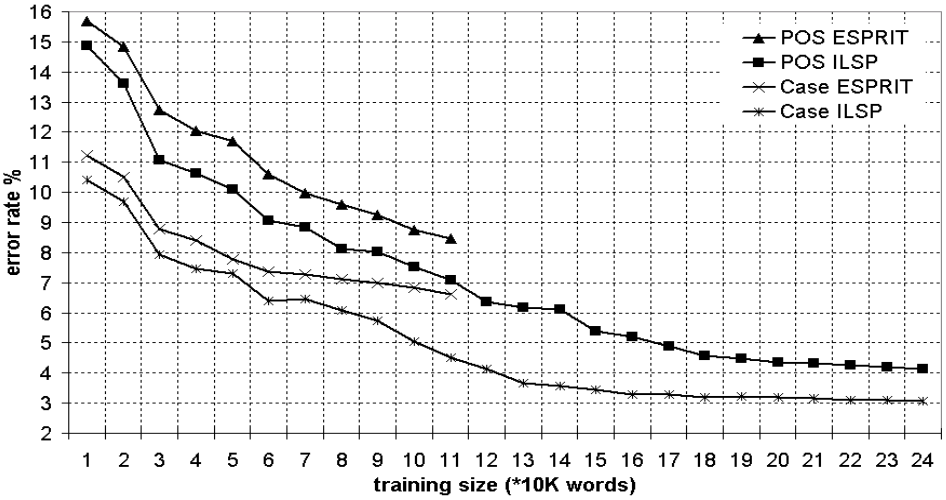

Figure 2: Percentage of POS and case error rate in the test set of 10,000 words for fluctuant training text

ILSP							**Total Accuracy: 96.9%**
CASE	*accusative*	*dative*	*genitive*	*nominative*	*none*	*vocative*	*%PCA*
accusative	**5464**	-	47	534	-	-	90.4%
dative	-	**7**	-	-	-	-	100%
genitive	27	-	**3238**	21	-	-	98.5%
nominative	279	-	26	**3755**	-	-	92.5%
none	-	-	-	-	**11597**	-	100%
vocative	-	-	-	-	-	**4**	100 %
ESPRIT							**Total Accuracy: 93.4%**
CASE	*accusative*	*dative*	*genitive*	*nominative*	*none*	*vocative*	*%PCA*
accusative	**2744**	-	24	286	43	-	88.6%
dative	-	**4**	-	-	-	-	100%
genitive	28	-	**1588**	37	31	-	94.3%
nominative	158	-	21	**1850**	49	-	89%
none	54	-	23	58	**5799**	-	97.7%
vocative	-	-	-	-	-	**2**	100%

Table 3: Confusion matrix of the case tagger for both corpora.

As can be observed from Table 3, the high degree of ambiguity regarding the nominative and accusative case is the only reason case tagging accuracy drops below 95%. Because of the fact that the most declination categories nouns, adjectives and numerals share common suffixes, POS tagging accuracy for adjectives and numerals reaches 75% and 82.5% respectively (numerals are a more restricted group of words than adjectives, which explains the higher accuracy). Nouns, on the other hand, usually follow adjectives within a noun phrase. Therefore, POS information of the previous word helps increase noun POS tagging accuracy to over 90%. For the rest of POS categories, accuracy does not fall below 97%. The decrease in accuracy in ESPRIT could be attributed to the basically automatic annotation of the corpus.

ILSP												**Total Accuracy: 96%**	
POS	*ADJ*	*CON*	*INT*	*NUM*	*N*	*PN*	*ADV*	*PRE*	*ART*	*PAR*	*V*	*RES*	*%PCA*
ADJ	**1575**	-	-	10	533	-	-	-	-	-	-	-	74.4%
CON	-	**1977**	2	-	-	-	-	-	-	-	-	-	99.9%
INT	-	-	**7**	-	-	-	-	-	-	-	-	-	100%
NUM	19	-	-	**548**	99	-	-	-	-	-	-	-	82.3%
N	299	-	-	35	**6067**	-	-	-	-	-	-	-	94.8%
PN	-	-	-	-	-	**1497**	-	-	-	-	-	-	100%
ADV	-	-	-	-	-	-	**1325**	-	-	-	-	41	97%
PRE	-	-	-	-	-	-	-	**2239**	-	-	-	-	100%
ART	-	-	-	-	-	-	-	-	**4142**	-	-	-	100%
PAR	-	1	1	-	-	-	-	-	-	**530**	-	-	99.7%
V	-	-	-	-	-	-	-	-	-	-	**3169**	-	100%
RES	-	-	-	-	-	-	28	-	-	-	-	**855**	96.8%
ESPRIT												**Total Accuracy: 91.9%**	
POS	*ADJ*	*CON*	*INT*	*NUM*	*N*	*PN*	*ADV*	*PRE*	*ART*	*PAR*	*V*	*RES*	*%PCA*
ADJ	**699**	-	-	11	371	-	-	-	-	-	-	-	64.7%
CON	-	**1003**	13	-	-	-	-	-	-	-	-	-	98.7%
INT	-	-	**3**	-	-	-	-	-	-	-	-	-	100%
NUM	10	-	-	**281**	51	-	-	-	-	-	-	-	82.4%
N	153	-	-	181	**2989**	-	-	-	-	-	-	-	89.9%
PN	-	-	-	-	-	**766**	-	-	-	-	-	-	100%
ADV	-	-	-	-	-	-	**676**	-	-	-	-	21	96.9%

PRE	-	-	-	-	-	-	-	**1146**	-	-	-	-	100%
ART	-	-	-	-	25	-	-	-	**2087**	-	-	-	98.8%
PAR	-	20	6	-	-	-	-	-	-	**258**	-	-	90.8%
V	-	-	-	-	-	-	-	-	-	-	**1623**	-	100%
RES	-	-	-	-	-	-	54	-	-	-	-	**353**	86.7%

Table 4: Confusion matrix of the POS tagger for both corpora.

Conclusion

The Bayesian network model is a flexible technique 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* words 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, shown in Tables 3 and 4.

The taggers have been evaluated 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 wide coverage lexicon of known words 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 accuracy improvement, as was expected.

Acknowledgements

Our thanks go to Tommi Tsilander, whose contribution to the process of effectively dealing with Bayesian network learning issues throughout both the construction and evaluation phase was of paramount importance.

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.

Cooper G, and Herskovits E. 1992 *A Bayesian method for the induction of probabilistic networks from data*. Machine Learning, 9, pp. 309-347.

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

Elenius K. (990 *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.

Glymour C. and Cooper G. (eds) 1999 *Computation, Causation & Discovery*. AAAI Press/The MIT Press, Menlo Park.

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

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.

Mitchell T. 1997 *Machine Learning*. Mc Graw-Hill.

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".

Petasis G., Paliouras G., Karkaletsis V., Spyropoulos C. and Androutsopoulos I. 1999 *Resolving part-of-speech ambiguity in the Greek language using learning techniques.* In "Proceedings of the ECCAI Advanced Course on Artificial Intelligence (ACAI), Chania, Greece, 1999".

Ratnaparkhi A. 1996 *A Maximum Entropy Part-Of-Speech tagger*. In "Proceedings of the Conference of Empirical Methods in Natural Language Processing, May 17-18, 1996, University of Pennsylvania".

Vountilainen 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.