AN ASYMMETRIC STOCHASTIC LANGUAGE MODEL BASED ON MULTI-TAGGED WORDS

J. Pastor, J. Colás, R. San-Segundo, J.M. Pardo

Grupo de Tecnología del Habla - Departamento de Ingeniería Electrónica

E.T.S.I. Telecomunicación - Universidad Politécnica de Madrid

Ciudad Universitaria s/n, 28040 Madrid, Spain

ABSTRACT

Tag definition in stochastic language models (n-grams and n-pos) is based on grouping together words with similar right and left context behavior.

In [2] a modification of the n-gram model for French using multi-tagged words and unsupervised clustering was introduced. Their corpus was millions of non-tagged words.

We present a variation of bi-pos language model where two tag sets are defined and assigned to each word (multi-tagged model) using grammatical information. Each tag set is based on different context behavior.

We use linguistic expert knowledge and a simple automatic clustering procedure to obtain groups of words with similar left context behavior (first set of tags) and with similar right context (second set of tags).

We propose a grammatical based model useful when no big text corpus is available and a performance increase has been observed when multi–tagged words are used because of its better adaptation to the language.

1.- INTRODUCTION

Some words in a language have similar right context but very different left context and vice versa. If only one tag set is defined and each word has only one tag, part of this information is missed. When two different tag sets are defined and two tags are assigned for each word the language model improve because more detailed transitions are modeled.

This idea was introduced in [2] for bi-gram models using full-unsupervised clustering methods. We apply grammar knowledge in the clustering procedure and class definition.

We first present the normal symmetric bi-pos model and afterwards the improved asymmetric one.

2.- SYMMETRIC BI-POS MODEL

2.1 General concept

A language model tries to estimate the probability P(W) of the sentence the speaker is going to produce.

$$P(W) = P(w_1, w_2, ..., w_n) =$$
$$= P(w_1)P(w_2 / w_1)P(w_3 / w_1, w_2)...P(w_n / w_1, w_2 ...w_{n-1})$$

where w_i is the word produced in instant i.

N-gram model reduces system complexity assuming that word probability is conditioned by only the N-1 previous words.

$$P(W) = P(w_1, w_2, ..., w_n) =$$
$$= P(w_1)P(w_2 / w_1)P(w_3 / w_1, w_2)...P(w_n / w_{n-N+1}, w_{n-N+2}...w_{n-1})$$

Due to the great number of words in the dictionary, words with similar context behavior are grouped together by manual or automatic methods. When groups are based on grammatical information, they are named parts of speech (POS) and the model is called n-pos.

For a bi-gram (bi-pos) language model, formulas are as follows:

$$P(W) = P(w_1, w_2, ..., w_n) =$$
$$= P(w_1)P(w_2 / w_1)P(w_3 / w_2)...P(w_n / w_{n-1})$$

If N_C different groups (C^1,C^2, ..., C^{NC}) of words are defined where each word can only be member of one group (and is tagged with its identification),

$$P(w_i / w_{i-1}) = \sum_{j=1}^{N_C} P(w_i / C_i^j) P(C_i^j / w_{i-1}) =$$
$$= \sum_{j=1}^{N_C} P(w_i / C_i^j) \sum_{k=1}^{N_C} P(C_i^j / C_{i-1}^k) P(C_{i-1}^k / w_{i-1})$$

to estimate these probabilities, it is necessary tagged test (each word one tag referring its group) that generate a tagged frequency dictionary and a group pair probability matrix.

The dictionary contains the number of occurrences of each word with its tag $N(w_i,C_i)$ and the matrix, the number of times a group C_i^j has follow the group C_{i-1}^k $N(C_i^j , C_{i-1}^k)$ (for all possible j and k).

$$P(w_i / C_i^j) = \frac{N(w_i, C_i^j)}{N(C_i^j)} \qquad P(C_{i-1}^k / w_{i-1}) = \frac{N(w_{i-1}, C_{i-1}^k)}{N(w_{i-1})}$$

$$P(C_i^j / C_{i-1}^k) = \frac{N(C_i^j, C_{i-1}^k)}{N(C_{i-1}^k)}$$

The matrix has the same number of rows than columns because each word has only one tag and the matrix estimate the probability a tag follows another tag. C_{i-1} is the y-axis index and C_i is the x-axis index of the matrix.

2.2 POS definition

Currently we have a bi-pos language model working in DIVO, a dictation machine developed in our group [4] [5] [1]. This language model is based on 74 parts of speech defined as follows:

Our corpus was manually tagged with a group of many very accurate tags what we call 'simple' tags. The definition of this tags where done in [11] [12] for many european languages. 'Simple' tags have a very precise class and subclass information of each kind of word and its aspects (gender, number, time of verbs, etc.).

As a result of Esprit/860 project, a 160 tag set was defined [10] [13] [14] and was said to be the best Spanish tag set. It was obtained manually by linguistic experts analyzing also some correlation figures between tags. The problem of this model was its coverage, most of the tag pairs did not have any occurrence.

From these 160 tags, we did an unsupervised clustering based on the similarity of rows and columns in the pair matrix. As criteria function to measure the distance between two groups we used the sum of the euclidean distance of its rows and columns.

$$D_y(C^j,C^k)=\left(\sum_{x=1}^{N_C}\left(\frac{N(C_i^x,C_{i-1}^j)}{N(C_{i-1}^j)}-\frac{N(C_i^x,C_{i-1}^k)}{N(C_{i-1}^k)}\right)^2\right)^{1/2}$$

$$D_x(C^j,C^k)=\left(\sum_{y=1}^{N_C}\left(\frac{N(C_i^j,C_{i-1}^y)}{N(C_i^j)}-\frac{N(C_i^j,C_{i-1}^y)}{N(C_i^j)}\right)^2\right)^{1/2}$$

$$D\ (C^j,C^k)=D_x(C^j,C^k)+D_y(C^j,C^k)$$

The automatically generated groups where analyzed by experts who reorganized some groups and decided what to do with tags with very few occurrences in the training test, obtaining finally a set of 74 tags.

Another 51 tag set were obtained by only unsupervised clustering for comparison purposes between the symmetric and asymmetric models.

2.3 Weakness of the model

When analyzing the clustering procedure used, we realized that two similar rows (low D_y) have the same right context behavior (in relation with the following category) and two similar columns (low D_x) means the same left context behavior (in relation with the preceded category).

The distance used for clustering (D) groups together tags with similar right and left context behavior (low D_x+D_y). This means that we did not join tags with very low D_x but very high D_y and vice versa.

3.- ASYMMETRIC BI-POS MODEL

We call the model explained before a symmetric model because the pair matrix is squared (each word has only one tag). The same tag is used when a word is used to predict the next one and when a word has to be predicted.

The simplest asymmetric model uses two categories per word and a non-squared matrix.

3.1 Justification

We developed the language model and the tag set definitions for Spanish but we are going to explain the concept by an English example. Look at the two following sentences:

I know that car is broken

I know those cars are broken

We can observe the dependence in number among the adjective *that/those*, the noun *car/cars* and the verb *is/are*. We are using a bi-pos model so only two words relationships can be modeled.

We can see that the words *those* and *that* have the same a priory probability of following the word *know* (the same left context). We could introduce both words in the same group and estimate the probability a demonstrative adjective has when follows a verb independently of the words number. But number information is very important to guess the following of *that* and *those* (different right context).

If we are working with a symmetric language model we have to decide if we want to include both words (*those* and *that*) in the same tag (less complexity and more coverage) or separate them in different tags.

The problem is solved if we assign two tags to each word related with its left and right context separately. For instance, when we have to decide the word that follows *know*, *that* and *those* should be in the same group (are probably equal), but when they are used to decide its following word, a separation in number is very important (each word in different groups).

Between the noun (*car*/*cars*) and the verb (*is*/*are*) it happens the opposite, it is important the noun number to choose the verb but irrelevant the verb number to choose its next word (*broken*).

In Spanish the best example can be seen between articles and nouns (or adjectives). The word before an article is number and gender independent but its following word is fully gender and number dependent. All articles should have the same left context tag but different right context tag.

3.2 Formulation

In the asymmetric bi-pos model each word has two tags: C^- when the word is known and used to guess the next one; and C^+ when the word is being guessed by its previous one. The bi-pos model is defined as follows:

$$P(w_i / w_{i-1}) = \sum_{j=1}^{N_C^+} P(w_i / C_i^{j+}) P(C_i^{j+} / w_{i-1}) =$$

$$= \sum_{j=1}^{N_C^+} P(w_i / C_i^{j+}) \sum_{k=1}^{N_C^-} P(C_i^{j+} / C_{i-1}^{k-}) P(C_{i-1}^{k-} / w_{i-1})$$

Where N_C^+ and N_C^- are the tag number of both tag sets and the probabilities are estimated as follows:

$$P(w_i / C_i^{j+}) = \frac{N(w_i, C_i^{j+})}{N(C_i^{j+})} \quad P(C_{i-1}^{k-} / w_{i-1}) = \frac{N(w_{i-1}, C_{i-1}^{k-})}{N(w_{i-1})}$$

$$P(C_i^{j+} / C_{i-1}^{k-}) = \frac{N(C_i^{j+}, C_{i-1}^{k-})}{N(C_{i-1}^{k-})}$$

To evaluate these probabilities it is necessary a bi-tagged dictionary with word frequencies, and a non squared pair matrix (of N_C^- x N_C^+ dimensions) with times C^{j+} tagged word follow C^{k-} tagged words in the training corpus (for all possible j and k).

When a word is being guessed C^{k-} tag is used and when the same word is utilized to guess the next one C^{j+} is applied.

3.3 POS definition

As said above, a 74 symmetric tag set was obtained by semi-automatic methods from a previous 160 symmetric tag set, and another 51 symmetric tag one by automatic methods.

The same procedure was used to obtain a 46 x 40 asymmetric tag set by semi-automatic methods and a 51 x 51 asymmetric set by fully unsupervised clustering.

To obtain the C^- final tag set, similar rows of the original 160x160 symmetric square matrix were joined together. This means that similar left context words were grouped together (D_y was the similarity measure). To obtain the C+ final tag set, D_x was used to measure the similarity between columns and words with similar right context were grouped.

4.- EXPERIMENTS

4.1 Test set perplexity

In order to compare both language models, we use the LogProb introduced by Jelinek [15] to evaluate his n-gram language model based on the Shannon Information Theory [16]. This measure is also called Test Set Perplexity [6] [7] [8].

$$LP = -\frac{1}{L} \sum_{i=1}^{L} \log_2 P(w_i / w_{i-1})$$

$$PP = 2^{LP}$$

L is the test corpus word number and $P(w_i/w_{i-1})$ is estimated for each word pair of the training text using formulas seen above.

As information theory says [9], a task with perplexity PP has the same complexity as another task with PP words all with equal probability. Therefore, if different language models are compared, the best one has the less perplexity measure.

4.2 Corpus

To validate the language model we used a training corpus of 165.000 tagged words and a test corpus of 5.500, both tagged with ‘simple’ tags. The corpus was re-tagged to form five different corpus.

4.3 Definition of tags for experiments

The five different corpus are related with five different clustering procedures and tag sets:

Sym160: symmetric tagged corpus with 160 different tags obtained by expert knowledge from ‘simple’ tags.

Sym74: symmetric tagged model with 74 different tags obtained from *Sym160* tags by unsupervised clustering followed by expert knowledge.

Sym51: symmetric tagged model with 51 different tags obtained from *Sym160* tags by only unsupervised clustering.

Asym51.51: asymmetric tagged model with two sets of 51 different tags obtained from *Sym160* tags by only unsupervised clustering.

Asym46.40: asymmetric tagged model with two sets of 46 and 40 different tags obtained from *Sym160* tags by unsupervised clustering followed by expert knowledge.

The clustering used in *Sym74* and *Sym51* is based on left and right similarity. In *Asym51.51* and *Asym46.40* a double clustering was done to obtain the double tag set, one based on left word context and the other on right word context.

4.4 Results

The test set perplexity was used to compare the five different language models explained above with the following results:

Language model	Perplexity
Sym160	322
Sym74	330
Sym51	401
Asym51.51	361
Asym46.40	350

Table 1: Perplexity measured for the five different language models.

The model with less perplexity is *Sym160* because the 160 symmetric tags are the base for the other models so it is the more accurate but the most complex and the most difficult to train.

Only *Sym51* and *Sym51.51* are comparable because both are obtained by unsupervised clustering (one symmetric and the other asymmetric) and have the same complexity (number of tags). *Asym51.51* has less perplexity than *Sym51* what shows that an asymmetric bi-pos model is better than a symmetric one obtained by the same method and with the same complexity.

If we multiply the perplexity measure by the number of rows and columns of matrices of each model, we have a measure that combines the complexity and the quality of the model.

The following table shows that *Asym46.40* has the best rate. This model has the least complexity and less perplexity than *Asym51.51* model. This shows that a mixture of unsupervised clustering and expert knowledge is the best solution for this kind on language models.

Language model	**Perplexity * #rows * #columns**
Sym160	8.243.200
Sym74	1.807.080
Sym51	1.043.001
Asym51.51	938.961
Asym46.40	644.000

Table 2: Perplexity * #rows * #columns measure.

5.- CONCLUSION

We have shown that multi-tagged word models has better perplexity results than simple-tagged ones because of its better adaptation to the language relations between words.

Supervised and unsupervised models have been compared with lower perplexity results for supervised tag definition. This method requires linguistic expert supervision of the automatic clustering results.

This technique is highly recommended when a big corpus is not available.

6.- REFERENCES

1. M.A. Leandro, A. Villegas and J.M. Pardo. "Efficient Isolated Word Recognition In Spanish Based On Static Modeling". *Eurospeech 95, vol 1, pp. 71-74. 1995*

2. M. Jardino. “A Class Bigram Model for Very Large Corpus”, *ICSLP’94 p.867-870. 1994.*

3. J. Ueberla "Analysing a simple language model: some general conclusions for language models for speech recognition" *Computer Speech and Language. Volumen 8, Number 2 April 94.*

4. J. Macias-Guarasa et al. "On the Development of a Dictation Machine for Spanish: DIVO". *ICSLP 94, S22-26. pp. 1343-1346. 1994.*

5. M.A. Leandro and J.M. Pardo. "Low Cost Speaker Dependent Isolated Word Speech Preselection System Using Static Phoneme Pattern Recognition". *Eurospeech 93, vol. 1, pp. 117-120. 1993.*

6. P. Witschel. “Constructing Linguistic Oriented Language Models for Large Vocabulary Speech Recognition*”. Eurospeech’93, p. 1199-1202. 1993.*

7. M.Jardino, G. Adda, “Automatic Word Classification Using Simulated Annealing”, *Proc. ICASSP’93, Vol. 2, p. 41- 44, 1993.*

8. M.Jardino, G. Adda, “Language Modelling for CSR of Large Corpus Using Automatic Classification of Words”, *Eurospeech’93. Pp. 1191-1194. 1993.*

9. T.M. Cover, J.A. Thomas. “Elements of Information Theory”. Wiley series in telecommunications. Cap 2.

10. [BU-WKL-0376] BU-Kugler, Unification of the word classes of the ESPRIT project 860, 22-02-89. *ESPIRIT/860 Project Documentation.*

11. [UN-CAT0588] UN-Casado, Rules for Spanish text categorization, 01-05-88. *ESPIRIT/860 Project Documentation.*

12. [UN-CAT1087] UN, Abbreviations for grammatical categories of the Spanish language, 22-07-87. *ESPIRIT/860 Project Documentation.*

13. [UN-CS0588] UN-Enríquez/Casado, Spanish ‘Optimal Set’ of Cover Symbols, 17-05-88. *ESPIRIT/860 Project Documentation.*

14. [UN-WCS20288] UN-Casado/Enríquez, Changes made to the Spanish Word-class system, 21-02-88. *ESPIRIT/860 Project Documentation.*

15. F. Jelinek, “The Development of an Experimental Discrete Dictation Recognizer”, *Proc. of the IEEE, Vol. 73, Nº 11, Nov. 1985.*

16. C. E. Shannon, “A Mathematical Theory of Communication”, *Bell Syst. Tech. J., vol. 27, 1948, pp. 379-423, 623,657.*