Internet Documents: A Rich Source for Spoken Language Modeling

Dominique Vaufreydaz, Mohamad Akbar, José Rouillard

Internet Documents: A Rich Source for Spoken Language Modeling

D. Vaufreydaz[*]*, M. Akbar*[*]*, J. Rouillard*

Laboratoire CLIPS-IMAG, équipe GEOD
Université Joseph Fourier, Campus scientifique
B.P. 53, 38041 Grenoble cedex, France

ABSTRACT

Spoken language speech recognition systems need better understanding of natural spoken language phenomenon than their dictation counterparts. Current language models are mostly based on written text and/or very tedious Wizard of Oz or real dialog experiments[1]. In this paper we propose to use Internet documents as a very rich source of information for spoken language modeling. Through detailed experiments we show how using Internet we could automatically prepare language models adapted to a given task. For a given recognition system using this approach the word accuracy is up to 15% better than a system using language models trained on written text.

1. INTRODUCTION

One of the main components of an automatic speech recognition system (ASR) is the language model (LM). The role of LM is to reduce the search space in order to accelerate the recognition process. In phoneme based ASR, LM is also the bridge between phonemic and textual representation of speech utterances. LMs are basically of three types: grammar based, statistics and a mix of them. To produce a grammar based LM one only needs to describe the language to recognize using well known grammar definition languages. However the use of fixed grammatical constructs reduces the flexibility of LM to respond to the real situations like spontaneous speech in which the written language constructs are not well respected. Statistical LMs cope with the situation by exploiting statistical properties of the language in a given context of two, three or more words. Using statistical LMs we achieve more flexibility than grammars in describing the language. They are also easier to integrate to HMM or NN driven recognition systems and provide more robust answers in the same contexts. Integration of statistical LMs with grammars, also stochastic grammars in which the probabilities are directly incorporated to the parser have already been studied. Though due to their simplicity and tractability statistical LMs are the natural choice of many current ASR systems.

Statistical LMs are calculated generally on the large text corpora and by controlling vocabulary size, LM context length and back off schemes used to deal with out of vocabulary (OOV) words. The context length is chosen based on the vocabulary size and the corpus size. The suitability of a given LM for a given task is generally described by perplexity value of LM calculated on a test corpus. In a few words, perplexity (branching factor) is a measure for average number of words that can potentially follow a given word on the corpus [4]. The techniques to calculate and adapt a LM for a given task are well known and we are not going to discuss them here.

As we mentioned earlier the size of learning text corpora is of major importance during the learning phase. In fact increasing its size generally increases the number of different contexts seen during the learning phase and hopefully yields more robust LMs. However the selection method during corpus construction is very important. It is very well known that written text, for example, cannot be directly used in calculating LMs appropriate for spontaneous speech recognition.

Nowadays, many people can access the Internet, either from work, school or home. Other than consulting existing documents on the Web pages, news servers and chat sessions, they participate actively on the Internet by creating, publishing and/or synthesizing contents. Depending on the writing context (professional, personal, educational…) these document are of very different nature. People generally use a simplified vocabulary and a set of ungrammatical expressions as they do in everyday life. This means that using the documents publicly available on the Internet we can obtain a very huge corpus that is a mixture of well-written text and also near spontaneous speech. Using this corpus to train LMs will be more appropriate in the context of dialog systems and/or spontaneous speech recognition.

In section 2.1, we describe how using our indexing engines and appropriate filters we could collect a very huge set of French documents that could directly be used in LM learning. Another question of importance is: “to what extent should we increase the size of the corpus and what are the practical limits on its size if there is any?”. In section 3 we describe a number of experiments we conducted on this matter by calculating “context count” and perplexity evolution by corpus size.

2. DATA ANALYSIS

2.1 Gathering Internet documents

As we already mentioned, one could use any source of Internet documents to collect the corpus. To simplify our task we started our experiments using only the most widely used Internet documents Web pages and news posts.

---

* Member of CSTAR-II project French team.

[1] Recording and their transcription are very expensive and resource consuming. They must be repeated for every given task.

2.1.1 Web documents

In collaboration with MRIM[2], another team of our laboratory, we have developed a Web robot, which is in charge of collecting Web pages from the Internet. This bot is called **CLIPS-Index**. This tool is a multithreaded application for windows 9x/NT. It takes one or several starting points on the Web and finds all pages and text documents it can reach from there. It filters out documents on their types, separating html and text documents from others (images, audio…). CLIPS-Index is an RFC-2068 compliant robot that respects privacy of documents explicitly described in "robot.txt" files of each visited server.

CLIPS-Index provides a good way to collect Web data quickly. During February 1999, we performed our first collection of Web data (HTML and text): **WebFr**. This corpus contains more than 1,550,000 different documents (URL checked with a no case sensitive comparison) amassed in 84 hours. They were found on more than 20000 servers from the French domain '.fr'. **WebFr** is a collection of about 10 gigabytes of HTML and text documents. This collection is the first part of our work.

2.1.2 News posts

The distributed nature of News servers in which they mirror the content of the others allows us to gather interesting posts by sinking only one server. **CLIPS-News** is another robot developed to this means. Most of news servers do not keep messages more than a few days. So to collect a substantial set of news documents our robot revisits the news server every 6 hours and sinks the new articles. A filter option in **CLIPS-News** allows us to choose the newsgroups hierarchy to download.

Our collection of news posts was gathered from 1st to 20th June 1999. In our studies, we need only French text. Therefore we restrain the download on the 234 newsgroups (at this time) of the 'fr.' hierarchy. The result is called **NewsFr**. It consists of about 400,000 posts and represents a 660 megabytes corpus. This is the second part of our work.

2.2 Text corpus generation

Extracted data from Web and News posts are not in a suitable textual form to be used directly in language modeling, headers, tags and other diacritics are superfluous and must be removed. In the other hand all the documents in NewsFr and WebFr are not in French language. Indeed, documents in any language can be freely posted to News or published on .fr domain. However the selection of .fr domain helped us to reduce the chance of gathering texts written in other languages to a minimum. So the first step, to filter out the documents, was to choose only French text from all the extracted documents. BDLex [1], a dictionary with 245,000 entries, was chosen as lexicon to get large vocabulary coverage. It was enlarged with another source from ABU [3] (Universal Bibliophiles' Association) to about 400,000 distinct lexical forms.

Extraction begins by a first filter, which takes a document (Web page, News post or text) and outputs a formatted text. A few tags have been used to make a better correspondence between input documents and output text. For example start and end points of sentences are not always clearly marked, using tags like headers, paragraphs, table rows and columns, etc. has been proved useful to mark these points.

The second filter increases text output quality by making several changes to the input stream. For example 'Ecole' becomes 'école'. In this case, the filter restores the accentuated form of the word, which had lost the first accent due to capitalization. Conceptually this filter tries to find closed matches with an accentuated one. At this stage, the text will also be converted to a lowercase. The next step is to output, based on a task specific vocabulary, what we called "minimal blocks". A minimal block of order *n* is a sequence of at least *n* consecutive words from the document with all words of the block in the given vocabulary. Table 1 gives a concrete example of how this filter works. This allows us to control the OOV phenomenon and also choose the desired subsets of resulting corpus based on a specific vocabulary, which is in general quite smaller. We also added an option to this filter so that only complete sentences get out of the filter while this option is active.

Minimal Block Length	Output sequences
2	<s>[3] bonjour monsieur comment allez vous </s>
3	comment allez vous </s>
4 and more	Ø
Complete sentence	Ø

Table 1: Output sequences for the sentence "bonjour monsieur Durand, comment allez-vous ?" given the word Durand is not in the vocabulary

To study the impact of the minimal block length and also complete sentence option, we conducted several experiments using WebFr and a 3000 word vocabulary. The results are represented in Table 2.

Minimal Block Size	Resulting corpus size in words
3	145 Millions
5	44 Millions
Complete sentences of minimum 5 words	46,500

Table 2: Impact of minimal block size on resulting corpus with a 3000 word vocabulary

We can see in this table that refining options alters drastically the size of output corpus. In this example, increasing constraints reduces number of words by a factor of 3000. Filtering parameters must be though carefully chosen.

Some other filtering options are also added to increase the output quality. For example an option controls the acceptable number of consecutive digits and numbers in output. This option reduces the number of such patterns that happens regularly in "Web statistics" which are visible on the Web sites.

[2] see MRIM Web page at http://www-clips.imag.fr/mrim/

[3] <s> and </s> are diacritics added to indicate respectively start and end of sentences in the output corpus. Note that in the output not any </s> is followed by a <s>.

In further experiments we chose the minimal block length of 5 that gives a reasonably sized corpus and also a context of sufficient length which helps to better interpolate linguistic properties.

3. EXPERIMENTS

In order to evaluate the adequacy of the resulting corpus, we conducted a series of experiments in which we tried to measure how Web and News could contribute to model a dialog based language; and also to which level the corpus is pertinent in calculating the LM for spontaneous speech recognition.

3.1 Contributions of Web and News

The terms and expressions used in natural dialogs differ from what is used in written texts. So estimating LMs adapted to dialog conditions may not be done using text samples extracted from journals, books or similar documents. To measure the adequacy of a corpus for dialog conditions we noticed that the frequency of pronouns used in natural dialogs does not correspond to their frequency in written documents. We calculated the frequency of French personal pronouns in three corpora, WebFr, NewsFr and a 20 Mbytes extract of "Le monde" journal used as a reference in GRACE [1] project. Figure 1 illustrates the results in which we notice that Grace corpus does not practically contain the French pronouns mostly used in dialogs (je, tu, vous). In the contrary, NewsFr and WebFr contain these pronouns more frequently. That shows a more important presence of dialog contexts. At the same time the other pronouns are mostly present with the same degree of importance in all corpora. This means that other contexts are also present in WebFr and NewsFr as any other type of written text documents and confirms that WebFr or NewsFr can be used as a base for calculating LMs.

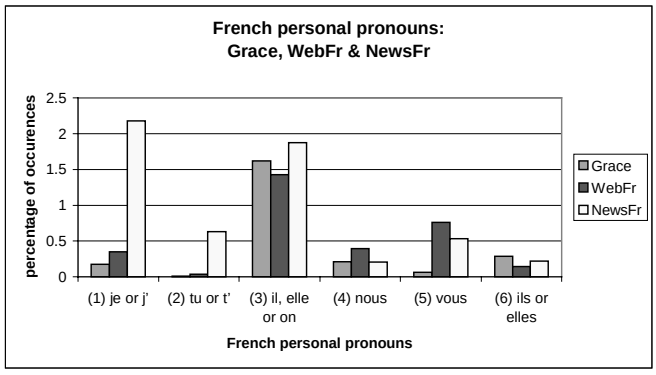

Figure 1: French personal pronouns occurrences in 3 different sources: Grace, WebFr, NewsFr

The preceding also ascertains our hypothesis that people regularly use everyday vocabulary and expressions over the Internet.

3.2 Language coverage

The question of to what extent a given corpus covers a specific language with a given vocabulary is a crucial one during LM design. Today, LMs are generally of 3-gram type. To answer the question we studied the evolution of the language coverage versus corpus size with a 3-gram context in mind. Our measures were 3-gram count and LM perplexity.

In the first experiment we gradually increased the corpus size and studied the 3-gram count evolution. The original corpus contains 45 million words obtained on a 3000 word vocabulary specific to reservation and tourist information domain.

To be sure of statistical validity of our experiment, we conducted a Leave-One-Out experiment (also known for H-test) in which we divided our original corpus to 20 smaller and identical sized parts. In each experiment we calculated 3-gram count evolution on only 19 of these parts. The mean average and the standard deviation evolution for 3-gram count of these sets of experiments are shown in Figure 2. The low standard-deviation in these sets of experiment shows how significant is the average curve. In fact we observe a slight sign of saturation at the end of our experiments but the current size of corpus does not allow a formal conclusion about this phenomenon.

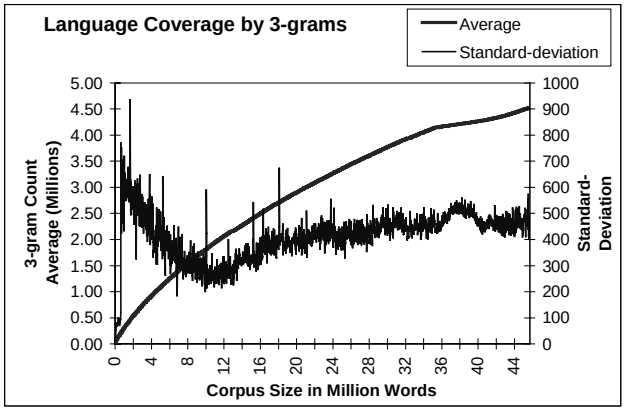

Figure 2: Corpus coverage evolution by 3-grams

The evolution of 3-gram count by increasing the corpus size can be a measure of how large is the LM coverage. At the same time 3-gram count increase may not lead to a substantial perplexity increase. To study the perplexity evolution over corpus size we conducted another experiment in which we gradually increased corpus size (steps of 2 million words) and we measured the perplexity of resulting LM. The corpus and the vocabulary were the same as the preceding experiment. A Turing discounting scheme was used to calculate LM back-offs. Figure 3 illustrates the perplexity evolution versus corpus size. The low standard-deviation in the last experiment confirmed that there is no need to conduct an H-test to statistically confirm these results. By studying the resulting curves it seems that by any practical measure the perplexity of the resulting LM are about 80 and there is no more chance to achieve better perplexities for the given task[4]. However the 3-gram count is steadily increasing. This is also a very good confirmation that perplexity is not an efficient measure of LM quality.

[4] Though a reasonable perplexity value for a 3000 word vocabulary.

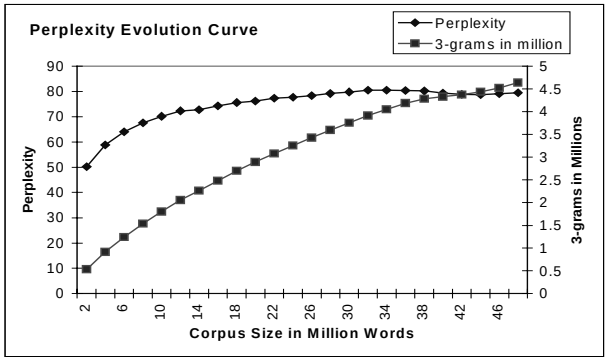

Figure 3: Perplexity measures versus corpus size

Another interesting measure, deduced from this experiment, is the n-gram hit factor evolution. By feeding the same subset of corpus used to calculate the LM as the test set during perplexity computation, we measured also the 1, 2 and 3-gram hit factors[5]. The Figure 4 illustrates the results. It clearly shows that however we are increasing the 3-gram count by increasing the corpus size (see Figure 2) this does not lead to a significant increase in 3-gram hit factor. We can deduce that our original corpus is a good candidate for calculating the LM for the given task.

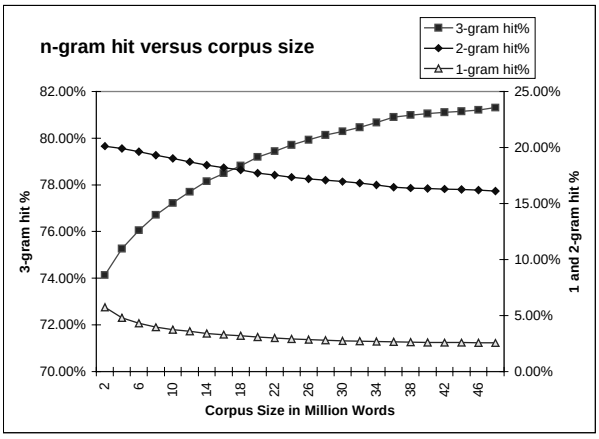

Figure 4: n-gram hit in percent versus corpus size

3.3 Speech recognition

To examine the usefulness of our approach in language modeling for spoken language speech recognition, we conducted a series of recognition experiments with a recorded dialog focused on reservation and tourist information task. Nothing but the corpus size has been changed in each experiment to study their direct influence on recognition performance. Grace and Wizard of Oz corpora recognition performances on the same task have been taken as base lines[6]. The results are shown in Figure 5. We can see that even for a LM with comparable size obtained from WebFr (the first point) the word accuracy is about 15 % better than the base lines. This confirms again our hypothesis about usefulness of the method.

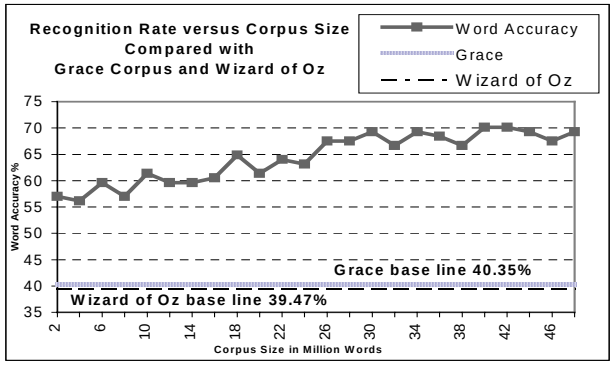

Figure 5: Recognition performance versus corpus size compared to base lines of Grace corpus and Wizard of Oz

4. CONCLUSIONS AND PERSPECTIVES

We have shown in this paper that Internet documents can be a very rich source of information for language modeling. However to be useful, one has to clean and filter out the extracted documents based on application in which LMs will be used. We successfully applied this method by integrating a trained LM to our spontaneous speech recognition module (RAPHAEL [6]). We tested a modified version of this system in real situation during CSTAR [7] demonstrations. Furthermore, we work on a task-oriented recognition engine adapted for information search and retrieval system, HALPIN [8]. Integrating vocal technology offers ease of use and provides more natural dialogs with HALPIN.

In the future, we will further investigate the use of this technique to automatically extract task and language based vocabularies that can be directly used in spoken language speech recognition systems.

5. REFERENCES

[1] 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.

[2] GRACE: see http://www.limsi.fr/TLP/grace/index.html.

[3] ABU: see http://cedric.cnam.fr/ABU/index.html.

[4] Rosenfeld R., "A maximum entropy approach to adaptive statistical language modeling", *Computer, Speech and Language* (1996).

[5] Clarkson P., Rosenfeld R., "Statistical Language Modeling using the CMU-Cambridge toolkit", *Eurospeech 97.*

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

[7] CSTAR: see http://www.c-star.org/

[8] Rouillard J., Caelen J., "HALPIN: a multimodal and conversational system for information seeking on the World Wide Web", *ESCA ETRW workshop: Accessing information in spoken audio*, pp 20-24, Cambridge (UK), April 1999.

[5] Note that using Turing discounting even with a closed vocabulary may lead to hit factors less than 100% for 3-grams.

[6] The size of language model obtained from Grace is comparable to that obtained from a 2 million word WebFr corpus. Wizard of Oz experiment has been conducted for CSTAR-II project.