A first evaluation campaign for language models

M Jardino, F Bimbot, S Igounet, Kamel Smaïli, I Zitouni, Marc El Bèze

A FIRST EVALUATION CAMPAIGN FOR LANGUAGE MODELS

M. Jardino [1], F. Bimbot [2], S. Igounet [3], K. Smaïli [4], I. Zitouni [4], M. El-Beze [3]

(1) LIMSI/CNRS, BP 133, Bât 508, Univ. Paris-Sud, 91403, Orsay Cedex, France
(2) IRISA, Campus Universitaire de Beaulieu, 35042, Rennes Cedex, France
(3) LIA, BP 1228, Agroparc, 339 Chemin des Meinajariès, 84911, Avignon Cedex 9, France
(4) CRIN/INRIA Lorraine, BP 239, 54506, Vandoeuvre Les-Nancy, France

Abstract

This article describes a comparative evaluation campaign for language models which has been set up by AUPELF-UREF [1], an agency in charge of the promotion of the French language. Three laboratories have participated to the first part of this campaign. The language models have been compared with an original scheme derived from the Shannon game. The results of this evaluation as well as the description of the method and the evaluated language models are presented.

1. Introduction

In order to compare language models out of the site where they have been built, an original scheme has been drawn (Bimbot, 1997). The method, depicted in the section 1 is derived from the Shannon game and based on a gambling approach : the system has to guess the next word to come. The corresponding protocole which have been adopted by each participant is written in the section 2. Three laboratories have participated to this campaign. Their language models are described in (Cerf-Danon, 1991; Jardino, 1996; Smaili et al, 1997). Results and comments are then given as well as the problems encountered.

2. Shannon guessing game and perplexity estimation

Usually, performances of language models are related to the perplexity of a test text but this approach is not appropriate for a comparative evaluation campaign. It would require either that each site compute itself the perplexity value or that the software of each participant be re-implemented within the site in charge of the evaluation. To overcome this difficulty, we have proposed to estimate the perplexity of a small word set, with a gambling scheme inspired by the Shannon game.

Consider the successsive sentence fragments, obtained by discovering progressively each word of the n-word sentence W :

$$W =< s > w_1 w_2 ... w_k ... w_n < /s >$$

For each truncated sentence W_0^{k-1}, $< s > w_1 w_2 ... w_{k-1}$, the probabilistic language model puts a bet $\beta(v_j)$ on each possible vocabulary entry v_j such that :

$$\beta(v_j) = p(v_j / W_0^{k-1})$$

where $p(v_j / W_0^{k-1})$ is the conditional likelihood to find v_j knowing the truncated sentence W_0^{k-1}. Furthermore, the next condition has to be filfulled :

$$\sum_{j=1}^{V} \beta(v_j) = 1$$

where V is the size of the vocabulary.

The sentence perplexity $PP(W)$ is then defined as the inverse of the geometrical mean of the bets :

$$PP(W) = \frac{1}{\sqrt[n]{\prod_{k=1}^{n} \beta(w_k)}}$$

This scheme has been adapted in three aspects.

1 Randomization of the truncated sentences
Instead of predicting successive words in a text, we have chosen the test data W as a set of distinct sentences which have been truncated at a random position. The participant has to predict the next word that comes immediately after each truncated sentence. With this approach, we can obtain a good representativity of the language with a relatively small sample.

2 List of hypotheses
In order to limit again the volume of data to be handled, the number of hypotheses for each truncated sentence has been reduced relatively to the size of the vocabulary, V. Only the most predictible words must be given by each participant with their corresponding bets. Assuming a uniform distribution for the words outside this set, their common bet is:

$$\beta_k = \frac{1 - \sum_{r=1}^{nhyp} \beta_k(v_r)}{V - nhyp} \qquad (1)$$

where nhyp is the number of the most probable hypotheses which has been chosen.

3 Out-of-vocabulary words
In order to deal with an open vocabulary, a certain part of the bets must be reserved for the probability that the word to be predicted is not in the vocabulary. This implies that a particular OOV-word is included in the vocabulary and that each participant must guess this word, in order to avoid a null term in the geometrical mean.

[1] Association des Universités Partiellement ou Entièrement de Langue Française - Université des Réseaux d'Expression Française

3. Protocol

A set of truncated sentences is provided to every participant. For each sentence, the participants must give a list of candidate words. A capital of 1.0 is distributed between the words of the vocabulary and each participant must bet on the word coming just after the truncation. The perplexity is then evaluated outside of the participant's site as the inverse of the geometric mean of the bets placed on the correct words.

At this point, we had to find the best trade-off between perplexity evaluation reliability and the size of the participant's files. We have chosen to reduce the list of hypotheses to 5,000 and to use 10,000 randomly selected sentences in the test set defined below. This was done according to the results published in (Bimbot, 1997). We will check later how much these choices are suitable for the three evaluated models which are more sophisticated than those described in (Bimbot, 1997).

The size of the vocabulary has been fixed to 20,000 words. An efficient language model is charaterized by its hability to reduce the branching ratio from the vocabulary size to a minimum value, the idealistic and smallest value being 1. So, this constraint seems a minimum condition to get fair evaluations.

4. Evaluation

In this preliminary phase, the different tasks have been dispatched among the participants.

The test set has been extracted from *Le Monde Diplomatique* (1990-1996), a corpus that none of the participants has used to train its models. The word segmentation has been processed as explained in (Adda et al, 1997). The punctuation has been removed from the data, but the segmentation in sentences has been kept. It corresponds to a situation close to the one encountered in a speech recognition task. A subset of 10,000 truncated sentences has been randomly selected in the corpus and distributed to each participant.

In response, each participant had to give the 5,000 best candidates selected by his language model for each sentence, in all, 10,000 times 5,000 words with their respective bets.

Then, out of each participant site, the references have been searched among the proposed candidates. The observation rank and the bet put on the reference permit to compare the models in course.

5. Evaluated models

5.1. Language Model A (LM A)

LM A is a stochastic model expanded with a formal grammar which uses both n-classes and n-grams.

The n-class model is based on an interpolated tri-class model built with 233 syntactic classes (including punctuation) and adjusted by hand. This high number of classes has been chosen in order to build a model which is sufficiently predictive and highly selective. In this classification, each word can belong to several classes (up to 4 classes). This involves that some words have to be duplicated in the dictionary if they have more than one syntactic category. The original dictionary is made up of 41 000 entries from which the 20,000 most frequent words of *Le Monde (1987-1988)* have been extracted. The stochastic model is trained over a corpus of 42M words. In a first step, a corpus of 0.5M words has been accurately labelled with this set of classes. Then, the model learned during this step has been used to label automatically each word of the entire corpus, with a modified Viterbi algorithm.

The n-gram component is an interpolated tri-gram computed from the same corpus. The formal grammar is a list of hand written French grammatical rules which are modelled with a unification grammar.

Prediction strategy

To explain the different steps of predicting the n best words in this experiment, it has been assumed that the truncated sentence is $w_1 \ldots w_{k-1} w_k$, the vocabulary classes are $c_1 \ldots c_{233}$ and the vocabulary words are $v_1 \ldots v_N$. Each word v_i belongs to one or several syntactic classes. The prediction is conducted as follows:

- Each word w_i of the truncated sentence has been labelled. For doing that, a home-made labelling tool has been first used, but unfortunately, the results were very bad because this tool is efficient when it disposes of the whole sentence. Therefore, to deal with this problem, it has been decided to take into account only the last two words ($w_{k-1} w_k$) of the truncated sentence. That means that all the classes of the words w_{k-1}, w_k have been kept.
- The 233 classes have been affected to the word to be discovered w_{k+1}. In other words, the word to be discovered can belong to each class. For each class of w_{k-1}, for each class of w_k and for each word w^i_{k+1} of the vocabulary classes, we compute the quantity:

$$Q = P(c_{k+1}/c_k c_{k-1}) * P(w^i_{k+1}/c_{k+1})$$

- In this step the n-gram model is used to rescore w^i_{k+1}. In fact, the positional word model has been used in order to give more weight to the word sequences.
- In the fourth step and for each discovered word, the unification grammar has been used to examine the linguistic validity of each partial sentence $w_1 \bigwedge w_k w^i_{k+1}$. Thus, only the n best words, those which are not eliminated by the unification grammar, have been kept .

5.2. Language Model B (LM B)

LM B is a linear combination of statistical trigram and tri-class based models (respectively 40% and 60%).

The class-based model has been built with the Viterbi algorithm, applied on the tagged corpus of *Le Monde (1992-1993)* and using the BDLEX dictionary (200,000 entries). The model has been built with 101 classes including the unknown word class, each vocabulary word can be tagged with at the most nine classes.

The trigram model has been trained on the newspaper *Le Monde (1992-1993)* (about 35 million words), with the first version of the CMU toolkit, using the Katz back-off method to interpolate unseen events. It takes into account a reduction to 63,667 forms of the BDLEX vocabulary plus several Proper Nouns which have been added later. Finally, 555,942 bigrams and 786,299 trigrams occurring more than 4 times have been used to train the trigram model. The reduction of the size of the vocabulary to 20,000 words has been done dynamically for each truncated sentence through formula 1. The probabilities of the corresponding out-of-hypotheses words have been spread over the 20,000 most predictible words associated with the truncated sentence.

5.3. Language Model C (LM C)

This is a statistical tri-class-based model, built with untagged words. The mapping of these untagged words is realized through an iterative Monte-Carlo process, in order to reduce the distance between the distribution given by a word model and the distribution given by the class model.

The word model is supposed to be entirely described by the set of consecutive words observed in the training text, $\{w_i w_j\}_{TT}$, and by the conditional probabilities :

$$p(w_j/w_i) = \frac{N[w_i, w_j]}{N[w_i]}$$

where $N[w_i, w_j]$ is the frequency of the sequence $w_i w_j$ in the training text, and $N[w_i]$ the frequency of the word w_i.

With the class model, these probabilities are averaged to:

$$q(w_j/w_i) = \frac{N[w_j]}{N[C(w_j)]} * \frac{N[C(w_i)C(w_j)]}{N[C(w_i)]}$$

where $C(w_i)$ and $C(w_j)$ are the classes which respectively contain the words w_i and w_j. $N[C(w_i)C(w_j)]$ is the frequency of the sequence $C(w_i)C(w_j)$ in the training text and $N[C(w_i)]$ the frequency of the class $C(w_i)$ in this text.

The algorithm searchs which mapping leads to the smallest cross-entropy of the two distributions, defined as:

$$D(p||q) = \sum_{\{w_i w_j\}_{TT}} p(w_j/w_i) * \frac{log[p(w_j/w_i)]}{log[q(w_j/w_i)]}$$

The training text contains about 300 million words and is composed of articles of *Le Monde* and of *AFP* wires, written between 1987 and 1996. The vocabulary is composed of the 20,000 most frequent words of *Le Monde (1987-1988)*. The number of classes is 1,000 and has been determined as the interpolation parameters on held-out data. Deleted interpolation has been used to predict events, unseen in the training text.

6. Results

Models	LM A	LM B	LM C
reference words	10,000	10,000	10,000
PP_{Sh}		294	167
words in list	8,100	9,576	9,853
mean rank	238	305	246
words at rank 1	1,650	1,313	1,490
words at ranks 1 to 5	3,446	2,969	3,565
PP_{is}	437	283	162

Table 1: Comparative results in terms of observation ranks and perplexity for the 10,000 words to be found, including unknown words.

Models	LM A	LM B	LM C
reference words	9,382	9,657	9,451
PP_{Sh}		292	192
words in list	7,483	9,233	9,304
mean rank	258	314	260
words at rank 1	1,290	1,296	1,267
words at ranks 1 to 5	2,875	2,938	3,090
PP_{is}		300	186

Table 2: Comparative results in terms of observation ranks and perplexity for the reference words to be found, excluding the unknown words from the initial set of 10,000.

The results for the three models are given in tables 1 and 2. We have distinguished results with unknown words (table1) from the ones without unknown words (table 2). PP_{Sh} is the perplexity calculated with the protocol defined above, with the bets $\beta(w_r)$ put on the N reference words, w_r, and defined as:

$$PP_{Sh} = \frac{1}{\sqrt[N]{\prod_{r=1}^{N} \beta(w_r)}}$$

The tables include the number of words found in the 5,000 hypothesis lists, the mean rank of observation of the references in these lists, the number of words observed at the first rank and the number of words observed from the first rank to the fifth one. After the evaluation campaign, the complete reference sentences have been given to the participants. They have evaluated their models on this test (about 400,000 words) and the perplexities calculated *in situ* are written on line PP_{is}.

In order to verify a posteriori the choice for the number of references and hypotheses, we have plot the variations of the perplexity, PP_{Sh}, against the number of reference words on the figure 1 and the variations of the perplexity, PP_{Sh}, against the number of hypotheses on the figure 2.

When the number of references grows, the perplexity tends to a relatively stable value, showing the efficiency of the randomized selection of the tested words on a relatively

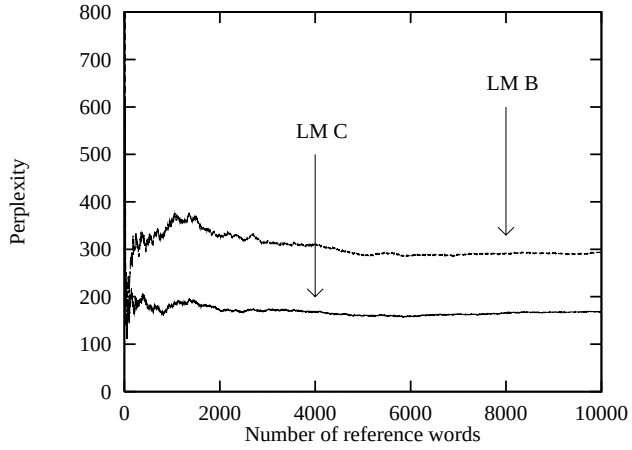

Figure 1: Variations of the perplexity, PP_{Sh}, against the number of reference words

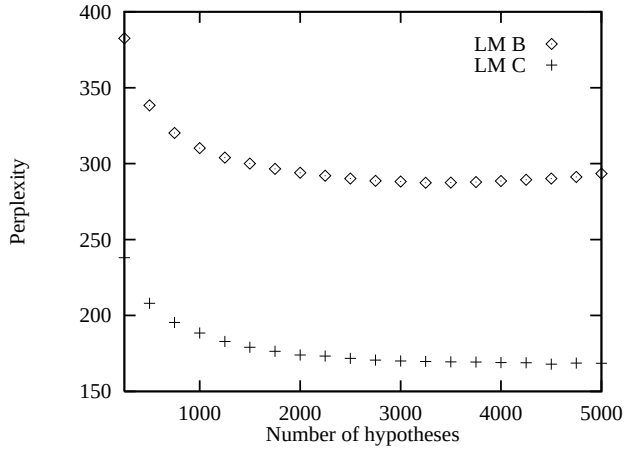

Figure 2: Variations of the perplexity, PP_{Sh}, against the number of hypotheses

small sample. When the number of hypotheses grows, we also observe that the perplexity tends to a constant value.

7. Evaluation's problems

This evaluation experiment has permitted us to figure out some uniformity problems in the training and test corpus. Every participant in this experiment has constructed his models using a corpus which has been treated long time before evaluation, leading to using his own criteria for rewriting compound words, apostrophes and numbers, etc. The text resources used in this evaluation campaign have been distributed by one of the participating laboratories with the description of formatting rules (Adda et al, 1997). They have been used without any modification in LM C. Nevertheless, some problems have been noticed due to the rewriting used in treating the test corpus. Some of which coming from the original text, the others coming from the preprocessing of this text. These problems have lead to the non-identification of some correct words especially for LM A and LM C.

7.1. Problems in the original text

Some words in the text were misspelled (fanastisme, EXPLICATIOB, ...), incorrectly segmented (MERESINDUSTRIES, PROJETSLIEUCAPITAL, URSScierie ...) or mistakenly written as in the sentence " pour le gouvernement de Mr William Clinton cela **sign ifiait** que les bénéfices ..." where the verb signifier has been truncated.

The non uniformity of putting accents for the same letter in the same word (specially for capital letters) gives misleading results when accented words are considered for prediction (Egypte instead of Égypte, Taiwan instead of Taïwan).

Long strings of foreign languages remain in the test, this arises every time one cites a foreign text.

7.2. Problems dues to the preprocessing

Words with capital letters

Rewritten rules for Mr, MR and M. are not the same in the training test and in the test.

Several words in the test have been uncorrectly written with their first letter as a capital.

One of us writes the words containing only one capital letter like Z., U., Q. while they are written Z, U, Q in the test.

Deletion of punctuation

The decision of removing the text punctuation in this campaign has also affected the evaluation. For instance, certain sentences have lost their syntactic validity after removing parenthesis and/or quotation marks. An example is given in the following sentence: " aux Etats-Unis, en mille neuf cent soixante-dix-huit puis mille neuf cent quatre-vingts, deux lois (deux) ont ouvert la période dite des " cieux ouverts " (open skies). " For which, removing the punctuation results in the incomprehensible sentence: " aux Etats-Unis en mille neuf cent soixante-dix-huit puis mille neuf cent quatre-vingts deux lois deux ont ouvert la période dite des cieux ouverts open skies."

Writing prefixes

Prefixes like (agro, pro, ...) were in many cases separated from the root by a hyphen and a space. In other words, these prefixes became lexical entities. However, except for some of them, the prefixes which end with o are attached to the following word if it begins with a consonant. This rule has not been respected in the test corpus as in: auto-référentielle, auto- flagellation, ...

Writing expressions and compound words

Some expressions were separated into two or more units, for example : au fur et à mesure... Some of us usually considered these expressions as a single word. This problem can be taken into account either a priori, with the definition of a common vocabulary or a posteriori, during an adjustment phase in order that each participant adapts its programs to the treatment of this kind of expressions.

In summary, for evaluating language models according to the same criteria, we have to consider these points and to evaluate their relative importance. Some of them can be

considered as noise, others can lead to misleading results. The work done by others in tokenization evaluation (Habert et al, 1998) would help us in this task.

8. Conclusions

In this paper we have described the efforts which are necessary to set up an evaluation campaign in a new domain. Tools for building the test data to be evaluated and tools for evaluating the models have been developed. The analysis of the results of this preliminary campaign with the detailed description of each system, shows the importance of the word segmentation and the necessity either to adapt the different models to a common vocabulary list or to take into account the discrepancies due to the different written forms, in particular to choose non misleading truncations in the sentences.

This work can be considered as a development phase in the evaluation of language models. It can be considered as the bases for a future complete comparative campaign.

9. References

Adda G., Adda-Decker M., Gauvain J.L., Lamel L. (1997). Text Normalization and Speech Recognition in French. EUROSPEECH'97, Rhodes, Greece, p.1711, 1997.

Bimbot F., El-Bèze M. and Jardino M. (1997). An alternative scheme for perplexity estimation. Proc. on ICASSP 1997, Munich, Germany.

Cerf-Danon H., El-Beze M. (1991). Three different Probabilistic Language Models: Comparison and Combination. Proc. on ICASSP 1991, Toronto.

Habert B., Adda G., Adda-Decker M., Boula de Mareuil P., Ferrari S., Ferret O., Illouz G. (1998). The need for tokenization evaluation. Proc. on LREC 1998, Grenade, Spain.

Jardino M. (1996). Multilingual stochastic n-gram class language models. Proc. on ICASSP 1996, Atlanta, USA.

Smaïli K., Zitouni I., Charpillet F. and Haton J.P. (1997). An Hybrid language model for a continuous dictation prototype. Proc. on EUROSPEECH 1997, Rhodes, Greece, p.2755.