Linguistic features modeling based on Partial New Cache

Kamel Smaïli, Caroline Lavecchia and Jean-Paul Haton

INRIA-LORIA, Speech Group
B.P. 101 - 54602 Villers les Nancy, France
Tel.: +33 (0)3 83 59 20 83 - Fax: +33 (0)3 83 27 83 19
e-mail: smaili,lavecchi, jph@loria.fr - http://www.loria.fr/equipes/parole

Abstract

The agreement in gender and number is a critical problem in statistical language modeling. One of the main problems in the speech recognition of French language is the presence of misrecognized words due to the bad agreement (in gender and number) between words. Statistical language models do not treat this phenomena directly. This paper focuses on how to handle the issue of agreements. We introduce an original model called Features-Cache (FC) to estimate the gender and the number of the word to predict. It is a dynamic variable-length Features-Cache for which the size is determined in accordance to syntagm delimitors. This model does not need any syntactic parsing, it is used as any other statistical language model. Several models have been carried out and the best one achieves an improvement of more than 8 points in terms of perplexity.

1. Introduction

In current statistical language models, it is difficult to take into account the agreement between two words, especially when they are not close. In such models, the agreement is hidden in the probabilities assigned to the sentence n-grams.

In French, each word has several linguistic features, and some of them may be incompatible with the features of another word. For instance, the production of a word is affected by the gender and number of its left context. Statistical language models handle the gender and number agreement inadequately, and that contributes to reduce the performance of the French speech recognition systems. For instance, the sentence *Les pommes que j'ai mangé**es** étaient vertes*[1] is often recognized as *Les pommes que j'ai mang**é** était verte* [2]. French is an inflected language with several homonyms words, consequently linguistic features are very useful to reduce speech recognition errors due to this phenomena.

Few research works have been conducted in this area (Bilmes and Kirchhoff, 2003) (Smaïli et al., 2004), and we consider that the introduction of such information will improve the statistical language models. Long distance dependencies in statistical language modeling have been widely explored in the litterature even by using syntactic structure (Chelba and Jelinek, 1998). Our work is related to long distance features dependencies without introducing any parsing nor syntactic rules.

The model we propose, *Feature-Cache* (FC), is inspired from the classical cache model (Kuhn and DeMori, 1990). Henceforth, in our model a word depends not only on its word left context, but also on its gender and number left contexts. The idea is to capture the left context features in order to predict if a word is compatible in terms of features.

2. The Feature-Cache model

2.1. The Cache Model

The Cache model supposes that a word which occured in the recent past is much more likely to be used sooner than indicated by its frequency in the language. That leads to conclude that a classical n-gram is less powerful than a Cache to predict the recent uttered words. The Cache model estimates the probability of a word from its recent frequency of use.

$$P(w_i) = \frac{1}{N} \sum_{j=1}^{N} \delta(w_i, w_j) \quad (1)$$

where N is the length of the cache and $\delta(w_i, w_j)$ is the Kronecker function which is equal to 1 if $w_i = w_j$ and 0 otherwise.

2.2. An outline of the Feature-Cache

Since a Cache is more efficient to predict a word which occurred in the recent past, we have extended this idea to word features. We propose a new model which takes into account the recent word features in order to predict a compatible word in terms of features. For instance, in the sentence: *Les pommes sont vertes*[3], the feature "number" of *vertes* is compatible with its past, the gender of this word is also compatible with *Les* and *pommes*.

In French, some words are insensitive to gender or number, that means that some words may have the same orthographic form in singular and plural as *corps, souris, etc.* Other words are invariant in gender as *égoïste, tranquille, etc.* Consequently, the mass of words having one of both the features could be unbalanced. For this reason, we split the feature-cache model into two feature-Cache model: the gender Feature-Cache and the number feature-Cache. Under this assumption, we propose to estimate the features-Cache probability of a word as follows:

[1] The apples I ate were green

[2] The words mangées and mangé are acoustically identical

[3] The apples are greens

$$P_{FC}(w_i) = \delta \frac{N(G(w_i))}{\sum_{w_j \in V} N(G(w_j))} + \lambda \frac{N(U(w_i))}{\sum_{w_j \in V} N(U(w_j))} \quad (2)$$

where $N(f(x))$ is the occurrence of the feature f of a word x occured in Cache, G the gender feature and U the number feature and V is the vocabulary. Table 1 presents features we used in our model.

Feature	Example
FS (Female - Singular)	porte
MS (Male - Singular)	stylo
FP (Female - Plural)	portes
MP (Male - Plural)	stylos
Fi (Female - Invariant in number)	souris
Mi (Male - Invariant in number)	tapis
iS (Invariant in gender - Singular)	égoïste
iP (Invariant in gender - Plural)	ces
ii (Invariant in gender and number)	beaucoup

Table 1: Features list used in the Cache-feature model

	Mean	Min	Max	σ
α	0.933	10^{-6}	1	0.151
β	0.064	10^{-6}	1	0.151

Table 2: Statitics on EM parameters interpolation

Obviously, this model cannot be used alone; it is linearly interpolated with a classical n-gram. The estimation of a word w_i given a left context h is calculated as follows:

$$P(w_i|h) = \alpha P_{ngram}(w_i|h) + \beta P_{FC}(f(w_i)|Cache) \quad (3)$$

where Cache is a sequence of m features, and α, β are the interpolation parameters. In table 2 some statistics on the interpolation parameters (mean, min, max and standard deviation) are presented. Figure 1 illustrates the evolution of α and β obtained by EM algorithm.

3. Data description

Experiments were performed on Le Monde newspaper corpus. The training corpus contains 32 million words, the development 8 million words and the test 1,8 million words. The vocabulary is made up of the 57000 most frequent words. The features of words are extracted from a French lexical database (BDLEX distributed by ELRA) [4] which contains 430000 words. This database contains the inflected words derived from the canonical words. Each entry includes spelling, pronunciation, morphosyntactic attributes and a frequency indicator.

The length of the Cache-features has been set experimentally to 5. This is due to the fact that the agreement in gender and number has to be done in a close context.

[4]Base de Données LEXicales

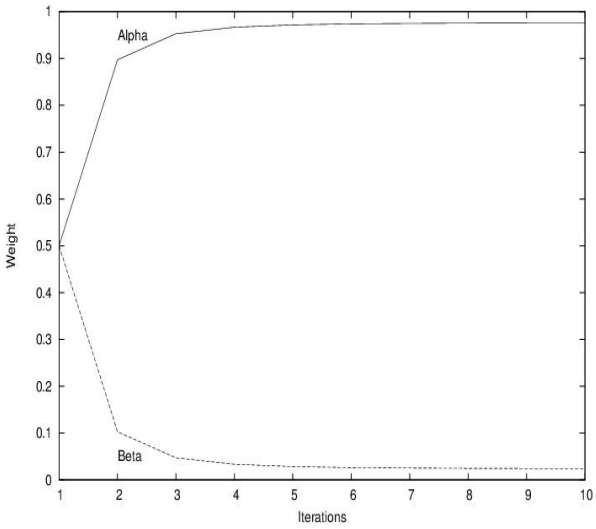

Figure 1: Evolution of interpolation parameters

4. Results and discussion

In order to perform the Feature-Cache language model, we labeled each word with its features by using BDLEX database. In the following, interpolation parameters have been optimized with EM algorithm (Dempster et al., 1977). Two sets of parameters have been calculated: one parameter by model in the mixture (1byM) and one parameter by history (1byH) for interpolated bigrams as shown in table 3. Experiments show that the model is promising and the use of features bring an improvement of 8 points for an interpolated bigram and almost 6 points for an interpolated trigram.

	Baseline	FCache model	
n-gram size	n-gram	1byM	1byH
$PPL_{n=2}$	212.83	206.85	204.75
$PPL_{n=3}$	165.35	159.42	-

Table 3: Results on interpolated Feature-Cache model

Several experiments have been achieved on different corpora and each time a reduction on test-perplexity has been observed. That shows the potential of our approach in terms of test-perplexity reduction, and we have succeed to introduce linguistic features in statistical language models. This is done without introducing any linguistic rule nor parsing technical.

We have to mention that the use of a classical word Cache outperfoms the results we obtained with the combined FC model (by 3 points).

Our objective is to conduct several experiments and to develop other models in order to improve more again the perplexity and hope to outperform our own speech recognition system (Brun et al., 2005).

In the next section we show how to go further by introducing a dynamic feature cache that we call feature-partial-Cache.

5. Feature-Class model

In this section we define a Feature-Class model as a n-gram of feature classes, the probability of word w_i is then defined as:

$$P(w_i|h_{w_i}) = P(w_i|C_{w_i})P(C_{w_i}|h_{C_{w_i}}) \quad (4)$$

where C_x denotes the feature class of x, h_{w_i} is the history of w_i, $h_{C_{w_i}}$ is the class history of C_{w_i}, and $P(w_i|C_{w_i})$ computes the ratio of the word w_i to the number of words having the same features as w_i:

$$P(w_i|C_{w_i}) = \frac{N(w_i)}{N(C_{w_i})} \quad (5)$$

where $N(x)$ is the occurrence of x in the training corpus. We used 9 feature classes described previously in table 1. The conditional probability $P(C_{w_i}|h_{C_{w_i}})$ is estimated as a classical n-gram, where grams are replaced by their corresponding features.

	Baseline	FClass model	
n-gram size	n-gram	1byM	1byH
$PPL_{n=2}$	212.83	211.82	209.78
$PPL_{n=3}$	165.35	163.39	-

Table 4: Perplexities for baseline and Feature-Classe

Unfortunately the class feature model leads to a weak improvement. Thus we decided to drop this model.

6. Partial features-Cache

Experimental results show the interest of the Feature-Cache, but this model may introduce errors in the relationship between word features. Actually, in the sentence *Le portefeuille bleu de ma grande soeur est beau* [5], the word *soeur*, which is singular female, will be affected by the dominant features in the Cache (singular male) and consequetly its probability will decrease. In fact, because no parsing technique is conducted, we should only take into account the agreement between words inclosed on buckets or syntagms.

In the previous example, the compatibility of features has to be checked inside the syntagms *le portefeuille bleu* or *ma grande soeur*, the word *beau* has to be treated with a distant model (Huang and al., 1993), (Langlois and Smaïli, 1999). Actually, the features of the word *soeur* have to be checked inside the last syntagm. With this assumption, the cache has to be splitted into buckets (or syntagms). The decomposition of a word history leads to what we call a Partial Feature-Cache. For that, we have to find out the limits of linguistic groups or syntagms.

To deal with this issue, we use a list of tool words as prepositions, conjunctions, etc, which are unvarying in gender and number, in order to set the limits of groups. These separators permit to retrieve dynamically the adequate size of the feature-Cache, leading to what we call a partial-feature-Cache.

In the previous example, the word *de* separates two syntagms. The results presented in table 5 were obtained by using the separators *de* and *du*. Despite the slight improvement, we are convinced that the agreement in gender and number has to be considered inside a group of words delimited by separators.

	Baseline	Partial-Fetaure-Cache	
n-gram size	n-gram	1byM	1byH
$PPL_{n=2}$	212.83	206.57	204.43
$PPL_{n=3}$	165.35	159.19	-

Table 5: Results on Feature-Partial-Cache using separators *de and du*

In order to go further, we introduce other separators (table 6).

de	du	mais	ou	et	donc	or
ni	car	dans	avant	depuis	que	qui

Table 6: Syntagm word separators

The introduction of these separators leads to an interpolated bigram perplexity of 203.95 (table 7) and an interpolated trigram perplexity of 159.18.

Overall, our approach allows to decrease the bigram perplexity by almost 9 points. We have to continue our investigation in order to delimit more correctly the boundaries of syntagms in a Feature-Cache.

	Baseline	Partial-Fetaure-Cache	
n-gram size	n-gram	1byM	1byH
$PPL_{n=2}$	212.83	206.56	203.95
$PPL_{n=3}$	165.35	159.18	-

Table 7: Perplexities on Feature-Partial Cache using an extended list of sepators

We decided to analyse the weights assigned to the Feature-Cache model, this study shows that only 3655 histories have a weight greater than 0.3 and only 736 among them have weights which exeed 0.6. Table 8 gives some histories and the corresponding Feature-Cache ponderation which are significant. Despite the weak number of histories, the contribution of Partial-feature-Cache brings an improvement over a n-gram.

7. Conclusion

In this paper, we presented an original statistical language model based on features of word. The idea is to consider a word not only as an orthographic form, but as a linguistic unit which has several attributes. We focused in this work on only two features, gender and number. Several feature models, based on statistical language formalisms have been developed, in order to find the best one. The features have been considered inside a left short window of the word to predict. Significant performance have been achieved with a variable-length Cache (Partial-Feature-Cache). The interpolated bigram test-perplexity has been decreased by almost 9 points. For the interpolated trigram, despite using

[5] The blue wallet of my older sister is beautiful

Word	FC weight	Word	FC weight
Derrick	0	apportée	1
Tiozzo	0.06	conceptuels	1
increvable	0.19	concertation	0.99
fuseaux	0.21	concurrente	0.99
défaits	0.26	voies	0.90
Sun	0.27	restaient	0.78
votre	0.31	ressentons	0.76
arrière-goût	0.36	verdeur	0.61

Table 8: Some histories leading to significant weights (left part)

a sub-optimal weights for both combined models, an improvement of more than 6 points has been obtained. With these results, we showed the feasibility of the features language model concept and the easiness way to formalize them. We just integrate features in classical and baseline statistical language models. In a near future, we will conduct experiments in order to take into account the agreement between the subject and its verb, this will be done by introducing other features. As for, the use of Partial feature cache model in a speech recognition system, this will be done on the second pass decoding.

8. Acknowledgement

This research is supported by EADS foundation in the framework of speech-to-speech translation Ph.D thesis. This work begun before the official start day of Ph.D.

9. References

J. A. Bilmes and K. Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *Proceeding of Human Language Technology Conference*, Edmonton, Canada.

Armelle Brun, Christophe Cerisara, Dominique Fohr, Irina Illina, David Langlois, and Odile Mella. 2005. Ants le système de transcription automatique du loria. In *Workshop ESTER, Avignon, France*, Mar.

C. Chelba and F. Jelinek. 1998. Exploiting syntactic structure for language modeling. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 225–231.

A. Dempster, N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B*, 39:1–38.

X. Huang and al. 1993. The sphinx speech recognition system: an overview. *Computer speech and Language*, 2:137–148.

R. Kuhn and R. DeMori. 1990. A cache-based natural language model for speech recognition. *IEEE Trans. PAMI*, 12(6):570–582.

D. Langlois and K. Smaïli. 1999. A new based distance language model for a dictation machine : application to maud. In *Proc. EUROSPEECH*, pages 1779–1782.

K. Smaïli, S. Jamoussi, D. Langlois, and J. P. Haton. 2004. Statistical feature language model. In *Proc. ICSLP*, Jeju.