A Stochastic Language Model for Automatic Generation of Arabic Sentences

Si Lhoussain Aouragh [1], Jelloul Allal [2] and Abdellah Yousfi [3]

[1,2] Département de Mathématiques et informatique, Université Mohamed premier, Oujda, Maroc,
[3] Institut d'études et de recherches pour l'arabisation, Rabat, Maroc
[1] aouragh@hotmail.com, [2] Allal@sciences.univ-oujda.ac.ma, [3] yousfi240ma@yahoo.fr

Abstract

Language modeling aims to summarize general knowledge related in natural language. To this aim, the automatic generation of sentence is an important operation in the automatic language processing. It can serve as the basic for such various applications such as automatic translation, continuous speech recognition.

In this article, we present a stochastic model that allows us to measure the probability of generating a sentence in Arabic from a set of words. This model is based on the fact that a sentence is based on syntax and semantic level that are independent, and that allows us to model each level with the appropriate model. The estimation of the parameters of this model is made on a corpus of training labeled manually by the syntactic labels.

Keywords: *Training corpus, tagging syntactic, generation of sentences, p-context model, Hidden Markov model, Arabic.*

1. Introduction

The most used of the language models are of nature probabilistic . The n-gram model [1] that originated in information theory [2] remains the base for many other languages models, such as models based on decision trees [3] or models of structured language [4] or models n-multigrams [5] or models with hidden memory [6]. One of the disadvantages of these models is the large number of parameters that needs to be estimated. In additions, these models require large training set and well selected as to cover all the events of successive words.

These models are still in main use in continuous speech recognition. In this article, we have developed a stochastic model that allows us to calculate the probability of generating automatically a sentence from a set of words in Arabic language. This model combines two levels:

- Language models derived from models n-gram, that we call p-context models. The advantage of this model is that the number of parameters to be estimated is lower than those of n-gram models. In addition, it does not take into consideration the order of words in the sentence. We have used this model to model the semantic level.
- Syntax model: it allows us to manager the order in the sentence. It based on the calculation of the optimal path of syntax labels $s_{i_1}, s_{i_2}, ..., s_{i_n}$ of the words $w_{i_1}, w_{i_2}, ..., w_{i_n}$

2. Probabilistic Model for Sentence Generation

A sentence can be viewed as a linguistic element made of two levels: a syntax level and a semantic level. In our approach, we assume that these two levels are independent (which allows us to treat each level independently of the other).

To generate a sentence $w_{i_1} w_{i_2} .... w_{i_n}$, we have modeled these two levels by using the following two conditions :

i. The first condition is that each word w_{i_j} $j \in \{1,...,n\}$ should appear in a context of size p $(1 \le p \le n)$ with all the remaining words, that is:

$$\Pr\left(w_{i_j}, w_{i_{j_1}}, ..., w_{i_{j_p}} \text{ are in the same context}\right) \neq 0$$

We note later that this probability $\Pr\left(w_{i_j}, w_{i_{j_1}}, ..., w_{i_{j_p}} / context\right) \neq 0$

For each j and for each $j_1, ..., j_p \in \{1,.., j-1, j+1,.., n\}$

Is equivalent to:

$$\prod_{j=1}^{n}\prod_{j_1=j+1}^{n}\prod_{j_2=j_1+1}^{n}....\prod_{j_p=j_{p-1}+1}^{n}\Pr\left(w_{i_j},w_{i_{j_1}},...,w_{i_{j_p}}/context\right)\neq 0 \qquad (1)$$

We call this model: p-context model.

- For p = 1 : the model is called a bi-context model, and the formula (1) reduces to:

$$\prod_{j=1}^{n}\prod_{k=j+1}^{n}\Pr\left(w_{i_j},w_{i_k}/context\right)=\prod_{j=1}^{n}\prod_{k=j+1}^{n}l_{jk}\neq 0$$

where : $l_{jk}=\Pr\left(w_{i_j},w_{i_k}/context\right)$

- For p = 2 : the model is called tri-context model, and the formula (1) becomes :

$$\prod_{j=1}^{n}\prod_{k=j+1}^{n}\prod_{l=k+1}^{n}\Pr\left(w_{i_j},w_{i_k},w_{i_l}/context\right)\neq 0$$

- For p = n − 1 : the formula (1) becomes :

$$\Pr\left(w_{i_j},w_{i_{j_1}},...,w_{i_{j_{n-1}}}/context\right)\neq 0$$

ii. The second condition allows us to verfiy whether the order of words $w_{i_1},w_{i_2},...,w_{i_n}$ is correct. This is achieved based on grammatical knowledge of the words $w_{i_1},w_{i_2},...,w_{i_n}$. We have modeled this condition as a problem of the existence of optimal path $s^*_{i_1},s^*_{i_2},...,s^*_{i_n}$ of syntaxic labels of words like $w_{i_1},w_{i_2},...,w_{i_n}$:

$$\Pr\left(w_{i_1},...,w_{i_n},s^*_{i_1},...,s^*_{i_n}\right)\neq 0 \qquad (2)$$

The probability of generating a sentence $w_{i_1}w_{i_2}...w_{i_n}$ is then the product of these two probability (1) and (2) (As noted earlier, we assumed that the syntax and semantic levels are independent).

$$\Pr\left(w_{i_1},..,w_{i_n}\right)=\beta^*_{d_n}\prod_{j=1}^{n}\prod_{j_1=j+1}^{n}\prod_{j_2=j_1+1}^{n}....\prod_{j_p=j_{p-1}+1}^{n}\Pr\left(w_{i_j},w_{i_{j_1}},...,w_{i_{j_p}}/context\right) \qquad (3)$$
$$\text{x}\,\Pr\left(w_{i_1},...,w_{i_n},s^*_{d_1},...,s^*_{d_n}\right)$$

Where : $\beta^*_{d_n}=\Pr\left(s_{d_n}\text{ is final state}\right)$

Remark

We have added the probability of a final state to avoid the problem of generate incomplete sentences. If we take for example the sentence *"□□□ □□□□□ □□□ "*, the probability of generating this sentence without taking into consideration $\beta^*_{d_n}$ is not null because we have the sentence *" □□□ □□□□□ □□□ □□□□□□□"* in the training sample.

3. Application

As an application of this model, we took the case of p=1 (bi-context model). In this case, the probability of generating $w_{i_1}w_{i_2}...w_{i_n}$ is :

$$\Pr\left(w_{i_1},..,w_{i_n}\right)=\beta^*_{d_n}\Pr\left(w_{i_1},...,w_{i_n},s^*_{d_1},...,s^*_{d_n}\right)\prod_{j=1}^{n}\prod_{k=j+1}^{n}l_{jk} \qquad (4)$$

$s^*_{d_1},s^*_{d_2},...,s^*_{d_n}$: the optimal path of syntaxic labels associated with the sentence $w_{i_1}w_{i_2}...w_{i_n}$ is given by:

$$s^*_{d_1},s^*_{d_2},...,s^*_{d_n}=\arg\max_{s_{j1},..,s_{jn}}\Pr\left(w_{i_1},..,w_{i_n},s_{j_1},..,s_{j_n}\right) \qquad (5)$$

We use hidden Markov models [7] as in to solve equation (5).

We assume that the double process (X_t, Y_t) is Hidden Markov Model (HMM) of order one, satisfying the following conditions:

- $X_t = s_t$: is a Markov chain of order 1 with value is in the set of syntaxic labels $E = \{s_1,..., s_N\}$, and X_t satisfies:
  - $\Pr(X_{t+1} = s_j / X_1 = s_{i_1},..., X_t = s_i) = \Pr(X_{t+1} = s_j / X_t = s_i) = a_{ij}$
  - $\Pr(X_1 = s_i) = \pi_i \quad i \in \{1,..., N\}$
- $Y_t = w_i$ is a processes with value is in the set of words $V = \{w_1,..., w_M\}$ representing the vocabulary of our system, Y_t satisfies:

  $\Pr(Y_t = w_t / X_1 = s_{i_1},..., X_t = s_j, Y_{t-1} = w_{t-1},..., Y_1 = w_1) = \Pr(Y_t = w_t / X_t = s_j) = b_j(w_t) = b_{jt}$

  b_{jt} : is the probability that word w_t has label s_j.

Remark

Our model of sentence generation is entirely defined in a vector of prameters denoted by $\Theta = (\Pi, \beta, A, B, L)$

$\Pi = \{\pi_1,..., \pi_N\}$ is the set of the initial probabilities.

$\beta = \{\beta_1,..., \beta_N\}$ is the set of the final probabilities.

$A = (a_{ij})_{1 \le i, j \le N}$ the matrix of transiations probabilities.

$B = (b_{it})_{\substack{1 \le i \le N \\ 1 \le t \le M}}$ the matrix of probabilities that the word w_i has label s_i.

$L = (l_{ij})_{1 \le i, j \le M}$ is the matrix of probabilities that a word w_i appears in the same context as the word w_j.

3.1. The clacul of the optimal Path

We use the Viterbi algorithm to derive the optimal path.

We define $\delta_t(s_k)$ as the probability of the best path of reaching label s_k at time t

$$\delta_t(s_k) = \max_{s_{d_1},...,s_{d_t}} \Pr(w_{i_1},.., w_{i_t}, s_{d_1},.., s_{d_{t-1}}, s_k)$$

Bayes rule leads the following recursive formula:

$$\delta_t(s_k) = \max_{s_j} [\delta_{t-1}(s_j) a_{jk} b_k(w_t)] \quad \forall t \in \{i_1,..., i_M\} et \forall k \in \{j_1,..., j_N\} \qquad (6)$$

The optimal path is then obtained using the recursive calculation defined in formula (6).

3.2. Training

In general, three methods of estimating model parameters can be used: Maximum Likelihood estimation [8], Maximum a Posteriori Estimation [9] or estimation using maximum mutual information as in [3].

In our case, we used the maximum likelihood estimation. Let $R = \{ph_1,..., ph_K\}$ a set of arabic sentence, labeled with a set of syntax labels $E = \{s_1,..., s_N\}$, the estimation of Θ is given by:

$$\Theta^* = \arg\left[\max_{\Theta} \prod_{i=1}^{K} \Pr_{\Theta}(ph_i)\right] \qquad (7)$$

Solving this maximization problem leads to the following estimates:

1. $\pi_i = \dfrac{\sum_{j=1}^{K} \delta(s_i \text{ is a initial state in } ph_j)}{\mathrm{K}}$
2. $\beta_i = \dfrac{\sum_{j=1}^{K} \delta(s_i \text{ is } a \text{ final state in } ph_j)}{\mathrm{K}}$

3. $a_{ij} = \frac{\sum_{l=1}^{K} F(s_i s_j)}{\sum_{l=1}^{K} F_l(s_i)}$

4. $b_{ij} = \frac{\sum_{l=1}^{K} F_l(w_j \text{ has label } s_i)}{\sum_{l=1}^{K} F_l(s_i)}$

5. $l_{ij} = \frac{\sum_{l=1}^{K} F_l(w_i \text{ and } w_j)}{K}$

Where:

$\delta(s_i \text{ is an initial state in } pk_j) = \begin{cases} 1 \text{ if } s_i \text{ is an initial state in } ph_i \\ 0 \text{ otherwise} \end{cases}$

$F_i(\sigma)$ is the number of times where σ is in the sentence ph_i.

3.3. Experimental Results

3.3.1 Training set

We built a training set of 1449 sentences in Arabic (of different lengths), labeled with 186 labels of syntax types chosen to cover almost all the syntax events of Arabic language.

The evaluation of our model of generating sentences is done using a program written in Perl language, made of two modules:

• Training model: responsible for estimating the parameters of our model.

• Sentence generation model: responsible for generating sentence from the vocabulary in our system.

3.3.2 Experimental Results

To evaluate our model, we have generated all possible sentence made of four words and with non-null probability of generation.

The error rate in our work is defined as the percentage of incorrect sentences out of all generated sentences.

The exact error rate in these sentences is as follows:

Number of generated sentences	7426
Error rate	61,52%

We notice that the error rate in this case is very high. The majority of these errors are at the syntax level (the syntax structure of many generated sentences is extracted from the structure of phrases of length different than four words).

To remedy this problem, we introduced two approaches:

• The first approach uses only sentences of four words to training the model of generation model. Results obtained under this approach are:

Number of sentences generated	592
Error rate	7,43%

The error rate has been reduced considerably in comparison with the first case, but the number of generated sentences is reduced as well. Sentences generated in the second case stand at 8% of generated sentences in the first case.

• For the second approach, the procedure of learning of HMM is done using only samples with four words, while the training of the bi-context model is done using all the sentences in the learning sample. The results obtained are as follows:

Number of sentences generated	1193
Error rate	29.08%

We notice that the number of generated sentences has increased two times compared to the second case. The error case has been reduced (compared with the first case) with 52.73%.

The analysis of all these results shows that the majority of errors are caused by the following points:

• The order of the model we have used is one. Many errors would be eliminated with a model of order 2 or 3.

• Labels are not well specified, leading to sentence considered correct at the syntax level but incorrect at the semantic level.

4. Conclusion and Future Work

The results obtained in this work are in general encouraging (note that works in this direction are rare). To further reduce the errors of sentence generation, we plan to extend our work to p-context models. This will increase the number of parameters to be estimated, and in this case the use of the concept of classes and their choice becomes importance. It would be preferable to use classes combining the two levels of syntax and semantic for words.

Furthermore, it would be useful to do training of p-context models independently of HMM models using a larger sample size.

In theory, the parameters of sentence generation model should converge to constant value, and a large training set should indeed makes these parameters close enough to their theoretical value. This will allow us to deduce the probabilities (rate) of utilization of different sentences in Arabic languages.

REFERENCES

1. BAHL L.R., BAKER J.K., COHEN P.S., JELINEK F., LEWIS B.L., et MERCER R.L., *Recognition of a continuously read natural corpus*. In : Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.Tulsa. 1978
2. Jelinek F., *Continuous speech recognition by statistical models*. Proceedings of the IEEE, 1976.
3. BAHL P., BROWN P., DE SOUZA P. et MERCER R., *A tree-based statistical language model for natural language speech recognition*. Pages 507-514 of : A.WAIBEL et K.-F. LEE (eds), Readings in Speech Recognition. Morgan-Kaufmann, 1990.
4. CHELBA C. et JELINEK F., *Structured language modeling*. Computer, Speech and Language, 14(4), 283-332, 2000.
5. DELIGNE S. et BIMBOT F., *Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams*. In : Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Détroit, USA, 1995.
6. KUHN R. et DE MORI R., *A cache-based natural language method for speech recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(6), 570-582, 1990.
7. Yousfi A., Jihad A., *Etiquetage morpho-syntaxique*. RECITAL, Durbin, France, 6 10 Juin 2005.
8. FEDERICO M. et DE MORI R., *Language Modelling*. Chap. 7, pages 204-210 of : R. DE MORI (ed), Spoken Dialogue with Computers. Academic Press, 1998b.
9. Yannick E., *Intégration de sources de connaissances pour la modélisation stochastique du langage appliquée à la parole continue dans un contexte de dialogue oral homme-machine*. Thèse de Doctorat, Université d'Avignon, 2002

Article received: 2006-06-12