A Language Modeling Based on a Hierarchical Approach: M_n^ν

Imed ZITOUNI
LORIA / INRIA-Lorraine
Campus Scientifique B.P.239
F-54500 Vandœuvre-lès-Nancy, France
E-mail: zitouni@loria.fr
Tel: (33) 3 83 59 20 85, Fax: (33) 3 83 41 30 79

Abstract

In this work, we introduce the concept of hierarchical M_n^ν language model and we compare it to the based class multigram and interpolated class n-gram model. The originality of our approach is its capability to parse a string of class/tags into variable length dependent sequences. A few experimental tests were carried out on a class corpus extracted from the French “Le Monde” word corpus labeled automatically. In our experiments, M_n^ν outperforms based class multigram and interpolated class bigram but are comparable to the interpolated class trigram model.

1 Introduction

In the field of speech processing, as in many other domains, the efficiency of pattern recognition algorithms is highly conditioned to a proper definition of the patterns assumed to structure the data. Consequently, the set of units can either be defined explicitly, with the risk of a possible mismatch due to our lack of a priori knowledge, or can be learned from a large and representative enough set of data samples, like in data-driven approches. In fact, increasing effort is being dedicated to learn the structure of speech and language from the data itself, either at the lexical level for language modeling [6], [7], or at others levels of speech processing. In This context, we present in this paper a new approach of language modeling, M_n^ν, which is able to modelise a language by a dependent variable length sequences similar to a probabilistic finite state model. This model is computed stochastically, in an ascending way, by the use of a large and representative samples of the language. At the lowest level, we retrieve, in a corpus of text, typical variable-length sequences of words. The multigram model, presented in [2], aims at modeling these kinds of dependencies. As we move up, we consider the sequences of a lower level as forming the basic element of current level. For feasible modeling, we must specify the maximum length of a sequence, as well as the depth of the model. We denote a model having maximum length n and depth ν as M_n^ν. Using this notation, the traditional multigram [1] can be written as M_n^1.

To evaluate this model, we use the test perplexity [4] as a performance measure. If the value of the perplexity decreases, the performance and the recognition rate of the dictation machine probably increase [8].

In the following we first present the M_n^ν model (Section: 2). We give some terminology (Section: 3). After, a formulation of the model is given (Section: 4). Then, we report an evaluation of the M_n^2 model and a comparison with the class multigram model and interpolated class n-gram model [5]. Finally, we conclude and give some perspectives.

2 The M_n^ν model

Motivated by the success of class based approaches in traditional n-gram modeling and in order to cope with the sparsness of training data we want to explore their potential in our approach. To deal with the syntactic constraints in a language, we label the stream of words with 233 classes extracted from the eighth elementary grammatical classes of the French language. Then, the inter-word transition probability of the M_n^ν model is assumed to depend only on the classes.

The class n-gram model [4], which is the most used model in the speech community, assumes that the statistical dependencies between words, labelled by classes, are of fixed length n along the whole sentence. In the approach we propose, we use a hierarchical model which successively combines sequences of classes. We refer to these class sequences as "syntagmatic groups" (we hope that these extracted sequences coincide with those defined traditionaly in natural language). In this approach, a sentence is modeled by the concatenation of dependent variable-length sequences of "syntagmatic groups". These groups are obtained in the different levels, $j \in \{1 \dots \nu\}$, of the M_n^ν language model. At each level j, we apply the class n-multigram[1] model on a training corpus to extract a dictionary of class groups[2] set (sequences set) used at the upper level $(j+1)$. After tagging the training corpus with the most likely segmentation, obtained by this dictionary, the process of applying the class n-multigram model is repeated until $j = \nu$. The dictionary (set of sequences) is updated at each level and the set of sequences obtained at level ν is used to evaluate the sentence.

In figure 1 we present an example of applying the M_3^2 to the sentence: "Tunisia is a mediterranean country".

After tagging the sentence [5] (Level 0), the dictionary of level 1 is formed by a set of typical variable-length sequences of words (< 3 in our example). Level 2 contains a dictionary obtained by the set of best sequences computed on a corpus using the dictionary of level 1.

3 Terminology

Let $\mathcal{W} = (w_1, w_2, \dots, w_T)$ denote a sentence which is a sequence of T words, and $\mathcal{C} = (c_1, c_2, \dots, c_T)$ denote the description of sentence T in terms of syntactic classes. n

[1]class multigram such us the maximum number of classes in a sequence is equal to n

[2]which could be in the same cases considered as "syntagmatic groups"

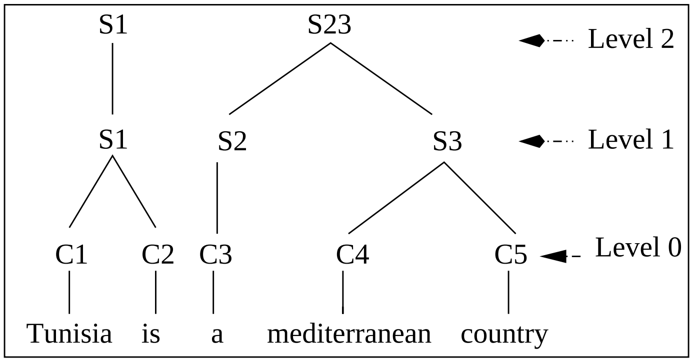

Figure 1: M_3^2 on the sentence: "*Tunisia is a mediterranean country*"

denotes the maximum length of a "syntagmatic groups" sequence, and ν the maximum depth order of the n-ary tree. C_0^i denote the number of occurences, at level i, above which a sequence of symbols is included in the initial inventory of sequences. Ω_j denotes the sequence of "syntagmatic group" or syntactic classes obtained at level j, $1 < j \leq \nu$ (depth order j). This sequence correponds to the most likely segmentation of Ω_{j-1} (recursively until Ω_1). Ω_1 corresponds to the sequence of syntactic classes $\mathcal{C}$. $|\Omega_j|$ denotes the number of units in the sequence Ω_j.

4 Formulation of the Model

Let L_j be a possible segmentation of the sequence of "syntagmatic groups" Ω_j: $s_j(1), s_j(2), \ldots, s_j(q_j)$. The dictionary of sequences, which can be formed by combining $1, 2, \ldots$ up to n symbols from Ω_j, is noted $D_{S_j} = \{s_j(i)\}$. The likelihood $\mathcal{L}(\Omega_j, L_j)$ of the sequence of "syntagmatic groups" Ω_j associated with segmentation L_j is the product of the probabilities of the successive sequences, each of them having a maximum length of n:

$$\mathcal{L}(\Omega_j, L_j) = \prod_{t=1}^{t=q_j} p(s_j(t)) \tag{1}$$

Denoting as $\{L_j\}$ the set of all possible segmentation of Ω_j into sequences of "syntagmatic groups" or syntactic classes, the likelihood of Ω_j is:

$$\mathcal{L}_{M_n^j}(\Omega_j) = \sum_{L_j \in \{L_j\}} \mathcal{L}(\Omega_j, L_j) \tag{2}$$

For the basic class n-multigram model the decision-oriented version that parses Ω_j according to the most likely segmentation is:

$$\mathcal{L}^*_{M_n^j}(\Omega_j) = \max_{L_j \in \{L_j\}} \mathcal{L}(\Omega_j, L_j) \tag{3}$$

where the most likely segmentation $L^*_{M_n^j}$ of Ω_j is :

$$L^*_{M_n^j} = \arg \mathcal{L}^*_{M_n^j}(\Omega_j) = \Omega_{j+1} \tag{4}$$

The decision-oriented version of the M$_n^\nu$ model parses $\mathcal{C}$ according to the set of most likely segmentations at each level, thus yielding the approximation:

$$\mathcal{L}^*_{M_n^\nu}(\mathcal{C}) = \mathcal{L}^*_{M_n^\nu}(\Omega_\nu) \tag{5}$$

Ω_ν is the most likely sgmentation $L^*_{M_n^{(\nu-1)}}$ obtained at level $\nu - 1$. This process is computed recursively from 1 until ν. Ω_1 denotes the sequences of syntactic classes $\mathcal{C}$. If $\nu = 1$ this model is similar to the basic class n-multigram model.

For instance, with $T = 4$, $n = 3$, $\nu = 2$, $\mathcal{C} = abcd$, and by denoting sequence borders with brackets. For $j = 1$, $\Omega_1 = abcd$:

$$\mathcal{L}^*_{M_3^1}(\Omega_1) = max \left\{ \begin{array}{l} p([a])p([bcd]) \\ p([abc])p([d]) \\ p([ab])p([cd]) \\ p([ab])p([c])p([d]) \\ p([a])p([bc])p([d]) \\ p([a])p([b])p([cd]) \\ p([a])p([b])p([c])p([d]) \end{array} \right\}$$

Assume that $\mathcal{L}^*_{M_3^1}(\Omega_1) = p([a])p([bc])p([d])$ and let X denote the sequence $[bc]$ ($X \equiv [bc]$) : $\Omega_2 = aXd$ and

$$\mathcal{L}_{M_3^2}(C) = \mathcal{L}^*_{M_3^2}(\Omega_2) = max \left\{ \begin{array}{l} p([a])p([Xd]) \\ p([aX]p([d]) \\ p([aXd]) \\ p([a])p([X])p([d]) \end{array} \right\}$$

The model is thus defined by the set of parameters Θ_j, $1 \leq j \leq \nu$, consisting of the probability of each sequence $s_j(i)$ in D_{S_j} : $\Theta_j = \{p(s_j(i))\}$, with $\sum_{s_j(i) \in D_{S_j}} p(s_j(i)) = 1$.

The most likely segmentation $L^*_{M_n^j}$ of a training corpus O_j is used to estimate the set of parameters Θ_{j+1}. Θ_1 is estimated on a training corpus O_1 of syntactic classes.

An estimation of the set of parameters Θ from a training corpus O can be obtained as a Maximum Likelihood (ML) estimation from data [3], where the observed data is the string of symbols O, and the unknown data is the underlying segmentation L. Thus, iterative ML estimates of Θ_j can be computed through an EM algorithm. The estimation details of these parameters are showed in [2].

5 Evaluation

In this section, we assess the M$_n^\nu$ model in the framework of language modeling. In our experiments, we decided to set ν to 2 in order to have reliable probabilities of sequences. We compared the M$_n^2$ model with the basic class multigram model and the conventional interpolated class n-gram model. Performance are evaluated in terms of class perplexity [4] on the test and training sets.

In order to evaluate these techniques, we labeled automatically [5] $2,5$ MW (extracted from the french newspaper “Le Monde”). We extracted 10% of this corpus for the test and more than $0,7\%$ for the developpment. The developpment corpus is used to optimise the maximum number (n) of "syntagmatic groups" in a M_n^ν model and the number of occurence (C_0^i) above which a sequence of symbols is included in the initial inventory of sequences at level i. The corpora of development and test do not appear in the training corpus.

To evaluate M_n^2 language model, we proceed as follow: first, we apply the basic class multigram model on a training corpus to compute the level 1 parameters of the model. In this step, we build the dictionary of sequences, which can be formed by combining 1, 2,... up to n syntactic classes $D_{S_1} = \{s_1(i)\}$ and the set of parameters Θ_1, consisting of the probability of each sequence $s_1(i)$ in $D_{S_1} : \Theta_1 = \{p(s_1(i))\}$, with $\sum_{s_1(i) \in D_{S_1}} p(s_1(i)) = 1$. We choose C_0^1 in order to avoid having a huge number of sequences, and in the same time keeping a reasonable computation complexity. These sequences become the dictionary of the model. Secondly, the training corpus is tagged by sequences dictionary of the first step D_{S_1} according to the most likely segmentation $L^*_{M_n^1}$, which can be formed by the use of parameters set Θ_1. Thirdly, the class multigram model is used again, on the tagged training corpus to compute the level 2 parameters (same way that first step). With the set of parameters Θ_2 and the new dictionary of sequences D_{S_2}, we evaluate the perplexity of the M_n^2 language model. In Level 1, we vary C_0^1 from 50 to 750 with a step of 20. To keep an acceptable number of parameters, we fixed the value of C_0^1 to 600 which gives a dictionary of 995 sequences. The experiment concerning the class n-gram model, on the same corpus, gives a perplexity of $13,46$ for the interpolated class bigram model and $11,66$ for the interpolated class trigram model. The comparison of perplexity of M_n^1, M_n^2 (table 1) and class n-gram indicates that from

n		$C_0 = 4$	$C_0 = 5$	$C_0 = 6$	$C_0 = 7$
3	$PP_{M_n^2}$	$14,77$	$13,95$	$14,68$	$15,01$
	$PP_{M_n^1}$	$14,61$	$14,64$	$14,68$	$14,70$
5	$PP_{M_n^2}$	$12,78$	$12,40$	$12,44$	$12,57$
	$PP_{M_n^1}$	$12,58$	$12,48$	$12,52$	$12,61$
7	$PP_{M_n^2}$	$12,41$	$\underline{11,95}$	$12,01$	$12,02$
	$PP_{M_n^1}$	$12,35$	$\underline{12,23}$	$12,28$	$12,32$

Table 1: **This table shows both the test perplexity of the M_n^2 model ($PP_{M_n^2}$) and the class multigram model ($PP_{M_n^1}$). n is the maximum number of words in a sequence and C_0 is the number of occurences above which a sequence of words is included in the initial inventory of sequences (C_0 refers to C_0^2 and C_0^1 for $PP_{M_n^2}$ and $PP_{M_n^1}$ respectively).**

$n = 5$ and $C_0 = 5$, the M_n^2 model is better than both the M_n^1 ($12,23$) and the interpolate class bigram model ($13,46$) but give less good results than interpolated class trigram model ($11,66$). It is important to note that the number of units is in the same order of magnitude for optimal M_n^2 (≈ 63000) and M_n^1 (≈ 67000) but less than the number of units in the class tigram model (80000). We think that our model give less good results than class trigram because of the choice of a great value to C_0^1 (600) at level 1. A smaller choice of C_0^1 (< 100) increases enormously the vocabulary of sequences used at upper levels. This increasement makes the complexity of the models very high.

6 Conclusion and Perspectives

We described in this paper a new language model based on an hierarchical multigram model. Our experiments show that high M_n^ν could be a competitive alternative to the class multigram and interpolated class n-gram. Unfortunately, this new concept is not yet very powerful. The arbitrary choice of C_0^i is not under control. We have to estimate them by using a development corpus. A better sequences labelling will improve the quality of this model. In fact, currently the choice of the i level vocabulary's depends only on the probabilities of the sequences of the $i-1$ level vocabulary. Improvements based on these remarks are under work.

Acknowledgments

We wish to thank Fréderic BIMBOT and Sabine DELIGNE for the package of multigram and for the faithful discussions we have on this new approach.

References

[1] S. Deligne and F. Bimbot. Language modeling by variable length sequences : Theoretical formulation and evaluation of multigrams. In *ICASSP95*, pages 169–172, 1995.

[2] S. Deligne and F. Bimbot. Inference of variable-length linguistic and acoustic units by multigrams. In *Speech Communication*, volume 23, pages 223–241, 1997.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc.*, 39(1):1–38, 1977.

[4] F.Jelinek. Self-organized language modeling for speech recognition. In Alex Waibel and Kai-Fu Lee, editors, *Speech Recognition*, pages 450–506. Morgan Kaufmann, 1990.

[5] K.Smaili, I.Zitouni, F.Charpillet, and J-P.Haton. An hybrid language model for a continuous dictation prototype. In *EUROSPEECH97*, Rhodes (GREECE), September 1997.

[6] K. Ries, F.D. Buo, and Y. Wang. Improved language modeling by unsupervised acquisition of structure. In *Proceding of International Conference on Acoustic Speech Signal Process*, 1995.

[7] B. Suhm and A. Waibel. Towards better language models for spontaneous speech. In *Proceding International Conference on Spoken Language Processing*, 1994.

[8] S. Young. A review of large-vocabulary continuous-speech recognition. In *IEEE Signal Processing Magazine*, pages 45–54, 1996.