Efficient linear combination for distant n-gram models

David Langlois, Kamel Smaïli, Jean-Paul Haton

LORIA
Université Henri Poincaré, Nancy I
BP 239, 54506 Vandœuvre-lès-Nancy, France
{langlois,smaili,jph}@loria.fr

Abstract

The objective of this paper is to present a large study concerning the use of distant language models. In order to combine efficiently distant and classical models, an adaptation of the back-off principle is made. Also, we show the importance of each part of a history for the prediction. In fact, each sub-history is analyzed in order to estimate its importance in terms of prediction and then a weight is associated to each class of sub-histories. Therefore, the combined models take into account the features of each history's part and not the whole history as made in other works. The contribution of distant n-gram models in terms of perplexity is significant and improves the results by 12.8%. Making the linear combination depending on sub-histories achieves an improvement of 5.3% in comparison to classical linear combination.

1. Introduction

In classical statistical language models (SLM), the whole history is used as a single block in order to recognize or to predict the next word (Fig 1). The aim of this paper is to show that the history contains words which have different influence and consequently weights for prediction (Fig 2).

In order to measure the usefulness of each history's part, distant n-gram models seem to be a good way to deal with this problem [1]. The following examples illustrate this. In the sentence "The book I bought is brown", "brown" is related to "book". The other words between them do not participate directly to this relationship. In the sentence "This changed conviction into certainty", we guess there is a relationship between the words "conviction" and "certainty". This is a semantic relationship: the two words cover the same idea of "evaluation of a fact".

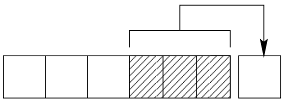

Figure 1: Contiguous relationship taken into account by a classical n-gram model (here a 4-gram model).

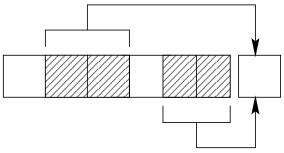

Figure 2: Example of possible distant relationships between the word to predict and several components of the history.

In this paper, we present:

- A simple linear combination between a classical model and a distant one. A unique weight is assigned to each model.
- A linear combination as in previous point but the models' weights here depend on each history. More exactly, a model's weight depends on the class of each sub-history (i.e. history's part).

The originality of this research is concerned by the development of a well chosen weights set for each language model. These weights are dependent on:

- the language model participating in the combination,
- the history's part used by the language model to predict the next word.

In other words, the weights are estimated not on the entire history but on each used component.

The reminder of this paper is organized as follows: section 2 gives an overview about distant language models dealing with the histories part by part. Section 3 presents a simple linear combination between classical and distant language models. In section 4, the proposed model is enhanced by integrating language model weights depending on histories or more exactly on sub-histories' classes. Experiments are presented in each section and the results confirm the thesis defended in this paper. Finally, conclusions are drawn in section 5.

2. Modelization of distance in SLM

The two most common distant models used in literature are the cache model and the trigger model [2].

The former deals with the self-relationship between a word present in the history and itself: if a word is frequent in the history, it has more chance to appear once again. This is a distant relationship because the history spans other a larger number of words: several hundred words for a cache model against one, two or three words for a n-gram model.

The latter modelizes the relationship between two words. It deals with couple of words $v \rightarrow w$ such that if v (the triggering word) is in the history, w (the triggered word) has more chance to appear. For example, "exportation" → "corn" could be a trigger. Such a model is distant because the respective occurrences of the two words can be distant. As just said, this last model deals with couple of possibly distinct words. But, in fact, the majority of triggers are self triggers ($v \rightarrow v$): a word triggers itself.

These two models deal with distant relationships but loose a great part of local and syntactic relationships. The relationship is just between a bag of words in the history and the word to predict. In addition, the purpose of this research is to study the usefulness of each history's part in terms of prediction. That is why we decided to work on distant n-gram models [3, 4]. These models use a distant part in history to predict the word to follow. Such models could be defined by two parameters: the size $n-1$ of the history's part or sub-history and the distance d between this sub-history and the word to predict. More formally, we define a d-n-gram model by:

$$P_d(w_i|w_1 \ldots w_{i-1}) = P_d(w_i|w_{i-n+1-d} \ldots w_{i-1-d}) = \frac{N_d(w_{i-n+1-d} \ldots w_{i-1-d}, w_i)}{N(w_{i-n+1-d} \ldots w_{i-1-d})} \quad (1)$$

where $N_d(w_{i-n+1-d} \ldots w_{i-1-d}, w_i)$ is the frequency of $w_{i-n+1-d} \ldots w_{i-1-d} \ldots w_i$. Note that a 0-$n$-gram model is the classical n-gram model.

This model allows us to study separately each history's part by varying d and n. In the following, we show the efficiency and the contribution of distant models when they are combined with classical models.

3. Evaluation of distant n-gram models

The following models are developed by using a vocabulary made up of 20 000 words, and a training corpus extracted from the French newspaper Le Monde (38 Million words). A corpus of 2 million of words has been devoted to development and another one to test (also 2 million of words). Performance of the baseline models are given in Table 1. Each model is linearly combined with its lower-order models to prevent from unseen events. The parameters of these linear combinations are estimated using the Expectation-Maximization algorithm [5] on the development data.

Models	Perplexity
unigram	739.9
bigram	132.4
trigram	97.8

Table 1: Baseline classical n-gram models: performance when using the linear combination.

3.1. Integration of distant n-gram models

It is obvious that a distant n-gram model could not be used alone. In fact, it takes into account only a part of the history. Experiments shown that when we use it alone the perplexity could reach a very high value (717 for $n = 2$ and $d = 4$, see [6]).

In the light of this remark and to take advantage of distant models, we decided to combine them with classical n-grams. Several models with distance up to d are combined with the baseline model.

Figure 3 shows the performance of such combinations for bigram models. In the same way, Figure 4 shows the performance for trigram models.

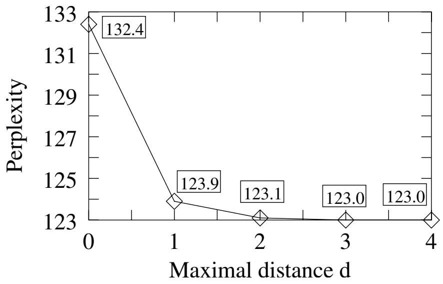

Figure 3: Combination of several distant bigram models with distance equal or less than d.

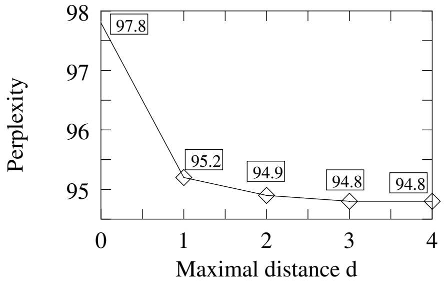

Figure 4: Combination of several distant trigram models with distance equal or less than d.

The use of different distant bigram models achieves an improvement of 7.1% in terms of perplexity. Also, distant trigram leads to an improvement, but it is less important than in the previous case (3.1%). We explain this difference by the overlap between the history of d-trigram and $(d+1)$-trigram. In other words, the intersection between two distant trigram models could be not null as in Figure 5.

To sum up, these experiments confirm that a distant language model provides information which is necessary for combination. But the utility of distant n-gram models

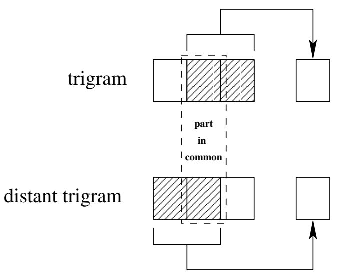

Figure 5: Overlap of the sub-histories used for prediction by a distant trigram model (here a 1-trigram) and a distant trigram model with lower distance (here a 0-trigram).

decreases with the distance: a distance greater than 2 does not provide more information.

3.2. Back-off smoothed models

The back-off [7] principle, as the linear combination, allows to prevent from unseen events. In this section, we test the same combination as in the previous experiment, but each model is smoothed with the back-off principle. Whereas, in the previous section all the models have been interpolated with unigram and zerogram. In our case, after several experiments, we used the absolute discounting method [8].

We propose in the following to combine two models, both are smoothed by using the back-off principle. The first one is a distant n-gram smoothed by lower-order models until zerogram. The second one is a classical n-gram smoothed also with lower-order models. This combination is formulated by equations (2) to (4) for a distant bigram model.

$$P(w_i|w_{i-1}) = \begin{cases} fr^*(w_i|w_{i-1}) \text{ if } N(w_{i-1}w_i) > 0 \\ \alpha_1(w_{i-1})P(w_i) \text{ elsewhere} \end{cases} \tag{2}$$

$$P(w_i|\cdot w_{i-2}) = \begin{cases} fr^*(w_i|w_{i-2}) \text{ if } N(w_{i-2} \cdot w_i) > 0 \\ \alpha_2(w_{i-2})P(w_i) \text{ elsewhere} \end{cases} \tag{3}$$

$$P(w_i|w_{i-2}w_{i-1}) = \lambda P(w_i|w_{i-1}) + (1-\lambda)P(w_i|\cdot w_{i-2}) \tag{4}$$

where α_1 and α_2 are normalization terms. fr^* is a discounting factor. We use the following notations: formula (2) defines the model $\mathbf{b\_u\_z}$, (3) defines the model $\mathbf{db\_u\_z}$ and (4) defines the linear combination $(\mathbf{b\_u\_z})\bullet(\mathbf{db\_u\_z})$. The classical linear combination is noted by $\mathbf{db}\bullet\mathbf{b}\bullet\mathbf{u}\bullet\mathbf{z}$. In the same way, $\mathbf{t}$ denotes the trigram model, and $\mathbf{dt}$ denotes the distant trigram model.

Table 2 presents the performance of the proposed models. We can conclude that our adaptation of the back-off principle leads to better results than the linear combination. Finally, the contribution of a distant bigram model amounts to 7.9% in comparison to the baseline. In the same way, the contribution of a distant trigram model amounts to 11.6%.

Combination	Perplexity
$\mathbf{b}\bullet\mathbf{u}\bullet\mathbf{z}$	132.4
$\mathbf{db}\bullet\mathbf{b}\bullet\mathbf{u}\bullet\mathbf{z}$	123.9
$(\mathbf{b\_u\_z})\bullet(\mathbf{db\_u\_z})$	121.9
$\mathbf{t}\bullet\mathbf{b}\bullet\mathbf{u}\bullet\mathbf{z}$	97.8
$\mathbf{dt}\bullet\mathbf{t}\bullet\mathbf{b}\bullet\mathbf{u}\bullet\mathbf{z}$	95.2
$(\mathbf{t\_b\_u\_z})\bullet(\mathbf{dt\_db\_u\_z})$	86.5

Table 2: Contribution of distant bigram and distant trigram models by using or not the back-off principle.

4. Efficient combination of distant n-gram

It is known that constant weights for interpolated models don't lead to the best performance. In order to improve the model one could use dependent history weights [8]. The model presented in the previous section is now enhanced by assigning a weight depending on the sub-history used by the combined model. More formally, in order to combine K models, $M_1, \ldots, M_K$, a set of weights $\alpha_1, \ldots, \alpha_K$ is defined and the combination is expressed by:

$$P(w|h) = \sum_{i=1}^{K} \alpha_i(h)P_i(w|h) \tag{5}$$

The problem, here, is that the weights are estimated using a finite development corpus. This corpus is not sufficient to estimate a huge number of parameters. The idea is then to classify histories and to set a weight to each class. Let $\mathcal{C}(h)$ be the class of the history h. Then (5) can be rewritten as:

$$P(w|h) = \sum_{i=1}^{K} \alpha_i(\mathcal{C}(h))P_i(w|h) \tag{6}$$

4.1. Classification of sub-histories

A classification based on the entire history does not take into account the features of each part which constitutes it. What we propose in the following is to break the history into the several parts or sub-histories used by the combined models. Each sub-history is associated to a specific model. Each sub-history is analyzed in order to estimate its importance in terms of prediction and it is then put into a class. Such a class is directly linked to the value of the sub-history frequency: this class gathers all sub-histories which have approximately the same frequency. The final objective is to classify the whole history by combining the sub-histories or more exactly by combining classes. EM algorithm is then used in order to find the best weight for each model and each history.

For example, let $h = w_1 \ldots w_{i-2}w_{i-1}$ be a history. The distant bigram model uses the part w_{i-2} and the classical bigram model uses the part w_{i-1}. These sub-histories correspond respectively to the classes $\mathcal{C}(w_{i-2})$ and

$\mathcal{C}(w_{i-1})$. Then, to combine these two models, we propose to define the class of the history h as to be the couple $\mathcal{C}(h) = (\mathcal{C}(w_{i-2}), \mathcal{C}(w_{i-1}))$. In our experiments, the set of sub-histories is split into MAXCLASS classes. In order to find the best number of classes, we experimented several values of MAXCLASS.

Figure 6 plots a history dependent linear combination of distant and classical bigram models. Whereas, figure 7 plots a history dependent linear combination of distant and classical trigram models. Figure 6 shows that the perplexity decreases and reaches 115.4 when the number of classes increases until 8000 classes and the performance is worse for more classes. We can achieve the same conclusion for trigram except that the optimal number of classes is 4000 and the perplexity reaches 85.2. When these models are compared to the baseline ones (bigram respectively trigram) we achieve an improvement for both about 12.8%. This original way to combine models leads to an improvement of 5.3% in comparison to a classical linear combination.

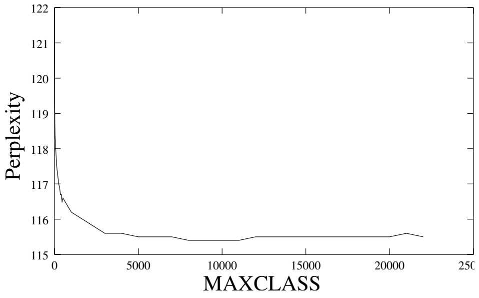

Figure 6: Performance of the history dependent linear combination of a distant bigram and bigram models.

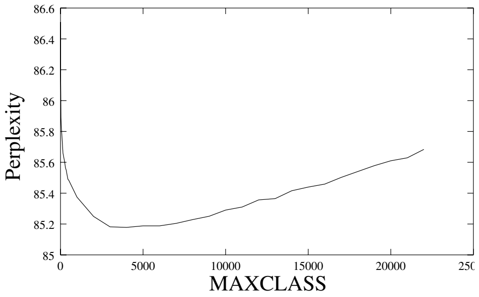

Figure 7: Performance of the history dependent linear combination of a distant trigram and trigram models.

5. Conclusion

In this paper, we present a study about distant n-gram models and an original efficient method to combine them. We quantify their performances using the perplexity measure. In our knowledge, such a study is achieved for the first time. To discover the sub-histories which are truly useful for prediction, we showed that distant n-gram models are well suited.

The contribution of distant bigram and trigram models is significant. This improvement is due to two reasons. The first one concerns the adaptation of back-off principle for both distant and classical models. The second one concerns a relevant choice of the linear combination parameters in order to combine the models. These parameters are dependent on each history. More precisely, these parameters are dependent on each sub-history used by each model. This new model achieves an improvement of 12.8% in terms of perplexity. Beyond this improvement, we show in this paper that it is possible to estimate more efficiently the linear combination parameters by taking into account the history's usefulness for each model. This method outperforms the classical linear interpolation by 5.3%. In order to make this method relevant in a speech recognition system, it has to be integrated in the decoding algorithm. This necessitates an adaptation of Viterbi's algorithm. Our objective is now to adapt this algorithm.

6. References

[1] D. Langlois and K. Smaïli, "A new distance language model for a dictation machine: application to maud," in *EUROSPEECH*, 1999, vol. 4, pp. 1779–1782.

[2] R. De Mori and M. Federico, "Language model adaptation," in *Computational models of speech pattern processing*, K. Ponting, Ed., vol. 169 of *F: Computer and Systems Sciences*, pp. 280–303. 1999.

[3] X. Huang, F. Alleva, H.-W. Hon, M.-Y. Hwang, K.-F. Lee, and R. Rosenfeld, "The SPHINX speech recognition system: an overview," *Computer Speech and Language*, vol. 2, pp. 137–148, 1993.

[4] R. Rosenfeld, "A maximum entropy approach to adaptative statistical language modelling," *Computer Speech and Language*, vol. 10, pp. 187–228, 1996.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistics Society*, vol. 39, no. 1, pp. 1–38, 1977.

[6] D. Langlois, *Notions d'événements distants et d'événements impossibles en modélisation stochastique du langage : application aux modèles n-grammes de mots et de séquences*, Ph.D. thesis, Université Henri Poincaré, Nancy 1. France, 2002.

[7] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE transactions on acoustics, speech, and signal processing*, vol. ASSP-35, no. 3, pp. 400–401, 1987.

[8] M. Federico and R. De Mori, *Spoken dialogues with computers*, chapter Language Modelling, pp. 199–230, Academic Press, 1997.