AUTOMATIC LEXICON ENHANCEMENT BY MEANS OF CORPUS TAGGING

Frédéric Béchet , Thierry Spriet , Marc El-Bèze
Laboratoire Informatique d'Avignon LIA - CERI
339, chemin des Meinajaries - BP 1228
84911 Avignon Cedex 9 - France
frederic.bechet@univ-avignon.fr

Abstract

Using specialised text corpus to automatically enhance a general lexicon is the aim of this study. Indeed, having lexicons which offer maximal cover on a specific topic is an important benefit in many applications of Automatic Speech and Natural Language Processing. The enhancement of these lexicons can be made automatic as big corpora of specialised texts are available.

A syntactic tagging process, based on 3-class and 3-gram language models, allows us to automatically allocate possible syntactic categories to the Out-Of-Vocabulary (OOV) words which are found in the corpus processed. These OOV words generally occur several times in the corpus, and a number of these occurrences can be important. By taking into account all the occurrences of an OOV word in a given text as a whole, we propose here a method for automatically extracting a specialised lexicon from a text corpus which is representative of a specific topic.

1 Introduction

With both Automatic Speech Processing and Natural Language Processing it is necessary to use a lexicon which associates each item with a certain number of characteristics (syntactic, morphologic, frequency, phonetic, etc.). In Speech Recognition, these lexicons are necessary in the lexical access phases and the language modelisation as they allow the association between lexical items and recognised sounds while maintaining syntactic coherence within the sentence under analysis. In Speech Synthesis, the grapheme-to-phoneme transcription phase uses morphological and syntactical information to constrain the phonetic transcription of the graphemes. In both cases, using lexicons which have the maximum information about the subject is an important benefit.

The actual performance of Automatic Speech Treatment systems often limits their application to smaller subject-areas of language (medical texts, economic articles, etc.). It is important to have specialised lexicons which cover these smaller subject-areas in order to optimise the synthesis or recognition applications. But although general lexicons are readily available now, this is not the case for specialised lexicons which contain, for example, technical terms relevant to a subject, or family and brand names as can be found in journalistic texts.

When working with corpora we are faced by the evolutionary aspects of a given language. The quicker the evolution of a specialised area, the more the dictionary will lack the ability to cover the subject, because a dictionary represents the state of a language at a given time. The words missing from a lexicon (which we refer to here as Out-Of-Vocabulary words or OOV words) represent a significant problem. In effect, whatever the size of the lexicon used, one can always find OOV words in texts. If, for a given word, the lexical access fails, this failure can affect the processing of the word as well as the processing of the contextual words.

It would be useful to have dynamic lexicons which evolve in accordance with the corpora processed in order to limit, as much as possible, the OOV words. Such an enhancement of lexicons could be automatic if big corpora of specialised texts were available : medical reports in an electronic form, newspaper available in CD-ROM, etc.

This interesting idea of automatically enhancing specialised lexicons from a general lexicon and a big corpus, is the aim of this paper. By using statistical language models, we show how to automatically assign one or several categories to the OOV words which are found in our corpora. Then, by taking into account all the occurences of each OOV word, we are able to automatically extract a new lexicon of OOV word with reliable labels associated to each word.

2 Processing OOV words

Various applications at LIA need a large lexicon, such as the automatic generation of graphical accents in a French text, language models for a dictation machine, the grapheme-to-phoneme transcription system, etc. As most of these applications process text corpus, the lexicon is mainly used through a syntactic labelling system developed at the laboratory (El-Bèze, 1995). This tagging system is based on a 3-class probabilistic language model which has been trained on a corpus of 39 million words contained in articles of the french newspaper *Le Monde*. The lexicon used is composed of 230 000 items.

The use of a big general dictionary allows us to limit most of the OOV words to one of these categories : proper names, composit words, unused flexions, neologisms, mistakes. The problem of missing roots becomes important when the texts processed belong to a different area than the one used during the building of the lexicon. This is the case in corpus dedicated to sub-areas of language, such as in technical documentation, for example.

Previous studies (Ueberla, 1995; Maltese, 1991) show that the modelling of OOV words improves significantly the performance of a language model. The presence of OOV words in the corpus can produce errors, not only in the form itself, but also in its context in the sentence. This is the reason why the syntactic tagging system has been endowed with a module, called Devin (Spriet, 1996), which proposes a category for each OOV word that is found.

The modules described here take into account all the simple OOV words, which are those composed with only alphabetical characters (no space, hyphen, digits, or special characters). A specific module dedicated to composite words is currently being developed. We classify these simple OOV words in two categories : the "proper-names", and the "common-words" which represent all the others ! By applying simple heuristics to a sentence we can separate the OOV words into proper-names and common-words.

3 Processing OOV common-words with the morpho-syntactic Devin

3.1 Out-of-context process

The goal of this module is to give a probability to syntactic labels which can represent the OOV common-words. These labels are distributed amongst 21 syntactic classes (adverbs, adjectives, names, verbs). It is commonly accepted that the ending of a word belonging to one of these classes influences strongly its syntactic category (Vergne, 1989; Guillet, 1989). Using this idea, we trained a statistical model with all the words from our dictionary. We make the hypothesis that this model will correctly work on unknown words, since these words should be governed by the same morphological principles.

The approach chosen is based on decision-trees (Breiman, 1984). An out-of-context evaluation of the morpho-syntactic Devin is presented in (Spriet, 1996).

3.2 Context analysis

The context analysis of OOV words permits the choice, from all the possible categories proposed by the Devin, of the one which best fits with the context of the OOV word. The hypotheses produced for each OOV word are inserted in the graph of possible categories generated by the language model. The 3-class analysis allows us to find the label which has the best probability.

We decided to test the module on a corpus containing "forced" OOV words. This means that we voluntarily removed from the lexicon a set of test words. The text corpus chosen contained 313 690 words of which 10 850 were "forced" OOV words (these 10 850 occurrences represent 3430 different forms).

In the first stage, we labelled this corpus without using the Devin. 1771 errors of context (as compared to the initial reference) were induced by the addition of 10 850 OOV words. Then we labelled again the same corpus, this time using the Devin. 88.3% of OOV words were correctly labelled (as compared to the initial reference) and 86.2% of induced contextual errors were corrected due to attributing a syntactic category to each OOV word. Thus, 87.5% of labelling differences with the initial reference were corrected by using the Devin.

It is important to point out that this type of evaluation does not take into account the errors which are intrinsic to the tagging system employed (about 4% as mentioned in (El-Bèze, 1995)). Indeed, the syntactic categories calculated by the Devin were compared to those produced by the tagger when these words belonged to the lexicon. Nevertheless the benefit of this technique is that it is automatic, which allows us to test our module on an important corpus of tests. A manual verification of a small corpus of "true" OOV words has also been carried out (Spriet, 1996), the results are appreciably similar.

4 Proper-names process

The second category of OOV words represents the forms which have been identified as proper-names. We separate these words into the following classes : family name, first name, town name, company name, country name. It is not possible to simply make a morphological module which allows us to process proper-names. Thus, the estimation of an out-of-context probability for each of these classes is independent of the graphical form of the proper-names. It is therefore the consideration of the context that allows us to attribute a reliable probability to the likelihood of an OOV proper-name belonging to a specific class. We present here a method based on a statistic 3-class model dedicated to OOV proper names.

4.1 Contextual Tagging using the Devin for proper-names

The general 3-class language model is, most of the time, unable to choose between the different categories of proper-names. In fact, when you have to decide whether an OOV word is a family name or a town name, the word-context of the OOV word is more useful than its syntactic-class-context. A 3-gram model seems natural for solving this problem. But, because we want to process OOV words, we use a 3-gram model specific to proper names where some categories of words are represented by their classes (all the proper names as well as punctuation and non-alphabetical words) while others are represented by their graphical form (all the other classes).

In the labelling process, when an OOV proper-name X_i appears at position i in the sentence, the label which is given to X_i represents the class which maximize $P(t/X_i)$, the probability of X_i belonging to the class t.

$$\tilde{t} = Arg \max_{t} P(t/M_1 \ldots X_i \ldots M_n)$$

$$\tilde{t} = Arg \max_{t} \frac{P_t(M_1 \ldots t \ldots M_n)}{\sum_j P(M_1 \ldots j \ldots M_n)}$$

We carried out similar experiments to those presented above. The test corpus was the same and we voluntarily removed 970 proper-names from the lexicon, which represented 5000 occurrences in the corpus. 86% of the OOV words had been correctly tagged by the proper-names language model.

It is important to point out that the average number of classes which can be attributed to a proper-name is very close to 1 (1.07 in our test corpus and 1.08 in the general lexicon). This shows that the comparison between the reference labels and the labels calculated is a true evaluation.

5 Automatic lexicon production

In studying all the occurrences, in all their contexts, of the OOV words of a corpus, we aim to automatically obtain new lexicons which represent the corpus studied.

As we have mentioned already, the syntactic tagger used was trained on a journalistic text corpus from the newspaper *Le Monde*. The test corpus chosen to validate our automatic lexicon enhancement method was composed with articles of the newspaper *Le Monde Diplomatique* from 1990 until 1995. This 6-million-word corpus contains a large amount of proper-names and technical terms relative to various subjects.

The test corpus contains 110 000 OOV words composed as follows :

- 22 766 OOV common-words (20.7%)
- 63 194 OOV proper-names (57.4%)
- 24 040 OOV composite words (21.8%)

The lack of static coverage of our general lexicon is 1.85% (0.38% for the OOV common-words and 1.06% for the OOV proper-names).

By tagging the corpus using Devin modules (for common-words and proper-names) we are able to automatically extract a lexicon of OOV words which contains, for each word, its number of occurrences as well as the list of labels which have been attributed to it during the tagging process. The list of labels given to each word of the lexicon is classified by frequency, as shown in the example below.

OOV word	Nb	C1	C2	C3	C4
tchétchène	41	AFS	AMS	NFS	NMS
		54%	32%	8%	6%

This frequency information allows us to filter the lexicon according to 2 criteria : number of occurences of each word ; percentage of occurences for each label given to a word.

5.1 Lexicon of common-words

For the OOV common-words, we reduce the lexicon to the words which have at least 4 occurences in the corpus, then we keep, for each word, only the syntactic labels which represent 80% of all the occurences of the word. We obtain a lexicon of 1032 items representing 44% of all the occurences of OOV common-words in our corpus.

5.2 Lexicon of proper-names

The lexicon of OOV proper-names is limited to the words which have at least 4 occurences in the corpus and for which the most frequent label has a frequency of at least 90%. Then we keep, for each word, only the most frequent label. The lexicon contains 2250 words representing 28.5% of all the occurences of OOV proper-names in our corpus.

5.3 Results

We verified manually the first 1000 most frequent OOV words of each filtered lexicon. The results are presented as follows : Table 1 shows, in the column "Correct", the percentage of OOV words where all the labels were correct ; the column "Wrong" indicates the percentage of words which were labelled with at least one incorrect tag.

Table 1	Correct	Wrong
Common-words	95.6%	4.4%
Proper-names	92.4%	7.6%

Table 2 details, for the common-words lexicon, the results obtained on the correct words. The column "All classes" shows the percentage of correct words which had all their possible syntactic categories in the lexicon. The column "Missing classes" indicates the percentage of correct words which could have received more syntactic categories than those stored in the lexicon.

Table 2	All classes	Missing classes
Common-words	79%	21%

These results show that the criteria used to filter the OOV lexicons allows us to produce reliable lexicons (only 4% of the OOV common-words contained label errors). By keeping the 1000 most frequent words of each lexicon, we reduced by 20% the lack of coverage of our general lexicon on all the test corpus.

6 Conclusion

The aim of this study was the automatic production of a lexicon from corpus dedicated to some specific areas. The results obtained satisfy this goal. Indeed, taking into account all the occurrences of the unknown words of a text corpus permits us to automatically produce lexicons containing, for each entry, a list of possible syntactic classes with frequency information.

The integration of these lexicons within a linguistic module, points out the problem of the dynamic adaptation of the language model. This should be dealt with by means of a cache-based language model (Kuhn, 1990). The resultant lexicons produced contain very few incorrect syntactic classes for each item which is represented in the corpus by a sufficient number of occurrences.

This lexicon-extraction module has been used within the Text-To-Speech system developed at LIA : before the grapheme-to-phoneme transcription phase, we first extract a lexicon of all the OOV words of the text to process. Then, we add this lexicon to our general lexicon and we use the syntactic labels given to each word to constrain the grapheme-to-phoneme transcription rules as well as the liaison-generation rules.

Finally, it is important to point out that the approach chosen in this study remains independent of the processed language, as long as the hypotheses made by the morpho-syntactic Devin are satisfied.

References

Breiman L., Friedman J., Olshen R., Stone C. 1984. Classification and Regression Trees Wadsworth Inc.

El-Bèze M., Spriet T. 1995. Integration de Contraintes Syntaxiques dans un Systeme d'Etiquetage Probabiliste In *TAL* , Vol. 6 N 1-2.

Guillet A. 1989. Reconnaissance des formes verbales avec un dictionnaire minimal In *Revue Langage* , Paris.

Kuhn R., De Mori R. 1990. A Cache-Based Natural Language Model for Speech Recognition In *IEEE* , Vol. 12 N.6, Juin 1990.

Maltese G., Mancini F. 1991. A technique to automatically assign parts-of-speech to words taking into account word-ending information through a probabilistic model In *Eurospeech 91* Genova, pp. 753-756.

Spriet T., Béchet F., El-Bèze M., de Loupy C, Khouri L. 1996. Traitement Automatique des Mots Inconnus In *TALN 96*, Marseille juin 1996.

Ueberla J.P. 1995. Analysing weaknesses of language models for speech recognition In *ICASSP 1995*, pp. 205-208.

Vergne J. 1989. Analyse morpho-syntaxique automatique sans dictionnaire These de doctorat de l'Universite de Paris 6, 8 juin 1989.