Improving language models by using distant information

Armelle Brun, David Langlois, Kamel Smaïli

IMPROVING LANGUAGE MODELS BY USING DISTANT INFORMATION

A. Brun[2], D. Langlois[1] and K. Smaïli[2]

INRIA Lorraine
BP 239
54506 Vandoeuvre-lès-Nancy Cedex
{brun, langlois, smaili}@loria.fr
1. IUFM of Lorraine   2. University Nancy 2

ABSTRACT

This study examines how to take originally advantage from distant information in statistical language models. We show that it is possible to use n-gram models considering histories different from those used during training. These models are called crossing context models. Our study deals with classical and distant n-gram models. A mixture of four models is proposed and evaluated. A bigram linear mixture achieves an improvement of 14% in terms of perplexity. Moreover the trigram mixture outperforms the standard trigram by 5.6%. These improvements have been obtained without complexifying standard n-gram models. The resulting mixture language model has been integrated into a speech recognition system. Its evaluation achieves a slight improvement in terms of word error rate on the data used for the francophone evaluation campaign ESTER [1]. Finally, the impact of the proposed crossing context language models on performance is presented according to various speakers.

1. INTRODUCTION

Statistical language n-gram models represent efficiently local constraints. However such models have limits due to the size of the history $(n-1)$: in the case of a larger relationship, such a dependence cannot be considered. It is well-known that the value of n is a serious limit for handling constraints of natural language. Growing up this value of n leads to sparseness data problems [2]. Class-based language models [3] can be viewed as a generalization of n-grams, lowering the missing data problem. Unfortunately, these models are less efficient than n-grams. One way to support more constraints without increasing the value of n is to take advantage from distant information. In previous work [4], we showed the improvement brought by distant n-grams. Distant n-grams allow to consider dependence between a word and a sequence of words which is more or less close.

In this paper an extension of this idea is presented. In n-gram models the probability of a word is always estimated in accordance to the $n-1$ preceding words. Given the sentence *le chien méchant aboie*[1], a 3-gram model uses the phrase *chien méchant* to estimate the probability of word *aboie*. Guess that in training corpus, *chien* is systematically followed by *méchant*. However, phrases *le chien aboie*[2], *le chien grogne*[3] could be likely in a test corpus. This means that the probability of the word *aboie* given *le chien* will be underestimated. This issue could be handled by using a distant n-gram for the estimation of a contiguous n-gram. Now, let assume that *chien aboie* has been met in a contiguous bigram and has never been seen in a distant context *chien ... aboie*. In this case it should be interesting to use a classical n-gram in a distant context.

---

[1] the nasty dog barks
[2] The dog barks
[3] The dog growls

In this paper we propose to use both distant and non distant language models by crossing left contexts, as pointed out in Table 1. For instance, the second line shows a typical n-gram used in a distant test context.

	Name	Training	Test
n-gram	M_{00}	$P(w_i|h)$	$P(w_i|h)$
n-gram	M_{0d}	$P(w_i|h)$	$P(w_i|h_d)$
distant n-gram	M_{dd}	$P_d(w_i|h_d)$	$P_d(w_i|h_d)$
distant n-gram	M_{d0}	$P_d(w_i|h_d)$	$P_d(w_i|h)$

Table 1. Crossing-context models

Where $h = (w_{i-n+1} \ldots w_{i-1})$ and $h_d = (w_{i-n+1-d} \ldots w_{i-1-d})$. $P(\cdot|\cdot)$ is the probability assigned by the classical n-gram model and $P_d(\cdot|\cdot)$ is the probability assigned by the distant n-gram model (with distance d).

The idea presented in this paper is related to the following general one: given a history, only parts are really useful for prediction. Each part provides information at different levels: gender, number, semantics, etc. We have to find a method that automatically retrieves useful parts from history, and assign the corresponding optimal model to each part. Such a work has been initiated with the Selected History Principle [13] and Feature Language models [14].

Section 2 presents an overview of distant language models. Section 3 formalizes crossing context models. Then experimental data is put forward. In section 5 evaluations of the perplexity of language models are computed. Then, results of the combination of these language models are shown. Section 7 integrates these new models in a speech recognition system. A conclusion and perspectives of this work are then discussed.

2. AN OVERVIEW OF DISTANT LANGUAGE MODELS

A distant language model deals with non contiguous context. In the following this distance is set to d. A distant n-gram [4, 5] estimates the probability of a word w_i given a sequence of words $h_d = (w_{i-n+1-d} \ldots w_{i-1-d})$ located exactly at d words before w_i. Let's remark that a distance $d = 0$ corresponds to a classical n-gram model.

In this case, the probability of a word w_i is:

$$P_d(w_i \mid h_d) = \frac{N_d(h_d, w_i)}{N(h_d)} \tag{1}$$

Where $N_d(h_d, w_i)$ is the frequency of $h_d w_i$ (w_i occurs at a distance d from h_d). Obviously, due to the distance between h_d and w_i, such a model is less powerful than a baseline n-gram. However, using distance is efficient when corresponding models are combined with classical n-grams [6].

3. CROSSING CONTEXT MODELS

This section describes how standard models (n-gram and distant n-gram) are used by crossing their contexts. For instance, for the bigram case, M_{01} is used in distant way by employing a standard bigram (see Table 1). This conducts to a non standard way of bigram usage. However these models continue obeying to a probability distribution. Actually, for each context and each model a probability distribution is assigned. In other words:

$$P_{01}(w_i|w_1,\ldots,w_{i-1}) \stackrel{def}{=} P(w_i|w_{i-2}) \qquad (2)$$

where P_{01} is the probability assigned by model M_{01} and:

$$\forall v \in V \quad \sum_{w \in V} P(w|v) = 1 \qquad (3)$$

where P is the probability assigned by the standard (non distant) bigram model and V is the vocabulary.
Obviously, this model remains statistically correct because P sums up to one for all histories, and thus, for this specific history w_{i-2}.

In the same way, this property is also true for model M_{10}.

4. EXPERIMENT MATERIAL

Training, development and test data are extracted from *Le Monde newspaper*. Twelve years (1987 to 1998) have been devoted to training (288 million words). The development corpus (79 million words) is made up of three years (1999-2001) and the test has been performed on 27 million words corresponding to year 2002. The vocabulary contains 60K words and corresponds to the one used for ESTER[4] evaluation campaign [1].

5. EVALUATION

In this section the crossing context models are evaluated in terms of bigram and trigram perplexity. The models have been smoothed by using the absolute discounting method [7]. d has been set to 1, in fact in [6] we showed that performance of distant models fall dramatically when distance increases. Table 2 presents the corresponding perplexities.

Model	Test Perplexity	
	bigrams	trigrams
M_{00}	**164.7**	**100.4**
M_{11}	499.4	390.6
M_{10}	2403.8	2838.9
M_{01}	20632.8	21716.7

Table 2. Evaluation in terms of perplexity

These results show obviously that baseline models (bigram and trigram) are widely better than crossing context models. The baseline distant models (bigram and trigram) give reasonable results. In the opposite, crossing context models M_{01} and M_{10} are not efficient. This is not surprising and was expected because of the mismatch between training and test parameters. Crossing context models may bring improvement only when they are used adequately in specific cases. The initial idea was to use these special models when standard ones are deficient. While crossing context models are not efficient in most cases, they should be satisfactory for specific histories. Consequently, we have to find the best way to use them with the aim to improve perplexity.

---

[4]Evaluation des Systèmes de Transcription d'Emissions Radiophoniques

6. COMBINATION OF MODELS

To take advantage from crossing context models M_{10} and M_{01}, we decided to combine them with baseline models M_{00} and M_{11}. Several combination methods are frequently used: maximum entropy [8], linear interpolation [9], *etc*. In the following, we use a linear interpolation. We will perform two experiments. The first uses one weight per model and for the second a set of weights is assigned to each model [9].

6.1. Linear combination independent on the history

In this experiment, a weight is assigned to each model. The set of weights is obtained by EM algorithm [10]. Different tests have been performed in order to study the impact of each model in comparison to the baseline model. Results of experiments on bigram and trigram models are given in Tables 3 and 4.

Used models	Weights				Perplexity
	M_{00}	M_{01}	M_{10}	M_{11}	
$M_{00} + M_{10}$	0.991	–	0.009	–	164.8
$M_{00} + M_{01}$	0.986	0.014	–	–	164.4
$M_{00} + M_{11}$	0.870	–	–	0.130	**157.2**
$M_{00} + M_{01}+$ $M_{10} + M_{11}$	0.878	0.002	0.003	0.117	**157.2**

Table 3. Bigram perplexity of various models linearly interpolated

Used models	Weights				Perplexity
	M_{00}	M_{01}	M_{10}	M_{11}	
$M_{00} + M_{10}$	0.985	–	0.015	–	100.3
$M_{00} + M_{01}$	0.990	0.010	–	–	100.1
$M_{00} + M_{11}$	0.921	–	–	0.079	**97.7**
$M_{00} + M_{01}+$ $M_{10} + M_{11}$	0.923	0.002	0.005	0.069	**97.6**

Table 4. Trigram perplexity of various models linearly interpolated

These results show that the only combination that improve the baseline (M_{00}) are the $M_{00}+M_{11}$, for both cases bigram and trigram, by respectively 4.6% and 2.6%. Other combinations either do not improve perplexity or improve it slightly. This kind of context independant weights does not take advantage from the contextual specificity of each model in the mixture. Consequently, these models have to be used judiciously in the appropriate context. That is why, in the following experiments, each model uses a set of weights depending on the history.

6.2. Linear combination depending on the history

Due to the huge number of left contexts, consistent coefficients for each history cannot be obtained. One solution consists in reducing the space of parameters. In the following we first propose to reduce this space by putting histories together according to their frequency. Secondly, the space is reduced by computing weights for only a subset of frequent histories.

6.2.1. Frequency dependent weights

In this case, a set of weights is assigned to each bucket made up of words which have the same frequency [9].This method leads to 9K buckets.

Used models	2-gram Test PP	3-gram Test PP
$M_{00} + M_{10}$	164.3	100.1
$M_{00} + M_{01}$	162.6	99.9
$M_{00} + M_{11}$	150.2	96.8
$M_{00} + M_{01} +$ $M_{10} + M_{11}$	**149.3**	**96.5**

Table 5. Perplexity tests based on frequency dependent weights

Table 5 illustrates the bigram and trigram perplexity results according to this classification.

The improvement of perplexity confirms our idea: in particular contexts, it is benefit to use a model with a part of history specific to another model. The combination $M_{00} + M_{01} + M_{10} + M_{11}$ improves the baseline models M_{00} for both bigram and trigram cases respectively by 9.4% and 3.9%. Thus the context dependent combination improves performance of context independent combination.

6.2.2. Rank dependent weight

In this solution, the space reduction is achieved by selecting the most frequent histories. Thus, for each model and each frequent history, a weight is assigned. A unique coefficient is attributed to remaining histories. Several experiments have been conducted in order to find out the optimal weights, resulting in 900K statistically significant weights.

Table 6 illustrates the bigram and trigram perplexity results according to this classification.

Used models	2-gram Test PP	3-gram Test PP
$M_{00} + M_{10}$	163.7	100
$M_{00} + M_{01}$	159.8	99.6
$M_{00} + M_{11}$	143.3	94.7
$M_{00} + M_{01} +$ $M_{10} + M_{11}$	**141.1**	**94.4**

Table 6. Perplexity tests based on rank dependent weights

The investigation carried out by this experiment shows the usefulness of these models. An important improvement (14.3%) of the baseline bigram is achieved by the mixture. The baseline trigram is outperformed by 5.6%.
We point out that the improvement obtained by the trigram mixture has been carried out by the combination of the baseline and the distant trigram. Crossing context models have only a slight impact in the global mixture. We also have to mention that this mixture model is more performant (4.6% improvement in terms of perplexity) than a linear combination between a classical trigram and a bigram.

The rank dependent combination slightly improves the performance of the frequency dependent combination, this may be explained by the number of classes which is 100 times more important.

These experiments show that both mixtures (frequency dependent and rank dependent) assign a weight greater than 0.9 to M_{00} in about 75% of the histories. However, the contribution of crossing context models is important in many histories: crossing context models are assigned a weight greater than 0.3 in more than 12% of contexts, among them 2% of the histories assign a weight equal to 1 to crossing context models.

7. SPEECH RECOGNITION RESULTS

7.1. An overview of the speech engine

In order to evaluate our approach in a speech recognition system, we integrate the corresponding language model in the ANTS system [11]. This system has been developed at LORIA and used for ESTER [1], the French broadcast news transcription evaluation campaign. It is based on four sequential stages:

- broad-band/narrow-band speech segmentation,
- speech/music classification,
- detection of silences and breath noises,
- large vocabulary speech recognition.

The aim of the three first stages is to split the audio stream into homogeneous segments with a manageable size and to allow the use of specific algorithms or models according to the nature of the segment. Four sets of acoustic triphones models are used according to the female/male and telephone/non telephone dimensions. Produced segments are automatically regrouped into clusters and a MLLR adaptation is applied on each cluster.

ANTS is based on JULIUS, an open source engine recognition originally developed by Akinobu Lee at Kyoto university [12]. 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 French newspaper *Le Monde* and radio data. In the second pass, we integrate the mixture ($M_{00} + M_{01} + M_{10} + M_{11}$) with and without context dependent weights.

The test data is made up of 30 minutes of French broadcast news extracted from the ESTER data set.

Table 7 presents recognition performance, in terms of correct words, substitution, deletion, insertion and word error rate, of the three following models:

- **A:** the baseline trigram model leading to a perplexity of 100.4 (section 5).
- **B:** the trigram mixture with a set of context independent weights, leading to a perplexity of 97.6 (section 6.1).
- **C:** the trigram mixture with a set of rank dependent weights, leading to a perplexity of 94.4 (section 6.2.2).

Models B and C lead to a slight improvement of the baseline model (A) in terms of word error, insertion and substitution rates.
In addition, we studied the performance of crossing context language models regarding to speakers. In the test data the speaker of each segment is known, 30 different speakers are referenced. For both models B and C, Table 8 indicates the number of speakers for which we observe an improvement (+) in terms of WER, the number of speakers with a decrease of performance (-), and the number of speakers without any change (=). In both models, the comparison is made in accordance to the baseline model (trigram).

The performance for a speaker depends on the environment, noise, speed of elocution, accent, etc. That is why we investigate how our crossing context language models might have an impact on the speaker WER. For instance, to recognize efficiently

Id	Models	Correct	Substitution	Deletion	Insertion	WER
A	baseline	75.5	18.4	6.1	5.2	29.7
B	cont. indep. comb.	75.2	18.4	6.3	4.8	29.6
C	cont. dep. comb.	75.5	18.1	6.4	4.9	29.4

Table 7. Performance in terms of word error

Condition	+	-	=
B	12	13	5
C	9	15	6

Table 8. Comparison of WER by speakers

a foreigner speaker with special language features, a specific language model should be used. Obviously, this issue is commonly handled by a speaker adaptation module. The table shows contrasted results. However, by the use of a speaker identification module, we could use the language model which makes his performance better.

8. CONCLUSION

In this study we examine the pertinence of using, in particular contexts, a part of history specific to another model. We apply this idea to classical and distant n-gram models with a distance equal to 1. This leads to a mixture of 4 models.

Experimental investigation of crossing left contexts of baseline and distant n-grams shows the feasability of this idea and its contribution in the improvement of perplexity. It outperforms the standard bigram and trigram models by respectively 14% and 5.6%. Its integration in a real speech recognition system achieves a slight improvement of the word error rate on broadcast news corpus.

In a future work, the space reduction of histories will have to be managed in a different way. Actually, we used a basic method which separates the histories according to both frequency and rank. We have to investigate other ways to cluster them judiciously in order to improve significantly the word error rate. For example, the used of POS tagging can be envisaged.

9. REFERENCES

[1] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, and K. Choukri, "The ester evaluation campaign of rich transcription of french broadcast news," in *Proceedings of the Language Resources and Evaluation Conference*, Lisbon, 2004.

[2] Sven C. Martin, Jörg Liermann, and Hermann Ney, "Adaptative topic-dependent language modelling using word-based varigrams," in *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes, Grce, Septembre 1997, vol. 3, pp. 1447–1450.

[3] K. Smaïli, A. Brun, I. Zitouni, and J.-P. Haton, "Automatic and manual clustering for large vocabulary speech recognition: a comparative study," in *Proceedings of the European Conference on Speech Communication and Technology*, Budapest, Septembre 1999, vol. 4, pp. 1795–1798.

[4] David Langlois, Kamel Smaïli, and Jean-Paul Haton, "Efficient linear combination for distant n-gram models," in *8th European Conference on Speech Communication and Technology - Eurospeech'03, Genève, Suisse* , Sep 2003, vol. 1, pp. 409–412.

[5] Xuedong Huang, Fileno Alleva, Hsiao-Wuen Hon, Mei-Yuh Hwang, Kai-Fu Lee, and Ronald Rosenfeld, "The SPHINX speech recognition system: an overview," *Computer Speech and Language*, vol. 2, pp. 137–148, 1993.

[6] David Langlois and Kamel Smaïli, "A new based distance language model for a dictation machine : application to maud," in *6th European Conference on Speech Communication and Technology - EUROSPEECH'99, Budapest, Hungary*, Sep 1999, vol. 4, pp. 1779–1782.

[7] Hermann Ney and Ute Essen, "On smoothing techniques for bigram-based natural language processing," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toronto, Mai 1991, vol. 2, pp. 825–828.

[8] R. Rosenfeld, *Adaptative Statistical Language Modeling: A Maximum Entropy Approach*, Ph.D. thesis, School of Computer Science, Carnegie Mellon University, April 1994.

[9] F. Jelinek and R.L. Mercer, "Interpolated estimation of markov source parameters from sparse data," *Pattern Recognition in Practice*, pp. 381–397, 1980.

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

[11] D. Fohr, O. Mella, I. Illina, and C. Cerisara, "Experiments on the accuracy of phone models and liaison processing in a french broadcast news transcription system," in *8th International Conference on Spoken Language Processing - ICSLP'2004, Jeju, South Corea*, Octobre 2004.

[12] A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine," in *Proceedings of the European Conference on Speech Communication and Technology*, 2001, pp. 1691–1694.

[13] David Langlois, Kamel Smaïli, and Jean-Paul Haton, "A new method based on context for combining statistical language models," in *Third International Conference on Modeling and Using Context - CONTEXT 01, 2001*, Dundee, Scotland, 2001, vol. 2116 of *Lecture Notes in Artificial Intelligence*, pp. 235–247, Springer.

[14] Kamel Smaïli, Salma Jamoussi, David Langlois, and Jean-Paul Haton, "Statistical feature language model," in *8th International Conference on Spoken Language Processing - ICSLP' 2004, 2004*, Jeju, Corée du Sud, 2004, p. 4 p.