

A SYSTEM FOR AUTOMATIC TEXT LABELLING

E. Dermatas, G. Kokkinakis

Laboratory of Wire Communications, University of Patras, Greece

ABSTRACT

This paper presents a system for automatic labelling of natural language texts according to a more or less detailed system of linguistic categories (grammatical, syntactical, etc.). A Markovian model is used to predict the label of each word of the unknown text. Several assumptions and restrictions improve the computational efficiency with a small decrease of the performance of the system. This has been measured by labelling 120.000 words of Greek newspaper texts with grammatical labels and proved to be satisfactory.

1. INTRODUCTION

Processing systems of natural languages rely on databases of texts labelled according to appropriate linguistic categories (grammatical, syntactical, etc.) [1]. Training such systems in the above databases, the necessary parameters of the probabilistic or deterministic models which they use are extracted.

This paper presents an automatic system which fully labels all the words of a text, including words which do not exist in the database (unknown words). The labelling is then checked for correction by the user of the system and the corrected text is appended to the database. This system is an improved version of a recently reported labelling system [2] in which unknown words were manually labelled.

The described system predicts the most probable sequence of labels in each sentence of the text on the basis of a 2nd to 5th class Markovian model. The user of the system has the possibility to choose the class of the Markovian model and the conditional probabilities used by the model for each class. The class and the conditional probabilities chosen for the labelling system have a significant influence on its performance.

The structure of this paper is as follows: The probabilistic model used by the labelling system is presented first. Then a detailed description of the labelling system is given. In the last section the performance of the system is discussed on the basis of several experimental results.

2. THE PROBABILISTIC MODEL OF THE SYSTEM

Given an unlabelled text and a set of possible labels for each word, the optimum sequence of labels is estimated by maximizing the probability of the sequence of labels.

Let $x(n)$, $n=1, N$ be a sequence of N words which build a sentence in the text. A sentence in a probabilistic model is a sequence of words which is statistically independent of the previous and following text. In natural languages a set of separators (punctuation marks and function words) defines the sentences. If the set $G(n)$, $n=1, N$ of the possible labels of $x(n)$ is known, the most probable sequence of labels (optimum solution) can be defined as:

$$V(n) = \underset{m_i}{\operatorname{argmax}} (P_{m_i}) \quad (1)$$
$$i=1, \prod_{n=1}^N a(n)$$

where:

$a(n)$ is the number of labels for $x(n)$,
 m_i is a sequence of N labels contained in the set M of all sequences of labels which can be produced by the sets $G(n)$, i.e.

$$M = G(1) \times G(2) \times \dots \times G(N) \quad (2)$$

P_{m_i} is the probability of occurrence of the sequence of labels m_i . The estimation of P_{m_i} is obtained by assuming that the labels of words which have a greater positional distance than the class of the model, are statistically independent. E.g. if we assume in the 2nd class model that every label depends only on the previous one, we can estimate the probability P_{m_i} from (3):

$$P_{m_i} = \prod_{n=2}^N P(g(n) | g(n-1)) \quad (3)$$

where $g(n) \in G(n)$.

From the sets $G(n)$ all the possible sequences of labels are created, (relation 2) and the sets of conditional probabilities are established from which the probabilities of the sequences of labels are estimated.

3. THE LABELLING SYSTEM

The labelling system consists of two processors:

- i. The learning processor which updates the parameters of the system by using already labelled text.
- ii. The labelling processor which labels new unlabelled text.

Figure 1 presents a block diagram of the procedures involved in the parameters and database updating.

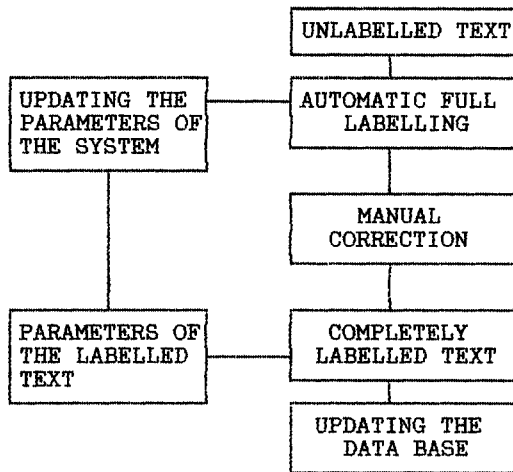


Fig.1 Block diagram of the parameters and the data base updating procedures.

3.1 Learning processor.

From a newly labelled text the learning processor updates:

- (a) The dictionary (L) of the different word forms extracted from the learning texts, with their labels.
- (b) The set (S) of F label sequences, (where F is the class of the model), with the (absolute) frequency of occurrence of each label. From the set (S) all the possible conditional probabilities are obtained which correspond to the label transitions included in set (S). For unknown label transitions the zero conditional probability is assumed in order to speed up the process of estimating the optimum solution.
- (c) The set (A) of different labels with their absolute frequency of occurrence.

The (L, S, A) sets include the parameters of the labelling system which are updated from the completely labelled text. The accuracy of the parameters is increased with the size of the learning texts. As a result, the performance of the system is increased.

3.2 Labelling processor.

The labelling processor can be defined as a set of parameters Mm:

$$Mm = (L, S, A, B, Se, mp, mu, Pz, Pu) \quad (4)$$

where:

(B) is the set of labels that are not used for labelling unknown words to the system. This set must be defined manually. By using the set (B) in the labelling processor, the number of possible labels for unknown words is decreased and the labelling process is speeded up.

(Se) is the set of separators. This set is also defined manually.

mp is the maximum number of words allowed to be processed in a sentence.

mu is the maximum number of unknown words allowed to be processed in a sentence.

Pz is a probability threshold. All the conditional probabilities estimated from set (S) which are less than Pz, are set to zero.

Pu is the absolute probability of a set of labels, the set (Lu) which includes possible labels for unknown words.

(Lu) is estimated as:

$$Lu = \{ a_1, \dots, a_l \}, \quad a_i \in A \quad (5)$$

$$\text{and} \quad \sum_{i=1}^{l-1} P_{a_i} < P_u \quad (6)$$

$$\sum_{i=1}^l P_{a_i} \geq P_u \quad (7)$$

$$P_{a_i} \geq P_{a_{i+1}} \quad i=1, l-1 \quad (8)$$

The above relations mean that the set (Lu) is a subset of the most probable labels of (A), defined from the threshold Pu.

With set (B), the final set of possible labels for unknown words is obtained from the relation:

$$Lu' = Lu - B \quad (9)$$

The 4 numeric parameters of the system, mp, mu, Pz and Pu have a significant influence on its performance.

4. PERFORMANCE OF THE SYSTEM

The described system was implemented in C-language on a micro-Vax II computer. Its performance was tested on a Greek newspaper text of 120.000 words which was previously labelled by analysts. This text was split into 12 equal parts. The first part was used to create the initial parameters of the model and the remaining parts for the automatic labelling and parameter updating. After each labelling procedure, automatic correction was performed and the correctly labelled (known and unknown) words were measured. The corrected text was used by the processor to update the learning parameters which were used by the system to label the next text.

A set of 120 grammatical categories was used to label the texts. Table 1, gives the numeric parameters of the labelling processor which were kept constant in the successive labelling runs. Table 2, gives the number of unknown

words in each text part, which were added to the dictionary (L) after completing the labelling of the corresponding part.

mp	mu	Pz	Pu
10	4	0	0.95

Table 1: Numeric parameters of the labelling processor.

2223	1316	854	756	647	509
1	2	3	4	5	6

909	1899	1665	1515	1030	WORDS
7	8	9	10	11	TEXT PART

Table 1: Number of new words added to the dictionary (L).

The graphs in fig. 2 to 8 present the results of labelling with the four different classes of the Markovian model. In all cases, forward transition probabilities were used.

As shown in fig. 2 to 4 the performance of the system is generally increased with the class of the model. Also increased is the performance of all classes from one text part to the next, when the texts have a similar syntactic structure. Thus, the rate of correctly labelled words (fig.2) starts with appr. 86% for all classes when a learning text of 10.000 words is used and reaches rates of 92% to 94% for learning texts of 50.000 words. A similar increase from 30% to a range between 57% and 63% shows fig. 3 of correctly labelled words which have more than one label (ambiguous words).

A steep decrease of the correctly labelled words and ambiguous words in the 6th and 7th text part is due to a different syntactic structure of the texts from these parts on. The existing training of the system proved to be inefficient to handle the new structure. Nevertheless, after updating the parameters of the system with the data extracted from these parts, an increase of the performance has been shown again.

The rate of correctly labelled unknown words (fig.4) is increased also with the class of the model, but seems to decrease as more and more labels and transitions are added to the system. In fact, there are strong fluctuations depending on the analyzed texts but the rate tends to reach a more or less stable mean value.

A picture of the increasing complexity of the system with increasing parameter updating is taken from fig. 5 to 8. The average number of possible label sequences per sentence (fig. 5), the number of transitions (fig. 6), the response time of the learning processor

(fig. 7) and the response time of the labelling processor (fig. 8) are generally increased with increased training of the system in all model classes although there are classes with a remarkably stable behaviour. In all cases but in that of fig. 5 the reported measure is increased with the class of the model.

4. CONCLUSION

The results have shown that an increase of appr. 3% to 4% of the labelling accuracy can be achieved at the expense of a much higher increase of the learning and labelling time, when the class of the Markovian model is increased from 2 to 5. If the labelling time is not crucial, this approach may be justified. Otherwise, a 2nd class model is appropriate which needs substantially less time.

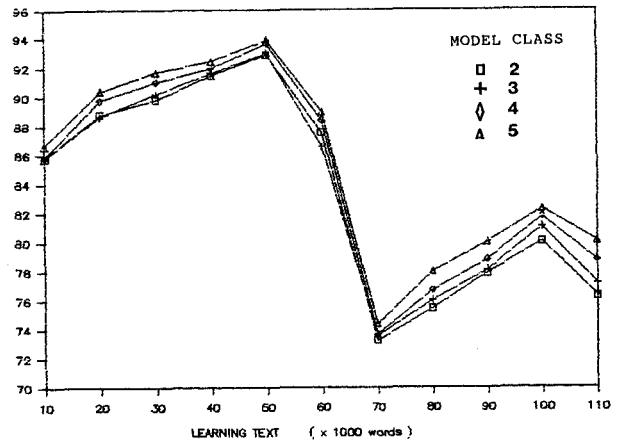


Fig.2 Correctly labelled words (%).

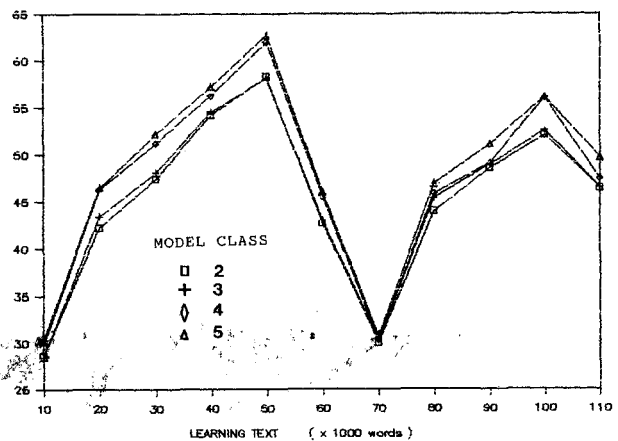


Fig.3 Correctly labelled words which have more than one labels (ambiguous words) (%).

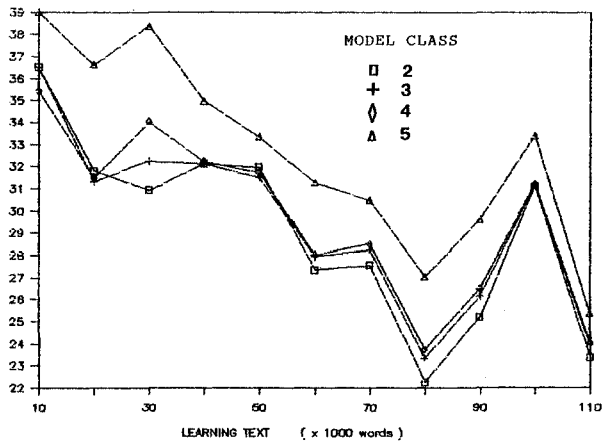


Fig.4 Correctly labelled unknown words (%).

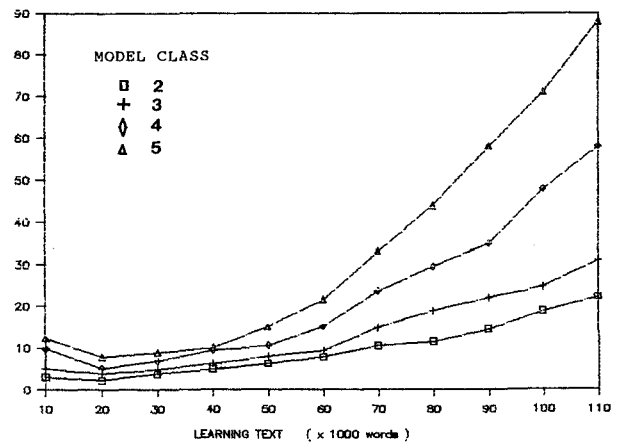


Fig.7 Response time of learning processor (x100 secs)

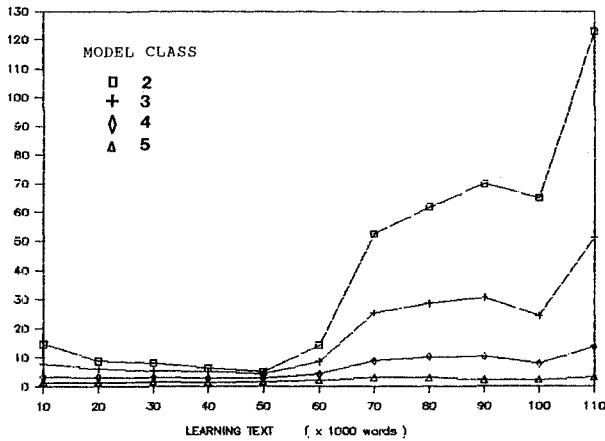


Fig.5 Average number of possible label sequences (possible solutions) per sentence.

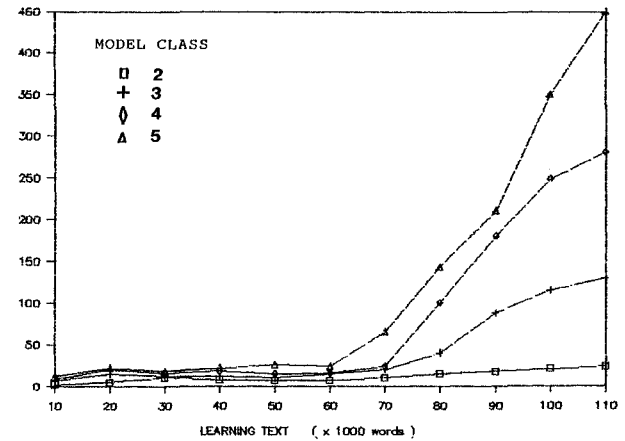


Fig.8 Response time of labelling processor (x100 secs)

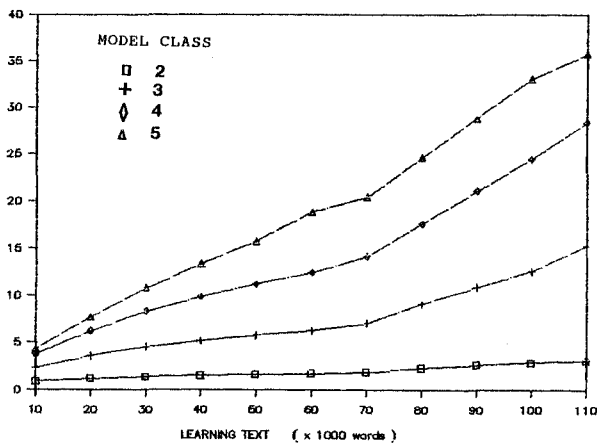


Fig.6 Total number of transitions (x1000)

REFERENCES

- [1] L. BOVES, M. REFICE, "The Linguistic Processor in a Multi-lingual Text-to-Speech and Speech-To-Text Conversion System", European Conference on Speech Technology, vol. I, p.385-388, Edinburg 1987.
- [2] E. DERMATAS, G. KOKKINAKIS, "Semi Automatic Labelling of Greek Texts", 7th FASE Symposium SPEECH'88, Book 1, p.239-245, Edinburg 1988.