INTERNET EVOLUTION AND PROGRESS IN FULL AUTOMATIC FRENCH LANGUAGE MODELLING

Dominique Vaufreydaz, Mathias Géry

Laboratoire CLIPS-IMAG, équipe GEOD et MRIM
B.P. 53, 38041 Grenoble cedex, France
{Dominique.Vaufreydaz, Mathias.Gery}@imag.fr

Abstract

The World Wide Web is the greatest information space unseen until now, distributed all over the world, in many languages, on many various topics. In a first part of this paper, we study the evolution of a French subset of this space during the last 3 years. During this time, the size of automatically extracted text for language modelling was multiplied by 6.5. Moreover, the French coverage has grown from 140,000 to 200,000 lexical forms. So, we show that we can get more and more reliable data in order to train our trigrams models. At last, recognition experiments, made on a French "state of the art" evaluation set, show that word accuracy increase from 51% up to 62.30% using two different models automatically calculated on Web corpora. The first corpus was gathered at the beginning of 1999 and the last one at the end of 2000.

1. INTRODUCTION

As we have already shown in a former paper [1], Internet can be a very interesting source for spoken language modelling, mostly n-grams, if we threat it in an appropriate way. To do that, we have previously described our "minimal block" method. But this method is efficient only if we can find needed words in input corpus. Indeed, our interest in speech recognition, and more precisely in language modelling, is language evolution on Internet and which words can potentially be found to train our language models.

Another point concerns the size of the training set for language modelling. Indeed, we need a very large amount of words to train 3-grams models. In the second paragraph, we will describe our Web crawler, Clips-Index and quickly explain his operating mode and his performances. In a third section, we will list our corpora, collected on the Web and details their size and French coverage.

In the next part, we will conduct some statistical experiment in order to check our first hypothesis. At the beginning of our work, we planed that Internet will grow and touch more and more different users. So, we predicted to get more and more reliable and various data to train our stochastic models. The last part, before the conclusion, will be about speech recognition experiment using two different corpora gathered on the Web.

2. CLIPS-INDEX

2.1. Overview

As we need to gather Internet documents for several researches, in information retrieval and language modelling for speech recognition, in different team of our laboratory, we decided to develop our own robot. So we built Clips-Index[1], a spider that crawls the Web, collecting and storing pages. Clips-Index tries to collect the largest amount of information in this heterogeneous context that is not fully respectful of the existing standards. It is a very interesting problem to collect the Web. We have made several improvements in order to correct usual errors, like bad written links, and to catch quickly more data. We do not want to work like usual search engine. So, we can not permanently get new or changed documents. Our aim is only to have a good snapshot of the Internet at once. So, Clips-Index is designed to get a large data collection in only a few days.

2.2. Technical description

Clips-Index is written in C++ for Windows™ platforms (Win9x, NT4 and Windows 2000). It is based on a multithreaded architecture that can have up to 500 simultaneous threads so up to 500 simultaneous connections to Web servers. Because of the overload a spider can cause to a server, timers are used and regulate requests. Moreover, Clips-Index respects documents privacy indicated by the robot exclusion protocol [2]. To limit the network bandwidth, Clips-Index uses a two HTTP requests method for collecting Internet documents. At first, it requests a header (HEAD command) to handle document types and do not download multimedia files - often very large - for example. Next, if type is correct (i.e. HTML or text), it downloads the document, checks its content by computing extra characters (i.e. control characters), because some misconfigured Web servers send multimedia documents with HTML type. Finally, it parses the file to extract new URLs and goes on to the next document.

[1] see http://Clips-Index.imag.fr/

2.3. Gathering performances

Clips-Index is quite efficient. For example, we have collected in October the 5th 2000, 38,994 pages on the ".imag.fr" domain. At the same time, Altavista indexes 24,859 pages and AllTheWeb 21,208 pages on the same domain. Tests made with *wget*, the GNU tool, give a worst score. So, Clips-Index seems to have a better parser than some other crawlers. Moreover, running on an ordinary low-cost 333 MHz PC with 128Mo RAM, Clips-Index is able to find, load, analyse and stock a maximum of 3 millions pages a day.

3. COLLECTIONS

3.1. Brief description

To reach our research aims, we gathered several Internet collections. These data were collected during the last 3 years. This paper only deals with our first and last Web corpora, i.e. the most representative of the evolution of Internet: *WebFr* and *WebFr4*. The first one is a corpus extracted on the Web during February 1999. The other one was gathered in December 2000.

3.2. Focus on French data

As we want to train French language model, we need to optimise the number of collected pages regarding the percentage of French data in it. We used AllTheWeb to get statistics about language in each Internet domain. The next table, show a subset of these results.

Domain	Nb Pages	% English	% French
be	1151946	31,62%	26,61%
ca	5228022	76,04%	20,29%
ch	2948797	25,60%	15,93%
com	113319060	83,21%	2,17%
de	17010456	18,03%	0,51%
edu	20744430	95,83%	0,23%
fr	3477169	19,98%	73,49%
gov	2308598	96,91%	0,11%
lb	41787	59,81%	29,83%
lu	112330	40,88%	32,55%
ma	39964	18,61%	76,22%
nc	21964	10,95%	85,69%
tn	17681	14,18%	66,98%

Table 1: Language used in documents of some Internet domains

The *Table 1* shows information about some Internet domains. We can find the total number of pages indexed by Altavista when we requested this information and the percentage of English and French pages in each domain. Now, these data are obsolete because of the Internet changes. All the domains are not listed here but we can find some interesting examples. In order to choose the list of domains to gather we must deal with several points. We defined several logical and technical rules that help use to define the Internet domains to parse.

Firstly, we must consider the amount of French data potentially in the domain. The first rule considers domains capabilities to provide adequate corpus in size for largest ones or in percentage for small ones. First if a domain has more than 1 million of French documents, it respects the first rule. Obviously, France (fr domain) respects this rule and contains more than 2,500,000 French documents. Commercial (com) is also a good provider with a few less than 2,460,000 French pages. Canada is too a good candidate with about 1,000,000 documents. After, for the other domains, we compared percentage of French. We set the threshold to 20%. Thus, all English domains, like uk, edu and gov, are not incorporate in our list. Switzerland (ch), a French/Italian/German domain, is also rejected.

Secondly, we must integrate technical facts in our decision. Indeed, whereas Internet indexing engine, we have large hard drives but we can not save all ".com" pages for example. So, even if we can get a lot of interesting data from this domain, we decided to ignore it. Another problem is the network proximity between Web servers and our crawler. For example, Canada is far from France and when we tried to collect pages, we had a lot of timeouts. When we changed timeout values, we obtained more pages but we also penalised gathering performance. Threads spent more time in waiting responses from servers than really working to get more data. So, even if we have a huge parallelism, as seen in 2.2, we decided to limit our gathering space to "nearby" servers.

Lastly, we decided to collect small French speaking domains, which can be useful for our linguistic researchers to compare standard French and idioms of these countries. We encounter a lot of problems with these Web servers. The last rule is "is the gathering overload, timeouts and network distance, negligible regarding the research interest of the collected data ?". Thus, we added domains like Morocco, New Caledonia, Tunisia, Gabon… The final list of Internet domains we used is: ad, af, ag, be, bf, bi, bj, bo, bt, cd, cf, cg, ci, ck, cm, cu, dj, dz, eg, fj, fm, fr, ga, gd, gf, gn, gp, gq, km, int, jm, jo, kh, km, ky, lb, lc, li, ls, lu, ma, mc, mg, ml, mq, mr, mu, nc, ne, ng, pf, qa, re, rw, sc, sn, st, td, tf, tg, tn, tv, va, vn, vu, wf, yt.

4. FIRST STATISTICAL STUDIES

In the next studies, we will present evolution of available data for French language modelling. The first one, we call Grace, is a 20 Megabytes corpus extracted from the French newspaper *"Le Monde"* [3]. It has been used for an evaluation of parsers for French texts. It will only be used as a reference corpus for the French coverage because many research teams in French automatic speech

recognition use data from *"Le Monde"* as training data for their language models.

At first, we will analyse the respective size of each corpus. After, we will examine the French coverage evolution of Web data compared to Grace.

4.1. Growing factor of the training set

WebFr is almost the quarter of *WebFr4*: 10 Gb for the first one and 44 Gb for the other. So we can estimate the growing factor of extracted text between *WebFr* and *WebFr4*. This text is somehow different from the original one, written by the author of the Web page. Indeed, we must, despite it is done in Information Retrieval, transcribe all numbers in their expanded text version to limit the number of entries in our speech recognition engine. So, our extracted text is bigger than the text written in pages. As we need to take care of sentences beginnings and endings, we extract some extra information from the structure of documents. For example, each sentence in a table, even if there is no written diacritics, are considered to be ended at the closure of the cells ("<TD>" and "</TD>" tags). Other tags are used but are not detailed here. Another important point concerns difference between methods of text extraction used with *WebFr* and *WebFr4*. At first, we try to correct unaccented words to their correct form, the nearest in term of numbers of substitutions needed. That causes a big computation overload because it is not simply looking for existing words in a dictionary but calculating distances with a sub-set of it. For *WebFR4*, after some experiments described later in this paper, we decided not to do that.

After these explanations, we can estimate a growing factor (*gf*). In *WebFr*, each document is, in average, 6.6 Kb long and contains 4.2 Kb of text after removing HTML tags and rewriting numbers. In *WebFr4*, each document is 8.5 Kb long in which we can get 3.6 Kb of data. The proportional reduction of extracted text can be explained by the development of advanced interfaces, using javascript for example, on the Web. Indeed, writing menu and animated content in such language is very long regarding the text really on the screen and in the document. Finally, the growing factor can be calculated:

$$gf = \frac{\left(\frac{3.6}{8.5}\right)}{\left(\frac{4.2}{6.6}\right)} * \frac{44}{10} = 2.92 \approx 3 \qquad (1)$$

This factor, rounded to 3, gives us first information about growing of interesting data for our task. It needs to be completed by French coverage information to be useful for analysis.

4.2. Study of the French coverage

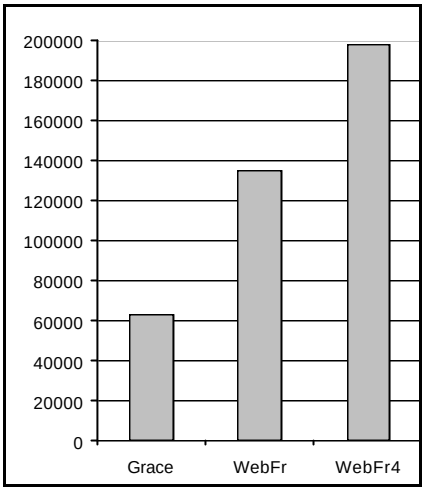

Figure 1: French coverage of the three corpora

The Figure 1 shows us the number of lexical forms we can find in the three corpora. These forms are obtained by computing word frequency on each corpus. The maximal list of French lexical words is constructed with two lexicons, BDLex [4] and ABU dictionaries [5], and consists of more than 400,000 forms. We can see that potentially, *WebFr* contains contextual information, used in n-gram models, on more than the twice number of words than Grace. *WebFR4* is more various than the others with a few less than 200,000 lexical forms.

In conclusion of these first studies, we can notice that not only we have more data, but also at the same time, these data become more various. Compared to Grace, the two Web corpora are very huge and provide more diversity in French forms. All these observations show an undeniable evolution of Web content from February 1999 to December 2000 and validate our hypothesis. We can now go ahead and test these data in speech recognition task in order to verify if, as we hope, *WebFR4* is a better reflect than *WebFr* of French language.

4.3. Extracted corpus

Now, we are going to study the impact of the evolution of Internet on our very specific use, the French language modelling. The table below indicates how many words (including special marks for beginnings and endings of sentences), with a 20,000 words lexicon and our "minimal block" method, we can get from each corpus to train our n-grams models.

Corpus	Number of words in training set
WebFr	245,525,254
WebFr4	1,587,142,200

Table 2: Number of words in training set

In this table, we see immediately a big difference between these two results. If we calculate a real growing factor of our training set, we obtain a value near 6.5. It is more than the double of our estimated factor seen in (1). This result corroborates our first analysis concerning the variety of *WebFR4*. If we consider the WWW in the past, we see that at the beginning of its development, the aim of the users was only to have a page on the Internet. But in the last months, we notice an evolution of the users' comportment. Most of the pages contain now advanced features, as we said before, but also more clean textual content. We think that all the new affordable and cheap tools for Web pages design, that often include word spelling verification, and the interest of most users for the new technologies can explain the better quality of pages on the Internet.

5. SPEECH RECOGNITION

5.1. Description of the evaluation material

Now we will test different language models calculated on Web data in speech recognition task. We will conduct tests on 300 signals from the Aupelf evaluation of French dictation system [6]. The vocabulary (used in 4.3) was given by the Aupelf. Signals contain OOV words. Raphaël [7], our French recognition engine, is used in tests. Its acoustic models were trained on only 12 hours of speech. The computation of language models by the "minimal block" method is full automatic. The process takes in input a corpus, a vocabulary and the cut-offs for bigrams and trigrams. In output, it generates an ARPA language model. We calculate two trigrams languages models with *WebFr* and *WebFr4*. We obtain the word accuracy of the system by inverting the Word Error Rate. We match recommendations[2] and use tools (sclite) from the NIST to compute results.

5.2. Recognition results

	WebFr	WebFr4
word accuracy	51,00%	62,70%

Table 3: recognition results using WebFr and WebFr4 to train our language models

We note in *Table 1* that the Word Error Rate decreases from 49% to 37.30%. So, these results corroborate our previous information concerning these corpora. *WebFr4* is not only larger than *WebFr* but it has also a better quality of content for language models training. We can compare Raphaël to the systems that made the Aupelf evaluation. Our word accuracy is almost the same as the other systems, except for the best system of the evaluation. We are working on better acoustic models train on more data, the complete BREF corpus like the other evaluation systems, in order to get similar experimental conditions and provide more reliable comparisons. As we conducted experiments in real-time conditions, we may improve recognition result by increasing search space. Moreover, the language models were not tuned at all to improve performance and some adaptation may lead in word error rate reduction.

[2] Compound words, like "aujourd'hui" in French; are considered as two distinct words.

6. CONCLUSION

In this paper, we have shown that Internet give us new interesting data for language model training. That confirms our first experiment described in [1]. Moreover, with less computation, because we do not anymore need to correct errors, we can obtain more reliable corpus than before. We are now investigating other properties of the Internet data in language modelling. For example, we are working on topic detection using such data in order to increase performance of very large vocabulary system. Besides, in [8], experiments on automatic aligned multilingual texts have already been done in information retrieval research. So, we work too on multilingual language modelling for international speech recognition engine and for speech to speech translation system.

7. REFERENCES

[1] Vaufreydaz D., Akbar M., Rouillard J., *Internet Documents: A Rich Source for Spoken Language Modelling*, ASRU'99 Keystone, Colorado (USA), p.277-280.

[2] Koster M, *A Method for Web Robots Control*, technical report of IETF, December 1996.

[3] see the LIMSI Web site about the GRACE action, a French evaluation of text parsers http://www.limsi.fr/TLP/grace/index.html

[4] Pérennou G., De Calmès M., *BDLEX lexical data and knowledge base of spoken and written French*, European conference on Speech Technology, pp. 393-396, Edinburgh (Scotland), September 1987.

[5] see http://abu.cnam.fr/

[6] Dolmazon J.M., Bimbot F, Adda G, El-Bèze M, Caërou JC, Zeiliger J, Adda-Decker M *Organisation de la première campagne Aupelf pour l'évaluation des systèmes de dictée vocale*, 1st jst Aupelf-Uref, Avignon (France), April 1997.

[7] Akbar M., Caelen J., Parole et traduction automatique : le module de reconnaissance RAPHAEL, COLLING-ACL'98, pp. 36-40, Montreal (Quebec), August 1998.

[8] Nie J.Y., Simard M., Isabelle P. Durand R., *Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web*, 22nd Annual International ACM SIGIR, pp. 74-81, Berkeley, CA, USA, August 1999.