AUTOMATIC ASSIGNMENT OF PART-OF-SPEECH TO OUT-OF-VOCABULARY WORDS FOR TEXT-TO-SPEECH PROCESSING

F. Béchet and M. El-Bèze
Laboratoire Informatique d'Avignon - LIA
339, chemin des Meinajaries BP 1228 - 84911 Avignon Cedex 9 - FRANCE
E-mail : (frederic.bechet,marc.elbeze)@univ-avignon.fr

ABSTRACT

Working with large corpora of text highlights the need for the special treatment of Out-Of-Vocabulary (OOV) words. This paper describes a strategy for processing OOV words within a Text-To-Speech (TTS) framework of the French language.

A probabilistic module, called "Devin", guesses a Part-Of-Speech (POS) for each OOV word according to the morphological structure of the word and the context in which it occurs. These POS can be either syntactic or semantic. The semantic labels represent the categories of each proper-name (family name, town name, etc.) and its linguistic origin which has a strong influence on its pronunciation.

According to these POS, the system chooses the correct set of rules which will be employed by the rule-based grapheme-to-phoneme transcriber of the TTS system.

1. INTRODUCTION

The pronunciation of a text involves dealing with the specific ambiguities of the grapheme-to-phoneme transcription process of a given language. The first step in Text-To-Speech (TTS) processing consists of converting the text into a string of phonological symbols which can be pronounced by a speech synthesiser.

Whatever the technique employed, the graphical form in itself is inefficient in solving all the transcription ambiguities. Accordingly, syntactic and semantic information must be integrated into the grapheme-to-phoneme transcription process in order to deal with these ambiguities. This information is usually coded using the Part-Of-Speech (POS) associated with each word. In the TTS system for the French language developed by LIA, the POS can be either syntactic or semantic.

Working with large corpora of text highlights the need for the special treatment of Out-Of-Vocabulary (OOV) words. This paper describes a strategy for processing OOV words within a TTS framework of the French language. A probabilistic module, called "Devin", guesses a POS for each OOV word according to the morphological structure of the word and the context in which it occurs. According to these POS, the system chooses the correct set of rules which will be employed by the rule-based grapheme-to-phoneme transcriber.

2. PROCESSING OOV WORDS

Before being transcribed into phonemes, every text is tagged with the POS of each word. A large lexicon is necessary during this phase in order to process texts belonging to different subject-areas. The lexicon is used through a syntactic tagger developed at LIA [4].

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 newspaper *Le Monde*. The lexicon is composed of 230 000 items and we use a set of 103 syntactic classes.

The use of a big general dictionary allows us to limit most of the OOV words to one of these categories : proper names, composite 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.

Previous studies [6] [9] 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 [8], which proposes a POS for each OOV word that is found.

The modules described here take into account all the simple OOV words, which are those composed of 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". By applying simple heuristics to a sentence we can separate the OOV words into these two categories. Each category is processed by Devin : Syntactic Devin for the common-words and Proper-Name Devin for the others.

3. OOV PROPER-NAMES

3.1. Specific problems due to proper-names

We have first to determine which category of proper-names we will deal with in this study. One possible definition has been given in the framework of the European project ONOMASTICA [7] : a word is considered as a proper-name if it can be an entry in a phone directory. We limit this definition by considering only the words composed of alphabetical characters and which can represent a name of a person, a company or a place. The acronyms are processed by another module which decides whether or not they have to be spelt or read.

The pronunciation of proper-names points out some specific problems [10] :

• The lack of normalisation in the historic evolution of proper-names together with the presence of some archaic forms increases the ambiguity of some sequences of letters. For example, the sequence "*is*" become ambiguous in the town name "*Isle-sur-Sorgue*" where it is pronounced /i/.

• The addition of a determiner or a prefix to proper-names complicates the morphological segmentation of some words like *Montredon* (*Mont+redon*) and *Montreux*.

• Finally, the pronunciation of a proper-name is strongly linked with its linguistic origin [2]. This phenomena occurs in some French dialects and also in proper-names which have a foreign origin. To pronounce a foreign name, you have to guess its linguistic origin, and then adapt the pronunciation according to the phonetisation rules representative of this origin.

The proper-name process module presented in this study has several sets of grapheme-to-phoneme transcription rules. Each of these sets is representative of a given linguistic origin. So, if a proper-name has been labelled as French, the corresponding transcription rules set will take into account all the French specific phenomenon presented above. Similarly, a word labelled as English will be phonetised by rules which take into account the way a French speaker pronounces English words. We describe now the labelling process which gives to each proper name a semantic label and a linguistic origin.

3.2. Labelling proper-names

The Proper-Name Devin is composed of two phases :

• The first step in the labelling process consists of pointing out all the proper-names of a text by giving them a semantic label. All these labels can be used during the phonetisation process. This stage is based on a statistical language-model dedicated to proper-names.

• In the second phase, a morphological module allows us to guess a linguistic origin to a proper-name according to a statistic analysis.

3.2.1. Semantic labelling

We separate the OOV proper-names into the following classes : family name, first name, town name, company name, country name. 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 statistical 3-class model dedicated to OOV 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. 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 maximise $P(t/X_i)$, the probability of X_i belonging to the class t.

$$\tilde{t} = \underset{t}{\text{Argmax}}\ P(t|M_1 \ldots X_i \ldots M_n)$$

$$\tilde{t} = \underset{t}{\text{Argmax}}\ \frac{P_t(M_1 \ldots t \ldots M_n)}{\sum_j P(M_1 \ldots j \ldots M_n)}$$

Formula 1 : proper-names language model

3.2.2. Guessing a linguistic origin

A n-gram probabilistic model, based on sequences of letters, calculates (for each proper-name) the probability of that name having a specific linguistic origin (French, English, German, Spanish, etc.).

This model has been trained using a corpus of 10 000 proper-names extracted from articles of the French newspaper *Le Monde*. These proper-names have been classified according to some features representative of their pronunciation. This classification empirically determines eight linguistic sets which differ by their pronunciation.

We have trained, for each of these groups, a 3-letter statistical model on the 10 000 proper-names corpus. After this learning process, it is possible to calculate the probability of a proper-name belonging to one of these sets.

This probability is calculated as follows :

The linguistic set chosen for a proper-name *m* composed of the letters $c_1c_2\ldots c_l$, is the one which maximise the probabilities *P(i/m)* for all the classes *i* :

$$\tilde{i} = \underset{i}{\text{Argmax}}\ P(i|m) = \underset{i}{\text{Argmax}}\ P(m|i) \times P(i)$$

with

$$P(m|i) = \prod_{n=0}^{l} \partial_1 \times P_i(c_n|c_{n+1}c_{n+2}) + \partial_2 \times P_i(c_n|c_{n+1}) + \partial_3 \times P_i(c_n)$$

and $\partial_1 = 0,7 \quad \partial_2 = 0,2 \quad \partial_3 = 0,1$

Formula 2 : the 3-letter model

As shown in formula 2, the 3-letter model is linearly combined to a 2-letter and a 1-letter model by means of coefficients experimentally obtained.

4. OOV COMMON-WORDS

In a TTS application, it is important to attribute a syntactic label to each OOV common word in order to eliminate two kinds of ambiguity :

• in French, many liaison-generation rules are based on syntactic criteria ;

• the pronunciation of some sequences of letters is dependent on the syntactic category of the word in which they occur (the suffix "-ent" for example).

The Syntactic Devin calculates the probability of an OOV word belonging to a specific syntactic class as follows.

4.1. 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. 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 [3]. An out-of-context evaluation of the morpho-syntactic Devin is presented in [8].

4.2. 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.

5. EXPERIMENTS

We carried out some experiments in order to evaluate our Devin modules. We present the tests performed concerning the tagging of OOV words and an evaluation of the contribution of the Devin modules to the Text-to-Speech system.

5.1. Tests with "forced" OOV words

We decided, as a start, to test our methods 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.
3430 common-words and 1000 proper-names were removed from the lexicon, causing 15 850 "forced" OOV words. In the first stage, we labelled this corpus without using the Devin. Then we labelled again the same corpus, this time using the Devin. 88.3% of the OOV common-words and 86% of the OOV proper-names were correctly labelled.

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 [4]). 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 [8], the results are appreciably similar.

5.2. Contribution to the TTS process

All the modules have been integrated into the TTS system of the LIA which is currently being evaluated in the AUPELF test campaign of French language processing systems. The results of this test campaign will point-out the contribution of our OOV words process modules to the grapheme-to-phoneme transcription of texts. Nevertheless, we present a first evaluation of the contribution of our Devin modules to the phonetisation process on two aspects of this process.

5.2.1. Pronunciation of the suffix "-ent"

In French common words, the suffix "-ent" is ambiguous : it is pronounced as a schwa if the word is a verb but it is pronounced /ã/ if the word belongs to another syntactic category.

We extracted from a 6-million-word test corpus (from the French newspaper *Le Monde Diplomatique*) a list of 426 OOV common-words which have this suffix. 13% of them were verbs in the 3rd person plural which means that the suffix must be pronounced as a schwa.

100% of these words were correctly separated into verbs and other categories by the Devin module. By using these labels, the grapheme-to-phoneme transcription system correctly processed this set of OOV common-words

5.2.2 Pronunciation of proper-names

The lack of proper-names corpus with phonetic and semantic information prevents us from making a large evaluation of our techniques. Nevertheless, we carried out an evaluation on the 100 most frequent sequence first name + family name found in the newspaper Le Monde *Diplomatique*. 90% of them were correctly phonetised by using the linguistic origin labels given by the DEVIN module.

6. CONCLUSION

The aim of this study was the attribution of POS to OOV words in order to process them within a TTS system. The Devin modules achieved this goal for two categories of OOV words : the proper-names and the common-words.

The Syntactic Devin gives to the OOV common-words a syntactic labels which is used in the liaison generation between words and the phonetic transcription of some sequences of letters.

The Proper-Name Devin guesses a linguistic origin to each OOV proper-name ; this label is then used by the grapheme-to-phoneme transcription module for choosing the correct set of transcription rules.

The tests showed the good performance of this tagging process. The contribution of the Devin modules to the global performances of the TTS system will be measured during the Aupelf test campaign.

The good results obtained by these techniques lead us to consider other fields in Natural Language Processing. By taking into account all the occurrences of each OOV word in a corpus specific to a given subject, we are be able to automatically extract a new tagged lexicon characteristic of this subject [1]. The lexicon obtained can then be integrated in a language model for speech recognition by using a cache-based model [5].

REFERENCES

[1] Béchet F., Spriet T., El-Bèze M. 1997 *Automatic lexicon enhancement by means of corpus tagging* ACL/EACL Workshop on Spoken Dialogue Systems, Madrid

[2] Belrhali R. 1995 *Phonetisation automatique d'un lexique general du francais : systemique et emergence linguistique*, These de l'Universite Stendhal de Grenoble, ICP 1995.

[3] Breiman L., Friedman J., Olshen R., Stone C. 1984 *Classification and Regression Trees* - Wadsworth.

[4] El-Beze M., Spriet T. 1995 *Intégration de Contraintes Syntaxiques dans un Système d'Etiquetage Probabiliste*, TAL, Vol. 6 N 1-2.

[5] Kuhn R., DeMori R. 1990 *A Cache-Based Natural Language Model for Speech Recognition*, IEEE Transactions on pattern analysis and Machine Intelligence Vol. 12 No. 6.

[6] 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*, Eurospeech 91, Genova.

[7] Schmidt M., Fitt S., Scott C., Jack M. 1993. *Phonetic transcription standards for European names (ONOMASTICA)*. *Eurospeech'93* Berlin.

[8] Spriet T., Béchet F., El-Bèze M., de Loupy C., Khouri L. 1996 *Traitement Automatique des Mots Inconnus* in TALN96, Marseille.

[9] Ueberla J.P. 1995 *Analysing weaknesses of language models for speech recognition*, ICASSP

[10] Yvon F. 1996 *Prononcer par analogie : motivation, formalisation et évaluation*. Thèse de doctorat en informatique de l'ENST.