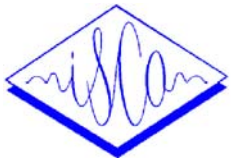

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A NEW BASED DISTANCE LANGUAGE MODEL FOR A DICTATION MACHINE: APPLICATION TO MAUD

D. Langlois and K. Smaïli
LORIA/INRIA-Lorraine
BP 239 54506 Vandoeuvre-Lès-Nancy France
E-mail: {David.Langlois, Kamel.Smaili}@loria.fr
Tel: (33) 03-83-59-20-74, Fax: (33) 03-83-41-30-79

ABSTRACT

This paper deals with the use of a stochastic language model based on the split of the words history into *d* words where *d* is the length of the history.
One of our aims is to modelise the semantic and syntactic relationships between words. This model can be considered as a first step for this goal.
We experimented our model through the Shannon game (on 10 000 truncated sentences) and implemented it in *MAUD*, our dictation machine. Tests on *MAUD* have been done on 300 sentences pronounced by several women and men. This model predicts more words (in the Shannon game) than any other methods we developed before in our team. However, these models are sophisticated in contrast to the one we describe. Moreover, when including unknown words, the results are better than the model ones we presented in a recent work in terms of mean rank, ranks from 1 to 5 and perplexity.
This work has needed to use two interpolation methods inspired from Markov model. Also, we discuss the problem of the unknown word modelling.

INTRODUCTION

Stochastic language models still come up against the problem of sparseness data. The first reason which urged us to work on a new model is precisely to get round this difficulty.
A *n*-gram model must be able to give a word utterance probability for each word in the vocabulary after each possible history. An history is made up of *n-1* words. If the vocabulary contains *V* words, the total history number is V^{n-1}. As *V* words can theoretically appear after one history, the total number of parameters to estimate is V^n. For a trigram model and a vocabulary of 20 000 words, this number is equal to 8E12 ! Even without forgetting that all these words sequences will never appear in the language ("the the the" for example), it's clearly impossible to estimate accurately all the parameters, whatever the training corpus size is.
To prevent from this problem, our model does not consider the history as a single block (maybe too much rare), but it deals with each isolated words composing it. These words may appear more often than the history.
The second reason is that our aim is to work on a stochastic syntactic and semantic level in language modelling.
Because of the data sparseness and instead of studying the relationship between a word and a single "block" history, we can look to the relationship between the current word and each word of its history. For instance, in "the third man", the utterance of "man" as a *NOUN* behind "the third" depends on "the" as an *ARTICLE* and separately on "third" as an *ADJECTIVE*. So it would be convenient to be able to study each of those syntactic agreements one by one. Moreover, at the semantic level, let us study this example: "the Empire State Building, this fabulous tower". There is a strong semantic relationship between "tower" and "building", and so, we could guess that the only presence of "building" in the history increases the utterance probability of "tower". So, to deal with a syntactic and semantic level in language modelling, we must be able to consider individually each of the words of the history.
In this paper, we describe formally the model and discuss the number of parameters it requires to be trained. Then, we present our experiments: test through the Shannon game [3], and in a real situation by including it in the dictation machine *MAUD* [1]. And finally, we give a description of our future work.

THE DISTANCE LANGUAGE MODEL

Our model, inspired from [2], is defined as a linear interpolation between distant bigrams, a bigram and an unigram models. To take into account an history of *n*-1 words, as a *n*-gram model, we use *n-1* distant bigram models, one for each distant word in the past. By this way, the model takes into account all the history words of the *n*-gram model. Because of that, we call it a *d-n*-gram model.

Formal definition

In the sequence of words $w_1w_2w_3\ldots w_i$, a distant *d* bigram model defines the probability utterance of the word w_i after the word w_{i-d} by:

$$P(w_i|w_{i-\delta}) = \begin{cases} \dfrac{N_\delta(w_{i-\delta}, w_i)}{N(w_{i-\delta})} & \text{if } N(w_{i-\delta}) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $N_d(v,w)$ is the count of times w appears d words after v, and $N(v)$ is the count of words v. These values are estimated from a training corpus.
So, a distant d bigram model considers the occurrence of a word being dependent on the word used d words before it. One can then take into account all the history handling an interpolation between the i-1 distant d bigram models:

$$P(w_i|w_1 w_2 w_3 \cdots w_{i-1}) = \sum_{\delta=1}^{i-1} \alpha_\delta \cdot P(w_i|w_{i-\delta})$$

where the parameters α_d vary from 0 to 1 and sum up to 1. Those parameters are estimated from a development corpus.
In practice, we have limited the maximum value of d, because our aim was to compare our model with the well known limited history n-gram and n-class models. This gives:

$$P(w_i|w_{i-n}\cdots w_{i-1}) = \alpha_0 \cdot \frac{1}{V} + \alpha_1 \cdot P(w_i) + \sum_{\delta=1}^{n} \alpha_\delta \cdot P(w_i|w_{i-\delta})$$

where $P(w)$ is the unigram probability of the word w in the training corpus. V is the size of the vocabulary. We add the parameter α_0 to prevent from the null probability of a word present in the vocabulary, but, which have never been met in the training corpus. The probability of a words sequence is then calculated as one would do with a n-gram model.

Discussion about the number of parameters

In the introduction, we have explained that the n-gram model requires too many parameters to be trained. Now, even if a words sequence $w_1w_2w_3\ldots w_{i-1}$ is rare, one can notice that each word of the history may has been appreciably more frequent in the training corpus, at the same distance from the current word. Obviously, each word at distance d has been met in several different histories. So, instead of dealing with the entire history as a single block, we guess it could be better to consider individually each of its words. Our model is dedicated to this idea: each distant bigram model involves with one of the words history. Then, the interpolation of these models allows to take into account the whole history.
A distant bigram history is made up of one word. Moreover, it must give an utterance probability after this history for each word of the vocabulary. In consequence, for a vocabulary size V, the number of parameters is V^2.
A d-n-gram model uses n-1 distant bigram models. To this value, we must add the V unigram parameters. Thus, the total number of parameters is $V+(n-1).V^2$. This value is considerably smaller than a n-gram model's.

Estimation of the interpolation parameters

To estimate the $(n+2)\alpha_i$ interpolation parameters, we used an adapted method from the Markov models (see [4]).
The first, noted A, is based on a detailed account in a development corpus (D), of the presence of the same events in this corpus and the training one (T). The estimation algorithm is:

Algorithm estimation parameters for method *A*
$\alpha_0 = 0$
$\alpha_1 = 0$
for each distant d bigram model
  $\alpha_{d+2} = 0$
for each $w_{i1}w_{i2}\ldots w_{in}$ in D
  for each distant d bigram model
    if distant bigram $(w_{i(n-d)},w_{in})$ has been met in T at distance d
      $\alpha_{d+2} = \alpha_{d+2} + 1$
  if no distant bigram has been met in T
    if w_{in} has been met in T
      $\alpha_1 = \alpha_1 + 1$
    else
      $\alpha_0 = \alpha_0 + 1$
normalise to 1 the α_i

Note that the vocabulary contains a word *UNK*, the unknown word. If a word met in T or D isn't in the vocabulary, it is replaced by *UNK*.
A gives more weight to distant bigram models which are often concerned by events in D. Note that the unigram model parameter is up to date only if no distant bigram model parameter has been.
The second (B) interpolation method favours the models which give, on average, the highest utterance probability to the word to predict. As A, the algorithm doesn't up to date the unigram parameter when it is possible.

Algorithm estimation parameters for method *B*
$\alpha_0 = 0$
$\alpha_1 = 0$
for each distant d bigram model
  $\alpha_{d+2} = 0$
for each $w_{i1}w_{i2}\ldots w_{in}$ in D
  if at least one distant bigram $(w_{in},w_{i(n-d)})$ has been met in T with $(1<d<n-1)$
    $dm = \text{argmax}_d(P(w_{in}|w_{i(n-d)}))$
    $\alpha_{dm+2} = \alpha_{dm+2} + 1$
  else if $P(w_{in}) > 0$
    $\alpha_1 = \alpha_1 + 1$
  else
    $\alpha_0 = \alpha_0 + 1$
normalise to 1 the α_i

EXPERIMENTS

We evaluated the d-n-gram model through the Shannon game and implemented it, in a selective processing in our dictation system: *MAUD*.
The vocabulary used is made up of 20 000 words, the corpus is composed of two years of *Le Monde*, a French newspaper (42 Million words). Twenty-two months are used for the training corpus. The remainder is devoted to the development corpus.

The vocabulary contains an unknown word, noted *UNK*. Each word in the corpus, not present in the vocabulary is likened to *UNK*.

The Shannon game

Description

This evaluation protocol (see [3]) is derived from the Shannon work on the capacity of people to guess missing letters from an unfamiliar text. In [3], the evaluation through the Shannon game consists in measuring the performance of the model in predicting the word which comes just after a truncated sentence. The test is made with 10 000 truncated sentences.

For each truncated sentence, the evaluated model proposes the more probably 5 000 words to appear after it. The model gives a score from 0 to 1 to each proposition. The 5 000 words scores must sum up to less than 1. If the real word to be found is not in the list, the model assigns a low bet which depends on the scores of the 5 000 hypothesis.

The evaluation step compares each 5 000 propositions set with the word which really follows the beginning of the sentence and gives the results following:

- **Number UNK**: is the unknown words number over all the real words to guess,
- **Number words in list**: is the number times the words to guess are present, over all the truncated sentences,
- **Mean rank when in list**: is the mean rank calculated over all the truncated sentences,
- **Number at rank 1**: is the number times the word to guess is proposed at the first rank,
- **Number at ranks 1 to 5**: is the number times the word to guess is proposed in five first ranks,
- **Shannon Perplexity (*PPSh*)**: is a Shannon game adapted measure of the perplexity,
- **Perplexity (*PP*):** the classical perplexity estimated on the corpus from which the truncated sentence has been extracted.

Results

We compared *d-3*-gram, *d-4*-gram and *LMA*, a model presented in [3] which is based on a linear combination of a *3*-gram and a *3*-class models. As for *LMA'*, it is a recent improvement of *LMA*. In this version the *n*-class model is used only to look for the likely candidate classes. When the classes are found, the language model bets on each word of each candidate class by using an interpolated *3*-gram.

As in [3], we give two tables in which we distinguish results with and without unknown words. Moreover, we give results for both the two interpolation methods.

Models	LMA	LMA'	d-3-gram	d-4-gram
Reference words	10 000	10 000	10 000	10 000
PP_{Sh}	—	119	161,07	170,59
Words in list	8 100	9 649	9 840	9 842
Mean rank	238	160	196,61	211,61
Words at rank 1	1 650	2 105	1 569	1 613
Words at ranks 1 to 5	3 446	4 346	3 534	3 420
PP_{is}	437	98	133,52	144,59

Table 1: Comparative results for the 10 000 words (including unknown words) to be found, by using the interpolation method *A*.

Models	LMA	LMA'	d-3-gram	d-4-gram
Reference words	9 382	8 697	8 697	8 697
PP_{Sh}	—	179,74	252,63	268,61
Words in list	7 483	8 346	8 537	8 539
Mean rank	258	186	226,38	243,69
Words at rank 1	1 290	1258	705	608
Words at ranks 1 to 5	2 875	3 095	2 243	2 127
PP_{is}	—	142,99	202,18	220,82

Table 2: Comparative results for the 10 000 words (excluding unknown words) to be found, by using the interpolation method *A*.

Models	LMA	LMA'	d-3-gram	d-4-gram
Reference words	10 000	10 000	10 000	10 000
PP_{Sh}	—	119	145,91	145,97
Words in list	8 100	9 649	9 835	9 847
Mean rank	238	160	183,91	188,12
Words at rank 1	1 650	2 105	1 540	1 534
Words at ranks 1 to 5	3 446	4 346	3 575	3 567
PP_{is}	437	98	120,92	121,48

Table 3: Comparative results for the 10 000 words (including unknown words) to be found, by using the interpolation method *B*.

Models	LMA	LMA'	d-3-gram	d-4-gram
Reference words	9 382	8 697	8 697	8 697
PP_{Sh}	—	179,74	225,68	225,63
Words in list	7 483	8 346	8 532	8 544
Mean rank	258	186	211,74	216,57
Words at rank 1	1 290	1 258	740	719
Words at ranks 1 to 5	2 875	3 095	2 288	2 275
PP_{is}	—	142,99	180,61	181,52

Table 4: Comparative results for the 10 000 words (excluding unknown words) to be found, by using the interpolation method *B*.

The first positive result of our model is its capacity to predict more words than any other model we developed before. In comparison with *LMA* and *LMA'*, this improvement is about 20% and 2% respectively.

Moreover, the *d-n*-gram model is better than *LMA* in terms of mean rank, words at ranks 1 to 5 and perplexity, in experiment including unknown words.

In the other hand, the performances are worse when excluding them (tables 2 and 4). The reason is words

probabilities are lower than the *UNK* one. It is not surprising: in the training corpus, different unknown words are numerous. In fact, *UNK* is not a word, but a class: the class of all unknown words. This set is very large ! The problem is that, nevertheless the model considers unknown class as an ordinary word in the vocabulary. To prevent from that, we defined *UNK* as a class. We estimated N_{UNK} (the number of distinct unknown words in the training corpus) ; and then, the probability of an unknown word from the class *UNK*, is defined as the uniform distribution $(N_{UNK})^{-1}$. By this modelling of *UNK*, the *d-n*-gram model is less sensitive to the presence or not of unknown words in test data. Moreover, in this last experiment our model gives better results when excluding unknown words.

Concerning the performances of the two interpolation methods, the perplexity, the mean rank and the number words at the five first positions are improved when using interpolation method *B*. These results were expected for perplexity because *B* is dedicated to favour the distant bigram model which maximises in average the probability of the current word.

We can notice that when *n* is high the model gives better results in term of discovered words. Unfortunately, this model predicts more words without increasing the first ranks.

Implementation in MAUD

We implemented a *d-3*-gram as a selective model in a words lattice generated by *MAUD*[1]. *MAUD* is a 20K words continuous dictation system using an acoustic model trained on the french *BREF* database. Each phoneme is modelled by an *HMM2*. The acoustic models used to construct the words lattice are independent (the ones used in [2] were context dependent). This words lattice is obtained by the combination of an acoustic and a simple bigram language models.

We tested a *d-3*-gram model (interpolation *B*) on 300 sentences uttered by female and male speakers. The *d-3*-gram operates on the words lattice proposed by *MAUD*. For each sentence, the model proposes one hypothesis. The evaluation on the 300 sentences test-corpus is made using *SCLITE*[1] alignment tool by comparing the reference sentence to the one proposed by our model.

Before evaluation we calculated a superior limit value of the accuracy word recognition given by lattices. By overestimating this value, in mean, the accuracy word does not exceed 80,25%. Therefore, it means that the performance of our model can not exceed this result. Moreover, a qualitative study of lattices showed us that they were strongly noised.

Finally, the acoustic models are not context dependent. For all these reasons, our accuracy word recognition is 55,4% on 300 sentences. To improve this result, the *d-n*-gram, should absolutely run during the acoustic recognition, in place of the bigram model.

[1] *SCLITE* is included in the Speech Recognition Scoring Toolkit distributed by the *SNLP* group (Spoken Language Natural Processing)

CONCLUSION

In this article, we have implemented a stochastic language model based on a linear interpolation of distant bigram models. We used two different interpolation methods.

Our model needs really fewer parameters to be trained than a *n*-gram model. Moreover, it allows dealing with the entire history but one word by one. As a consequence, our aim is to use it to model syntactic and semantic relationship between far words.

We test our model through the Shannon game with 10 000 truncated sentences extracted randomly from 6 years of *Le Monde Diplomatique* (A French newspaper). This model always recognises more words in list than any other models developed before in our team. It is also better than *LMA*, when including unknown words, for all measures except words at first rank.

The less good performances when excluding unknown words urged us to work on *UNK* modelling. First results are encouraging. We are still studying this problem.

Also, we are investigating why, when we increase the distance, the model improves words recognition but decreases their rank.

Concerning the implementation in *MAUD*, our model should show all its potential if it is included in the prediction stage, instead of in the post selective processing.

REFERENCES

[1] D. Fohr, J.-P. Haton, J.F. Mari, K. Smaili, I. Zitouni *"Towards an Oral Interface of Data Entry: The MAUD System"*, 3rd European Research Consortium for Informatics and Mathematics Workshop on "User Interfaces for All", 1997.

[2] X. Huang, F. Alleva, H-W. Hon, M-Y. Hwang, K-F. Lee, R. Rosenfeld "The SPHINX Speech Recognition System: an Overview", Computer Speech and Language, VOL 2, PP 137-148, 1993

[3] M. Jardino, F. Bimbot, S. Igounet, K. Smaïli, I. Zitouni, M. El-Beze "*A First Evaluation Campaign For Language Models*", Proceedings of the 1st International Conference on Language Resources and Evaluation, PP 801-805, Vol 2 Granada, 1998.

[4] F. Jelinek, R. L. Mercer, S. Roukos "*Principles of Lexical Language Modelling for Speech Recognition*", Advances in Signal Processing, PP 651-699, Marcel Dekker, 1992.