Discarding Impossible Events from Statistical Language Models

A. Brun, D. Langlois, K. Smaïli and J. P. Haton

LORIA BP 239
54506 Vandœuvre-les-Nancy, FRANCE
{brun, langlois, smaili, jph}@loria.fr

ABSTRACT

This paper describes a method for detecting impossible bigrams from a space of V^2 bigrams where V is the size of the vocabulary. The idea is to discard all the ungrammatical events which are impossible in a well written text and consequently to expect an improvement of the language model. We expect also, in speech recognition, to reduce the complexity of the search algorithm by making less comparisons. To achieve that, we extract the impossible bigrams by using automatic rules. These rules are based on grammatical classes. The biclass associations which are ungrammatical are detected and all the corresponding bigrams are analyzed and set as possible or impossible events. Because, in natural language, grammatical rules can have exceptions, we decided to manage for each of the retrieved rules an exception list.

1. INTRODUCTION

A language model is a fundamental component of a speech recognizer whose role is to estimate the hypotheses supplied by the speech decoder. In a statistical language model, each hypothesis is assigned a prior probability computed from traininig corpora. These probabilities depend obviously on the amount of data used for the estimation. For a vocabulary of V words, V^2 probabilities have then to be estimated in order to cover all events. Even those which do not occur in the training corpora, by smoothing or discounting the basic model. In the discounting process, the relative frequencies of seen events are discounted and the gained probability mass is distributed over the unseen events. In our experiment, we use a $20K$ vocabulary, the space of events thus covers 4.10^8 possible values. Actually, in the space of events, many events never happen in a dictation task, contrary to spontaneous speech. Indeed, dictating a text requires a correct grammatical delivery. Obviously, this is no longer true in spontaneous speech. The purpose of this paper is to present a new method which makes the difference between really unseen events and impossible events. The main objective is to purge the statistical language models from the impossible events and to improve the probability distributions of the possible events. Consequently, impossible events have to be found and preferably by automatic procedures.

The paper is organized as follows. In section 2, we propose to distinguish between unseen and impossible events in a speech corpus. Section 3 presents the polyclass model. Section 4 deals with the automatic extraction of rules for collecting impossible events, and section 5 with the integration of those rules in a language model. Finally, section 6 presents and discusses some experimental results.

2. UNSEEN EVENTS VERSUS IMPOSSIBLE EVENTS

Unseen events are the ones that have not been met in the training data, due to a lack of samples. Discounting methods handle this type of sampling error by reducing the probability of unreliable estimates made from observed frequencies, and by distributing this probability mass among words that did not occur in the training corpus [2]. In certain methods, the discounting is only applied to words which occur less than n times. In other methods, like the absolute or linear discounting, all the counts are discounted [3].

In classical methods, the space of events is divided into two parts: seen and unseen events. The idea we defend in this paper is that the space of events is split into three parts: seen events, unseen events and impossible events. If Ω is the space of events, we can write: $\Omega = S + U + I$, where S is the sub-space of seen events, U is the sub-space of unseen events and I is the sub-space of impossible events. In conventional language models, I is included in U whereas, impossible events will never occur in a well written text or in a dictation task. Our purpose is to discard all the noisy impossible events from the probabilistic space. The justification is as follows: anyone speaking a language possesses, a large amount of implicit knowledge about the language structure. Therefore, a fixed history cannot be followed by any other word of a dictionary [5].

3. DISCARDING IMPOSSIBLE EVENTS BY USING POLYCLASS MODEL

Discarding noise from unseen events will probably improve the language model; the probability assigned to each possible event will be increased. To discard impossible events, we decided in a first step to find automatically some syntactic rules. For that, we developed a polyclass model which can be compared to a n-gram model, except that the vocabulary is a class vocabulary and each unit of the corpus is a syntactic class. This model is used to extract rules modelizing the impossible events. Before explaining the process of discarding, let us give little details about the polyclass model.

In this approach, we estimate the probability $P(c_i/h)$ where hc_i is a sequence of classes which corresponds to a sequence of words $h^{'}w_i$. To do that, each word of the training corpus has to be tagged. Consequently the dictionary of the application needs a syntactic field for each entry. From the eight elementary grammatical classes of French, we built up about 230 classes including punctuation [7]. Each word can belong to several classes. These classes have been constructed by hand in accordance with linguistic criteria. The probability $P(c_i/h)$ is estimated by a relative frequency for the sequence of classes. In our approach the probabilities computed for this model are not relative to a dictionary word but to a dictionary of classes. To collect statistics, we labeled a small text by hand and we tagged automatically a corpus of 0,5 million of words extracted from *L'Est Républicain*, a French newspaper, by using a dictionary of $230K$ words split into 230 grammatical classes. This tagging has been checked by hand and the automatic labeling errors have been corrected.

4. AUTOMATIC RULES

The tagged corpus can be considered as a knowledge source which can be used to extract grammatical rules. To do that, we follow an idea which is inspired from the grammatical rules of French. Indeed, a syntactic class cannot be followed by any other syntactic classes. For instance, if a verb is followed by another one, the second must be in an infinitive form or after the preposition *à*, the immediate following verb must also be in an infinitive form.

In our experiments, we tried to discover automatically rules which can be used as a generative process of impossible events. Several rules have been used to retrieve a great number of impossible bigrams, some of them are listed below.

4.1. Zero frequency biclasses (R1)

Since the number of classes used is small, we can expect that, by using biclasses, most successions of two classes have occurred in the training corpus. Consequently, when the frequency of a biclass is equal to zero, we can consider this biclass as an impossible event. From all the 0-frequency biclasses, we find out all the impossible bigrams. To achieve that, we have to pay attention to the pair of words which will be considered as impossible events. Since a word can belong to several classes, a pair of words is considered as an impossible event, if the frequency of each biclass which can be constructed by assigning to each word its classes, is equal to zero. In other words:

$$(w_i, w_j) \in I\, if \quad \begin{array}{l} \forall C(w_i) \wedge \forall C(w_j) \Rightarrow \\ f(C(w_i), C(w_j)) = 0 \end{array} \qquad (1)$$

where $C(x)$ denotes the class of the word x, and $f(s,t)$ denotes the frequency of the biclass (s,t). To avoid the classes sparse data, we decided to take into account only those classes for which the frequency is high. In other words: a biclass (h, C_j) is a candidate for generating impossible events if and only if condition (1) is checked, and the following inegality is verified

$$\sum_C f(h, C) > \eta \qquad (2)$$

where η is a threshold determined experimentally. This rule assures that if the left context has been met very frequently then a missing class could be considered as an impossible class in such a context. First experiments using rules (1) and (2) show that the number of impossible bigrams is about $300K$ which constitutes only $0,0075\%$ of space Ω. Obviously, subtracting this small number of bigrams from the model is not sufficient and will not change efficiently the model. To deal with this problem, we decided to investigate other issues.

4.2. Infrequent biclasses (R2)

We remarked that a huge number of real impossible bigrams could be generated by low frequency biclasses. In fact, all the low frequency biclasses are either due to labelling errors or to infrequent syntactic structures. This last case is rare because the number of classes is small and the training corpus is very important. Nevertheless, we have to take into account the phenomena of unseen or infrequent biclasses.

This rule consists in generating impossible bigrams from all the statistically unreliable biclasses. It yields $0,27\%$ impossible bigrams.

4.3. Selecting impossible events by mutual information (R3)

This rule assumes that a biclass which has a very low class mutual information (MI) and whose count is above a given threshold is considered as a potential candidate for generating impossible bigrams. The justification of this rule is as follows: a low MI means that the amount of information provided by the first class on the second is small. In other words, the relationship between the two classes is weak. However, a weak MI does not necessary mean that the biclass is impossible, or more precisely that the corresponding bigrams are impossible. In order to take this into account, we decided to attach an exception list to each biclass, and we generalized this principle to all the rules presented in this paper. Thus, in the training step, we collect all the impossible biclasses by following the above constraints. In the development step, a corresponding corpus is used to build exception lists: if a bigram is met and its corresponding biclass is considered as impossible, the bigram is added to the exception list. At the end, the size of each exception biclass list should not be beyond a fixed value. Otherwise, the status of the biclass becomes definitively possible. Experiments show that the mean average of an exception list is about 115 bigrams per biclass.

4.4. Selecting impossible bigrams by using phonology (R4)

This rule, in opposition to the preceding ones looks for impossible bigrams without passing over the biclasses. The forbidden phonological association between two successive words are found out, and all the concerned bigrams become impossible. For instance, in French a word ending by the vowel $/a/$ cannot precede a word beginning with the same vowel (this rule is not true in all the cases). For example, we cannot say *ma armoire*.

Table 1 sums up the rate reduction of the space events in terms of impossible bigrams and biclasses. We can notice that the percentage of the impossible biclasses is very important in comparison to impossible bigrams.

	IBIGR%	IBICL%
R1	$0,075$	$65,8$
R2	$0,27$	$78,8$
R3	$10,5$	$80,0$
R4	$10,6$	$80,0$

Table 1: Rate reduction of the space events by different rules

5. MODEL ESTIMATION

In the previous sections, we described methods allowing to collect the impossible bigrams. Now, we give details about how to incorporate them in a language model. To take into account the unseen events, classical methods consist in discounting the original counts and the gained probability mass is redistributed over the less frequent and/or the unseen events. To train our model, we decided to use the back-off [2] and the absolute discounting methods [3]. The idea of impossible bigrams is to assign a zero probability to those bigrams which are declared as impossible. To do that, we retrieve from the unseen events those which are declared impossible by our method. Then, the initial probability mass which was reserved to those impossible events is redistributed among the less frequent events. Consequently, the probability of a word given an history (which is reduced to one word in our experiments), in accordance to Katz method, is defined as $P(w_i/h) =$

$$\begin{cases} \frac{n(h,w_i)}{n(h)} & \text{if } n(h,w_i) > k \\ d_r \frac{n(h,w_i)}{n(h)} + Q_h & \text{if } 0 < n(h,w_i) \leq k \\ \alpha(h)P(w_i/h^{-1}) & \text{if } n(h,w_i) = 0 \\ & and\ (h,w_i) \notin \{I\} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

with

$$Q_h = \frac{\sum_{w_i:(h,w_i)\in\{I\}} \alpha(h)P(w_i/h^{-1})}{\sum_{w:0<n(h,w)\leq k} 1} \tag{4}$$

The other parameters have the same meaning as what indicated in the litterature.

Our first experiments with this language model concerned the perplexity. Unfortunately, the test corpus contained few bigrams which have been declared as impossible by our method but which were actually present in the text. Some of these bigrams are grammatically wrong, they are due to errors which can be found in newspapers. For instance, we found duplicated words, accent missing, punctuation missing, etc. Other bigrams were correct, but our model set them as impossible. When we analyze these bigrams, we can notice that most of them concern particular language constructions. The amount of wrong impossible bigrams on a test corpus of 2 million words is about $0,1\%$. In concrete terms, perplexity computation is impossible, even if only one wrong impossible bigram is present during the test. So, we modified formula (3) in order to take into account the wrong impossible bigrams as follows: $P(w_i/h) =$

$$\begin{cases} \frac{n(h,w_i)}{n(h)} & \text{if } n(h,w_i) > k \\ d_r \frac{n(h,w_i)}{n(h)} + bQ_h & \text{if } 0 < n(h,w_i) \leq k \\ \alpha(h)P(w_i/h^{-1}) & \text{if } n(h,w_i) = 0 \\ & and\ (h,w_i) \notin \{I\} \\ \frac{(1-b)Q_h}{\sum_{w_i} \delta(h,w_i)} & \text{otherwise} \end{cases} \tag{5}$$

where

$$\delta(h,w_i) = \begin{cases} 1 & \text{if } (h,w_i) \in \{I\} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The constant b is determined experimentally and is used to take a small part from the probability mass assigned to impossible bigrams, in order to avoid the zero probability problem. To keep the added gained mass probability almost unchanged, the value of b is chosen close to 1.

6. EXPERIMENTAL RESULTS AND DISCUSSION

In order to evaluate the effectiveness of the proposed model, an experiment was conducted on test perplexity and with our MAUD speech recognition system [1]. In both cases the vocabulary is made up of $20K$ words. The test corpus contains $2M$ words extracted from the French newspaper *Le Monde*. The results of the speech recognition system have been analyzed with the *SCLITE* toolkit [4]. Results show that the perplexity has been improved slightly (3 points) over the baseline model and the word error has been decreased but not in a statistically significant ways. This weak performance could be explained in several way:

- The number of impossible bigrams is still very low: only 10% have been discovered. That means after each word $18K$ words are possible. It seems to us that this number is very high in comparison with what a human being can pronounce after a specific word. The weaker the number of impossible bigrams is, the lower the redistributed probability mass is.
- Several impossible bigrams found in the test corpus are due to a bad punctuation of the text. For instance, our model declared the pair words *près dès* (*near since*) as impossible, but in the test corpus this bigram has been met in a sentence for which a comma was missing.
- The scope of impossible bigrams is limited. After pronouncing a phrase, the number of possibilities in terms of words will decrease with the length of history. Consequently, longer histories have to be used.

7. CONCLUSION

In this paper, we presented a new approach based on the suppression from the space of events those which are grammatically impossible. A polyclass model has been used to extract automatic rules. These rules generate impossible bigrams which are discarded from the language model. Experiments shown that 80% of biclasses have been considered as impossible, which is very encouraging. Unfortunately, this is insufficient, since this rate, corresponds in fact to only 10% of impossible bigrams. All the reasons cited above make the performance of the method not as good as we would like them to be. However, the results give a clear indication that it is possible to do better, and the key of the problem is to discover automatically other impossible bigrams. We pursue several tracks which all have the same aim: increasing the number of impossible events. The first one consists in using another set of classes built automatically. We proved in [6] that an adapted classification gives better results in terms of perplexity. Another way consists in using larger impossible events which will be probably in better concordance with the idea presented in this paper. To do that, we will adapt the variable length sequences algorithm developped in [8] in order to detect automatically impossible phrases.

8. REFERENCES

1. D. Fohr, J.P. Haton, J.F. Mari, K. Smaïli, and I. Zitouni. Towards an oral interface for data entry: The maud system. In *3rd ERCIM Workshop on "User Interface for All" (ERCIM: Europea n Research Consortium for Informatics and Mathematics)*, Obernai, France, November 1997.
2. 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*, 35(3):400–401, 1987.
3. H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.
4. D. Pallett, J. Fiscus, W. Fisher, J. Garofolo, B. Lund, and M. Prysbocki. 1993 benchmark tests for the arpa spoken language program. In *Proc. ARPA*, pages 49–79, 1994.
5. C.E. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, 30:50–64, 1951.
6. K. Smaïli, A. Brun, I. Zitouni, and J.P. Haton. Automatic and manual clustering for large vocabulary speech re cognition: A comparative study. In *European Conference on Speech Communication and Technology*, Budapest, Hongrie, Septembre 1999.
7. K. Smaïli, I. Zitouni, F. Charpillet, and J.P. Haton. An hybrid language model for a continuous dictation prototype. In *Proceeding of the European Conference On Speech Communication and Technologie*, pages 2755–2758, Rhodes, Greece, 1997.
8. I. Zitouni, J.F. Mari, K. Smaïli, and J.P. Haton. Variable-length sequence language model for large vocabulary c ontinuous dictation machine: The n-seqgram approach. In *Proceeding of the European Conference On Speech Communication and Technologie*, pages 1811–1814, Budapest, Hongrie, Septembre 1999.