Using the Web for fast language model construction in minority languages

Viet Bac LE, Brigitte BIGI*, Laurent BESACIER*, Eric CASTELLI***

* CLIPS-IMAG Laboratory, UMR CNRS 5524
BP 53, 38041 Grenoble Cedex 9, France
** MICA Center, 1 Dai Co Viet, Hanoi, Vietnam
e-mail: (viet-bac.le, brigitte.bigi, laurent.besacier, eric.castelli)@imag.fr

Abstract

The design and construction of a language model for minority languages is a hard task. By minority language, we mean a language with small available resources, especially for the statistical learning problem. In this paper, a new methodology for fast language model construction in minority languages is proposed. It is based on the use of Web resources to collect and make efficient textual corpora. By using some filtering techniques, this methodology allows a quick and efficient construction of a language model with a small cost in term of computational and human resources. Our primary experiments have shown excellent performance of the Web language models vs newspaper language models using the proposed filtering methods on a majority language (French). Following the same way for a minority language (Vietnamese), a valuable language model was constructed in 3 month with only 15% new development to modify some filtering tools.

1. Introduction

There are more than 6000 languages in the world but only a small number possess the resources required for implementation of Human Language Technologies (HLT). Thus, HLT are mostly concerned by languages which have large resources available or which suddenly became of interest because of the economic or political scene. On the contrary, languages from developing countries or minorities were less treated in the past years. One way of ameliorating this "linguistic divide" is through starting research on portability of HLT for multilingual applications. This question has been increasingly discussed in the recent years. The SALTMIL[1] (Speech and Language Technology for Minority Languages), which is a Special Interest Group of ISCA, was created to promote research and development in the field of speech and language technology for lesser-used languages, particularly those of Europe. However, in SALTMIL, "minority language" mostly means "language spoken by a minority of people". We rather focus, in our work, on languages which have a "*minority of resources usable in HLT*". These languages are mostly those from developing countries, but can be spoken by a large population. In this paper, we will notably deal with Vietnamese, which is spoken by about 70 millions of persons, but for which very few usable electronic resources are available.

Among HLT, we are interested, in this paper, in Automatic Speech Recognition (ASR). We are currently investigating new techniques and tools for a fast portability of speech recognition systems to new languages. This topic has already been tackled in [1] and [2] but mostly for languages which already have large corpora available. Conversily, we particularly address languages, like Vietnamese, for which few signal and text resources are available. This activity includes different aspects:

- Portability of acoustic models: this can be achieved, for examples by using tools for performing a fast collection of speech signals [3] or by using Language Adaptive Acoustic Modeling [4].
- Language modeling for new languages: we propose to use web-based techniques [5] which have already shown ability to collect large amount of text corpora. For languages in which no usable text corpora exists, this is moreover the only viable approach to collect text data.
- Dictionnaries: collaborative approaches like in [6] could be also proposed for ASR.

This paper adresses particularly fast language model construction for ASR. The proposed method uses the web to collect large amount of data. In section 2, we first describe our text data collection tools and the filtering techniques associated which were first developped for French language modeling. Then, in section 3, we describe the modifications implied to adapt our collecting and filtering tools to Vietnamese. Section 4 is dedicated to experiments performed to validate our methodology; for comparison purpose, perplexity figures are simultaneously given for French and Vietnamese. Finally, section 5 concludes this work and gives some perspectives.

2. Language Modeling using Web resources

Language Modeling is one of the most important modules in a large vocabulary speech recognition system. Statistical language models (SLM), which describe probabilistically the constraints on word order found in language, are traditionnally used. However, it is difficult to construct a SLM because a large enough corpus which models all possible user input must be available. A large corpus tends to have more contexts for each word, and thus tends to produce more accurate and robust SLMs. N-grams based model is a useful one for solving this problem. In this model, an estimate of the likelihood of a word is made solely on the identity of the N-1 preceding words in the utterance. For more details, see [7].

With the development of the Internet and its services, the Web is the greatest information space distributed over the world, in many languages and on many topics. Web resources can be a very interesting source for spoken language modeling if it is

[1] http://www.cstr.ed.ac.uk/~briony/SALTMIL/

processed in an appropriate way. There are many solutions for a SLM construction using the Web. In particular, in the domain of information retrieval, we found some "web search query" based approaches [8, 9]. In our case, these solutions can not be applied at the moment because we have no tool for automatically generating the queries.

This section describes some techniques for language model construction. First, by using a web-robot (or web-spider), web pages can be collected and stored in the given language. And then, a text corpus is builded by filtering and analysing the web pages. Finally, all N-grams models are estimated from this text corpus.

2.1. Web pages collecting

Documents were gathered from Internet by some web robots (among them, one was developped in our lab[2]). From some starting points on the Web, the robots can reach and find all the text documents and web pages which have a direct or indirect link with these starting points. However the Web sites (Internet domain names), accessed by the robots, must be managed because only the documents in a given domain and in a given language must be collected.

2.2. Data preparation

Some filtering techniques are needed to construct the text corpus from HTML pages. The text parts from the HTML pages must be extracted and some document separators were inserted. The tokens <s> and </s> signal respectively the begin and end of a sentence. Web texts contain also a variety of "non-standard" token types such as digit sequences, words, acronyms and letter sequences in all capitals, mixed case words, abbreviations, roman numerals, URL's and e-mail addresses... These non-standard types cause problems for training language models. Normalizing or rewriting such text using ordinary words is a first important issue.

Then, for language modeling application, by using a compound word lexicon, we also compute compound words that are treated in the language model as one word. There are two benefits in this method: there is no biased usage of the word penalty of the recogniser and it increases the context taken into account in the language model [5]. The common words are then regrouped into classes for introducing them in the SLM. The choice of the classes depends on the task-specific application. For example: country name, city name, days of the week, month... Finally, numbers in context (date, money, etc) were also transcribed to their textual form (number-to-text).

2.3. Sentence filtering

There are many different solutions to extract the relevant sentences from a text corpus. Classically, the sentences exclusively made with words of the task-specific vocabulary preliminary defined are kept. The other method proposed in [10] is a text filtering algorithm based on character perplexity. However, it needs a "Standard Language Model" for reference and there is not any such reference model available in minority languages. The "minimal blocks" filtering method proposed in [3] can also be used. 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.

[2]http://slmg-index.imag.fr

Table 1: *List of the fixed and variable modules to adapt tools from French to Vietnamese*

Fixed modules	Variable modules
data collecting	character converting
html2text	case changing
token normalizing	number2text
sentence splitting	lexicon constructing
word splitting	
common word grouping	
data filtering	

We note that there is not a lot of research work which compares the performances of these methods. So, in the section 4, we propose, combine and compare several filtering methods applied in our experiments.

3. Portability to a new language

3.1. Methodology

As we show in the previous section, using the web to construct the SLMs implies to develop tools for text extraction and text selection. This development can be carried out specifically for each language but this work is laborious and time consuming. In the context of genericity, producing reusable components for language-and-task-specific development is an important goal. The aim is to create a set of tools that would represent the common text normalization for many languages. Consequently, we decided to construct a lot of small tools for French (the "source" language) and to estimate time consuming to adapt tools from French to Vietnamese (the minority target language).

First, because we want to construct SLMs in multiple languages, a unique character set for encoding all the documents and for covering all languages possible must be chosen. Universal Character Set (UCS) which is a part of Unicode international standard[3] provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. We have chosen Unicode standard for encoding all characters of our corpora. But there are hundreds of different character sets for encoding a character of the Web documents. We note that the French SLM tools we inherited are single byte (ASCII) based. Indeed, we construct a tool to convert a character in several character sets to the Unicode (UTF-8 encoding system).

Secondly, we decide to split the original tool for source language (French) to a set of modules (see table 1). And then, we have determined what are:

- *Fixed modules:* the modules which do not depend on the language.
- *Variable modules:* the modules which depend on each language.

This splitting and determination work is really important. For a new language modeling, we will inherit all the *fixed modules* and fastly adapt the *variable modules* to that language. It will economize the time consuming to build a complete language model. We propose these tools available on demand for any person who is interested in.

[3]http://unicode.org/

3.2. Application to Vietnamese language modeling

3.2.1. Vocabulary

To build a SLM and filter out the documents, it is necessary to have a vocabulary containing words. This vocabulary can come from a variety of resources in the Internet. We can use a bilingual or multilingual lexicon for generating this vocabulary.

In fact, there are many methods to construct a vocabulary. In the context of the Papillon[4] project, the construction of a lexical base for a new language may take several different ways depending on where the author has to start: collaborative approaches [6], dictionary recycling [11]... This project aims at creating a multilingual lexical database covering among others English, French, Japanese, Malay, Lao, Thai and Vietnamese.

¿From this Papillon project, we got a vocabulary for Vietnamese language (from French-Vietnamese and Vietnamese-French dictionaries). Then, we filtered this vocabulary to have a list of more than 40,000 unique words in Vietnamese: compound words, borrowed words and isolated words. By taking only the most frequent words, we can discount this size of vocabulary to 20,000 words. These were the highest frequency words which occur in the documents of our training corpus.

3.2.2. Data collecting

Text corpus for language modeling cannot be collected easily in the minority languages for some reasons:

- There are less pages and websites than in the majority languages.
- The debit of communication is often very low (several kilobits per second).

Consequently, we can not crawl all of the websites but we must focus on some which have more pages and higher debit than the others. So, a non negligible time was used to find out the websites to collect.

At the time of our work, there were about 2500 Vietnamese websites in Vietnam which publish: daily news, information, entertainment, e-commerce, forum... The daily news web pages introduced a constraint in the data collection, since we had to regularly access the same sites to get an acceptable amount of data. This is the major difference with web data collection for a majority language like French or English where there are enough web pages that can be collected at a given time.

3.2.3. Text-corpus filtering techniques

Our first positive result was the porting of our SLM tools to a new language. Indeed, we must have only a short time to modify and to adapt these variable modules to a new language. We have built a language model for Vietnamese in only three months with this methodology. A comparison of this minority SLM (for Vietnamese) with a majority one (for French) is proposed in the next section.

4. Experiments

4.1. Training corpora

The French data collection (called WebFR4) is a very large corpus containing a few less than 6 millions web pages representing 44 GB. This corpus was gathered in December 2000 and the collect was restricted to the *.fr* domain. The set of exploitable resources (after data preparation) is made of 12 GB, i.e. 184,738,292 sentences. This set of text is a very huge corpus and it is difficult to use all the corpus to learn the model. In our experiments, we choose to use only the first 700MB of this prepared data corresponding to a size comparable to what was obtained for Vietnamese.

To compare web-based language models with conventional language models, we also used a corpus extracted from the newspaper *Le Monde* from year 1997 to 2001. This corpus is made of 716 MB, i.e. 4,323,629 sentences. This newspaper corpus is used by the majority of French ASR systems.

The Vietnamese data collection is a corpus representing more than 2.5 GB of web pages. After data preparation, the text corpus is made of 858 MB, i.e. 10,020,267 sentences.

4.2. Test corpora

The French test corpus is made up of two dialogs extracted from the NESPOLE! project[5] database [5]. They are related to a client/agent discussion for organizing holidays in Italy. Only the 216 client turns were kept for our experiments. The French vocabulary (20,000 words) is made of this task specific words plus the most frequent French words of WebFR4 corpus.

The Vietnamese test corpus is a translation of the French corpus. The Vietnamese vocabulary (20,000 words) is obtained with the methodology described in 3.2.1.

4.3. Filtering and SLM construction

In these experiments, we tried some solutions to filter the training corpora. In all cases, we selected sentences without size restriction. To filter, the following solutions were tested:

1. *all-sentences:* take all the text corpus (without any sentence filtering).
2. *block-based:* take only blocks which have at least 5 in-vocabulary words by block.
3. *sentence-based:* take all sentences containing only in-vocabulary words (no unknown words).
4. *hybrid:* take all sentences containing only in-vocabulary words (3) and apply minimal blocks filtering (2) on the rejected sentences.

To learn our language models, we use the SRILM toolkit [12] with a Good-Turing discounting and Katz backoff for smoothing method. It is very important to note that with this toolkit, the unknown words are removed in our case, since we are in the framework of closed-vocabulary models.

4.4. Results

The perplexities of the language models with these data filtering solutions are given in table 2.

The last two filtering methods have the best perplexities in our dialogue test because test corpus contains many short sentences.

Table 2 also shows that the perplexities of the language models according to corpus collected from Web are better than from journalistic source in our context of dialogue test. That means that the Web is a very rich source for spoken language modeling and that it can be successfully applied to model minority languages like Vietnamese even if the correspondence between perplexities of Vietnamese and French language models is not very significant here because each language have a particular characteristic.

[4] http://bushido.imag.fr/papillon/

[5] http://nespole.itc.it/

Table 2: *Perplexities of the language models*

Expe.	FR: Newspaper		FR: Web		VN: Web	
	Size (MB)	PPL	Size (MB)	PPL	Size (MB)	PPL
all	716	673	686	539	858	260
block	642	796	366	637	667	359
sent.	92	**513**	156	580	370	**252**
hybrid	644	687	411	**509**	729	259

4.5. Redundancy

We also noticed that the Vietnamese Web ressources contain some redundant information (menus, references, advertisements, announcements...) which is repeated in different pages. This is due to the day by day collecting of daily news which may have a direct influence on the performance of the language modeling.

Therefore, we tried to evaluate this redundancy in the part of the Vietnamese corpus from *Vietnam News Daily*[6] website (called *VnExpress*). So, we applied a redundancy filtering method before html2text module. The new perplexity figures with and without this redundancy filtering are given in table 3.

Table 3: *Influence of the redundant information*

Expe.	VN:Web original filter		VN:Web redun. filter	
	Size (MB)	PPL	Size (MB)	PPL
all	868	260	402	201
block	667	359	357	282
sent.	370	**252**	226	**195**
hybrid	729	259	373	199

By filtering the redundant information contained in the web pages collected from the same site, the training corpus size is reduced by 54% in our experiments. On the other hand, the perplexity value is significantly improved by 26%.

5. Conclusions and perspectives

An effective methodology for fast language model construction in minority languages is introduced in this paper. It consists in collecting Web sites and filtering the web pages using some generic tools. This methodology has been tested and validated using the Vietnamese minority language. In a first step, we have built the set of tools for the French majority language and validated the Web-based language model comparing to the classical newspaper-based language model. In a second step, we have adapted our tools to Vietnamese and we have defined which modules are fixed and which are specific of the target language. By collecting regularly two daily news web sites, a language model for Vietnamese was obtained in only three months. In our experiments, we have also presented an evaluation of perplexities for some proposed filtering methods in majority language (French) and in minority language (Vietnamese).

[6] http://vnexpress.net/

As future subjects, we will focus on the task-dependency language modeling (for example using a Web search engine) and we will improve these data filtering methods. Fast construction of acoustic model for minority language is also a very important and challenging part of our future work.

6. References

[1] J. Kunzmann, K. Choukri, E. Jahnke, A. Kiessling, K. Knill, L. Lamel, T. Schultz, and S. Yamamoto, "Portability of automatic speech recognition technology to new languages: Multilinguality issues and speech/text resources," in *ASRU*, Madonna di Campiglio, Italy, 2001.

[2] L. Lamel, "Some issues in speech recognizer portability, workshop on portability issues in human language technologies," in *LREC*, 2002.

[3] D. Vaufreydaz, C. Bergamini, J. F. Serignat, L. Besacier, and M. Akbar, "A new methodology for speech corpora definition from internet documents," in *LREC*, vol. I, Athens, Greece, 2000, pp. 423–426.

[4] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, 2001.

[5] D. Vaufreydaz, L. Besacier, C. Bergamini, and R. Lamy, "From generic to task-oriented speech recognition: French experience in the nespole! european project," in *ITRW Workshop on Adaptation Methods for Speech Recognition*, Sophia-Antipolis, France, 2001.

[6] V. Berment, "Several technical issues for building new lexical bases," in *Workshop Papillon*, Tokyo, Japon, 2002.

[7] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling. computer," *Speech and Language*, vol. 10, pp. 187–228, 1996.

[8] R. Ghani, R. Jones, and D. Mladenic, "Building minority language corpora by learning to generate web search queries," CMU, Tech. Rep. CMU-CALD-01-100, 2001.

[9] G. A. Monroe, J. C. French, and A. L. Powell, "Obtaining language models of web collections using query-based sampling techniques," in *HICSS*, 2002.

[10] R. Nisimura, K. Komatsu, Y. Kuroda, K. Nagatomo, A. Lee, H. Saruwatari, and K. Shikano, "Automatic n-gram language model creation from web resources," in *Eurospeech*, 2001.

[11] H. Doan-Nguyen, "Techniques génériques d'accumulations d'ensembles lexicaux structurés à partir de ressources dictionnairiques informatisées multilingues hétérogènes," INPG, Grenoble, Tech. Rep., 1998.

[12] A. Stolcke, "Srilm - an extensible language modeling toolkit, international conference spoken language processing," SRI, Denver, Colorado, Tech. Rep., 2002.