How to handle gender and number agreement in statistical language models?

Caroline Lavecchia, Kamel Smaïli, Jean-Paul Haton

How to handle gender and number agreement in statistical language models?

Caroline Lavecchia, Kamel Smaïli and Jean-Paul Haton

INRIA-LORIA, Speech Group
B.P. 239 - 54602 Villers les Nancy, France
Tel.: +33 (0)3 83 59 20 83 - Fax: +33 (0)3 83 27 83 19
e-mail: {lavecchi, smaili,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 difficulties in 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 this agreement. 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. The size of the cache is automatically determined in accordance to syntagm delimitors. The main advantage of this model is that there is no need to 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 approximatively 9 points in terms of perplexity. This model has been integrated in a speech recognition system based on JULIUS engine. Tests have been carried out on 280 sentences provided by AUPELF for the French automatic speech recognition evaluation campaign. This new model outperforms the baseline one, in terms of word error, by 3%.

Index Terms : statistical language modeling, features, Features Cache, Partial Features Cache.

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 [1] [2], 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 [3]. Our work is related to long distance features dependencies without introducing any parsing nor syntactic rules.

The model we propose, *Features-Cache* (FC), is inspired from the classical cache model [4]. 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 Features-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 Features-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 Features-Cache model into two Features-Cache models : the gender Features-Cache and the number Features-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 green

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

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

TAB. 1 – Features list used in the Features-Cache model

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

TAB. 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

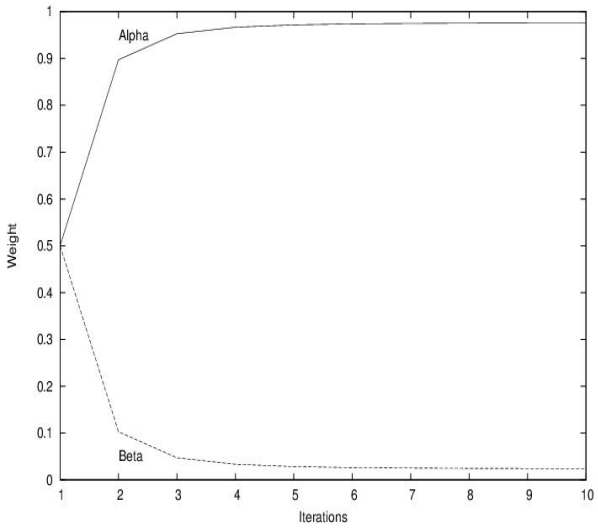

FIG. 1 – Evolution of interpolation parameters

4. Results and discussion

In order to perform the Features-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 [5]. 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. 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).

	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	-

TAB. 3 – Results on interpolated Features-Cache model

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 [6].In the next section we show how to go further by introducing a dynamic feature cache that we call Partial Features-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	-

TAB. 4 – Perplexities for baseline and Feature-Class

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 Features-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 [7], [8]. 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 Features-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 features-Cache, leading to what we call a partial Features-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. In order to go further, we introduce other separators (table 6). 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 Features-Cache. We decided to analyse the weights assigned to the Features-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 Features-Cache brings an improvement over a n-gram.

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

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

TAB. 5 – Results on Partial Features-Cache using separators *de* and *du*

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

TAB. 6 – Syntagm word separators

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

TAB. 7 – Perplexities on Partial Features-Cache using an extended list of sepators

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

TAB. 8 – Some histories leading to significant weights (left part)

7. Speech recognition results

7.1. An overview of the speech engine

In order to evaluate our approach in a speech recognition system, we used the JULIUS speech engine. JULIUS is an open source engine recognition originally developed by Akinobu Lee at Kyoto university [9]. Two passes are performed. In the first pass a tree-structured lexicon associated to a bigram is applied with the frame-synchronous beam search algorithm. This first pass produces a word lattice. The second pass is based on a trigram

model and researches the best sentence in the word lattice.

7.2. Implementation

In the first pass we use a standard bigram trained on ten years extracted from the French newspaper *Le Monde*. The acoustic model is based on HMM phone without adaptation to gender or speaker. In the second pass, we integrate our Partial Features-Cache presented in the previous section. Tests have been carried out on 280 sentences provided by AUPELF for the French automatic speech recognition evaluation program [10].

7.3. Analysis and Review

Table 9 shows the results obtained by integrating a Partial Features-Cache in the second pass instead of the standard trigram. This new language model has a real impact and achieves an improvement of 3% of the word error. We can also notice that all the bold rates of table 9 indicate slight improvements in comparison to the corresponding rates of the standard recognition engine.

Model	#Snt	#Wrd	**Corr**	**Sub**	Del	**Ins**	**Err**
Baseline	134	4076	**61.8**	**25.2**	13.1	**1.9**	**40.2**
P-F-C	134	4076	**62.4**	**24.0**	13.6	**1.4**	**39.0**

TAB. 9 – Performance in terms of word error

8. Conclusion

In this paper, we presented an original statistical language model based on features of words. The idea is to consider a word not only as an orthographic form, but as a linguistic unit with several attributes. In this work, we focused 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 Features-Cache). The interpolated bigram test-perplexity has been decreased by almost 9 points. For the interpolated trigram, despite using 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. The Partial Features-Cache model has also been integrated in the second decoding pass of the speech recognition system. An improvement of 1.2 points of the WER has been reached.

In future works, we will conduct experiments in order to take into account other agreements, such as the agreement between a subject and its verb. In this scope, new features will be introduced. Thereafter, we will be interested in automatically new agreements according to features considered.

9. Acknowledgements

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.
The authors whish to thank Dr. David Langlois and Dr. Armelle Brun for their precious help.

10. References

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

[2] K. Smali, S. Jamoussi, D. Langlois, and J. P. Haton, "Statistical feature language model," in *Proc. ICSLP*, Jeju, 2004.

[3] C. Chelba and F. Jelinek, "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*, 1998, pp. 225–231.

[4] R. Kuhn and R. DeMori, "A cache-based natural language model for speech recognition," *IEEE Trans. PAMI*, vol. 12, no. 6, pp. 570–582, 1990.

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

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

[7] X. Huang and al., "The sphinx speech recognition system : an overview," *Computer speech and Language*, vol. 2, pp. 137–148, 1993.

[8] D. Langlois and K. Smaïli, "A new based distance language model for a dictation machine : application to maud," in *Proc. EUROSPEECH*, 1999, pp. 1779–1782.

[9] A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, 2001, pp. 1691–1694.

[10] G. Adda, M. DeCalmes, L. Lamel, G. Perennou, M. Rajman, S. Rosset, and J. Zeiliger, "Ressources pour l'apprentissage, le dveloppement et l'valuation des systmes de dicte vocale en français : corpus de texte, de parole et lexical," in *Actes des premires JST Francil 1997*, Avignon, France, April 1997, pp. 305–309.