ISSN 1870-4069

Hidden Markov Framework for Lexical Tagging

C. Razo Hernández[1], J. M. Benedí[2]

[1] Universidad de Guanajuato,
Facultad de Ingeniería Mecánica, Eléctrica y Electrónica,
Mexico

[2] Universidad Politécnica de Valencia,
Departamento de Sistemas Informáticos y Computación,
España

jbenedi@dsic.upv.es

Abstract. In this work we present a lexical tagger for English using hidden Markov models with a probabilistic model of distribution of words in categories. The corpus used and set of labels corresponding to the categories is described. After, the model use is described and the form which it is estimated. Finally, we show the realized experiments and the obtained results.

Keywords: Tagger, POS tags, HMM, NLP.

1 Lexical Tagged

Tagged is the assignation of category to which the words in a corpus belong (POST = Part-Of-Speech Tagging). Its purpose is help to improve applications of the natural language processing in which its required to know the sense of the words, automatic translation, information retrieval, text classification and extraction of information among others.

The corpus used in the experiments realized in this work is the part of the Wall Street Journal that has been processed in the project Penn Treebank. It contains approximately a million words distributed in 25 directories. It was automatically labeled, analyzed and manually reviewed. The size of the vocabulary is greater to 49,000 words and the set of POS tags is 45 tags. For the experiments, the corpus is divided in training corpus (directory 00-20), tuning corpus (directory 21-22) and test corpus (directory 23-24).

2 Proposed Model

The used model, combines a probabilistic model of distribution of words in categories with hidden Markov models. For the probabilistic model, a list of words is used in which appears each word and the frequency of assignation of category (Cw). The hidden Markov model used is a model of the left to right.

In order to find the set of lexical labels corresponding to a sequence of words, the sequence of observable symbols is calculated (labels) that will be emitted by the most

probable states sequence. This calculation is carried out using HMM and the distribution word-category as in the equation (1):

$$\bar{C} = argmax_{s.e.} \prod b_1(e_1)\,\Pr(w_1|e_1)a_{12}b_2(e_2)\,Pr(w_2|e_2)a_{23}\ldots a_{n\ 1,n}\,b_n(e_n)Pr(w_n|e_n). \quad (1)$$

$Pr(w_i|e_i)$ calculates using distribution word-category Cw and the rest with the hidden Markov model. A modification to the Viterbi algorithm that allowed us to calculate the expression (1) and to obtain the corresponding set of labels [1].

The estimation of the Markov model and the model based on categories is made separately by simplicity. The hidden Markov model is training with a corpus tagged. A detailed description of the methods of training for HMM can be found in [3].

The parameters of the distribution word-category, C_w = Pr(w|e), calculates agreement with the equation [2]:

$$Pr(w|e) = \frac{N(w,e)}{\sum N(w',e)}, \quad (2)$$

where N(w, e) is the number of times that the word w has been tagged with the POStag e and $\sum_w N(w',e)$ is the sum of all the words that have been tagged with that POStag.

The word w can belong to different categories. It can also occur, that in the training set does not appear a word and therefore its probability $Pr(w|e)$ is not defined. This is solved adding the term $Pr(UNK|e)$ to all the categories, which represents the probability for words unknown in the test set. In order to estimate this probability, three approaches are used.

The first approach consists of assigning a small probability equal to $Pr(UNK|e)$ for all the categories.

The second approach consists of the supposition of that the distribution $Pr(UNK|e)$ in the test set, is very similar to the corresponding one in a tining set. For that reason, the distribution of the frequency of appearance of the words unknown tagged with $Pr(UNK|e)$ in this set of tuning is considered. For the possible null values of the considered distribution, a very small probability is added. These first two approaches are described in [1].

The third approach, is based on a study of Demartas and Kokkinakis [4], they concludes that the probability distribution of the unknown words is very similar to the one of which they frequently is equal to 1 and very different from the distribution of the well known words. Based on this, the estimation becomes by means of the equation (3):

$$Pr(w|e) = \frac{Pr(e|w^{unkonwn})P(w^{unkonwn})}{P(e_i)}. \quad (3)$$

$Pr(e|w^{unkonwn})$ and $P(e_i)$ are calculated from the training set and $P(w^{unkonwn})$ it is calculated from the tunning set. For the null values of the considered distribution, a very small probability is added.

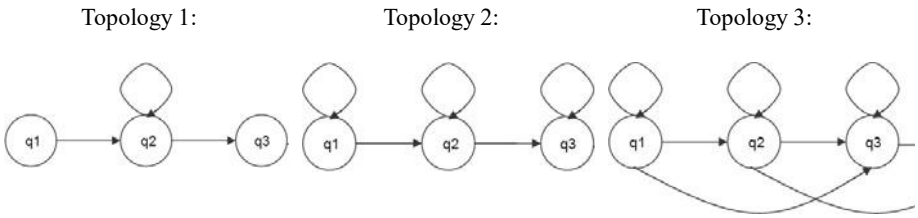

Fig. 1. Topologies used in the HMM.

Table 1. Baseline for both used.vocabularies.

Size Vocab.	Precision	Exactitude
1000	80.46%	80.46%
37075	85.63%	85.63%

Table 2. Result of the best model.

Model	Precision	Exactitude
C_w	80.46%	80.46%
□,C_w	87.50%	87.21%

Table 3. Results for test corpus.

Model	Precision	Exactitude
C_w	81.12%	81.12%
□,C_w	87.25%	86.79%

Table 4. Better result for vocabulary complete.

Model	Precision	Exactitude
C_w	85.63%	85.63%
□,C_w	91.69%	91.69%

Table 5. Results for test corpus with complete vocabulary.

Model	Precision	Exactitude
C_w	84.09%	84.09%
□,C_w	90.02%	89.13%

3 Results

In order to find the best model, experiments were made using different topologies and number of states in the hidden Markov model. In addition two sizes of different vocabularies were used considering the three approaches for unknown words.

Like baseline, the so large distribution was made tagged of corpus of test using only word-category for of vocabularies both used in table 1.

1. Experiments on topology and states:

Experiments were done changing the number of states in the HMM and using 3 different topologies.

2. Experiments on approaches of unknown words:

Considering the 3 topologies, different number of states and using the three approaches for unknown words, the best result is obtained using topology 1, 8 states in the HMM and using approach 3 for not known words as it is in table 2. Using this model, test corpus was tagged; the obtained results we show in table 3.

3. Experiments with complete vocabulary:

Using a vocabulary of 37,075 words, the results improve really. The best model is obtained with topology 3 and 3 states for the HMM and using approach 3 for unknown words, the results are in table 4.

4 Conclusions

The obtained results are good comparing with baseline. The precision was 81.12% like baseline, value that our model improves up to 87.25% for the vocabulary of 1000 words. For the vocabulary of 37,075 words, baseline is 84,09% and our model it improvement up to 90.02%.

References

1. Sánchez, J.A., Nevado, F., Benedí, J.M.: Lexical decoding based on the combination of category-based stochastic models and word-category distribution models. In: Proceeding of COLING (2006)
2. Sánchez, J.A., Benedí, J.M.: Combination of n-grams and stochastic context-free grammars for languaje modelling. In: Proceeding of COLING (2000)
3. Juang, B.H., Rabiner, L.: Fundamentals of Speech Recognition. Prentice-Hall (1993)
4. Molina, A.: Desambiguación en procesamiento del lenguaje natural mediante técnicas de aprendizaje automático. Technical report, Universidad Politécnica de Valencia (2004)