LARGE SPAN STATISTICAL LANGUAGE MODELS : APPLICATION TO HOMOPHONE DISAMBIGUATION FOR LARGE VOCABULARY SPEECH RECOGNITION IN FRENCH

Frédéric Béchet, Alexis Nasr, Thierry Spriet, Renato de Mori
LIA - University of Avignon
BP 1228 – 84911 Avignon Cedex 9 - France
frederic.bechet@lia.univ-avignon.fr
www.lia.univ-avignon.fr

ABSTRACT

Homophone words is one of the specific problems of Automatic Speech Recognition (ASR) in French. Moreover, this phenomenon is particularly high for some inflections like the singular/plural inflection (72% of the 40.7K lemma of our 240K word dictionary have inflected forms which are homophonic). In order to take into account word-dependencies spanning over a variable number of words, it is interesting to merge local language models, like 3-gram or 3-class models, with large-span models. We present in this paper two kinds of models : a phrase-based model, using phrases obtained from a training corpus by means of a finite-state parser; a homophone cache-based model, using derivation of constraints from word histories stored in a cache memory.

1. INTRODUCTION

Words having different orthography with the same phonetic transcription are *homophone* among themselves. Words in the same homophone class may differ because of their number and gender, the syntactic category they belong to or their meaning.
Quantity, size and frequency of homophone classes are language-dependent parameters. In French, an analysis of a large corpus of newspaper articles has shown that each word, on average, belongs to an homophone class of 2.2 elements.

A word acoustic model provides the same likelihood for all the words in a class of homophones. Thus, it is up to the language model (LM) to select a linguistically correct sequence of words, given a sequence of hypothesised homophone classes, in such a way that just one word is selected for each of these homophone classes.

N-gram LMs provide sequences of local constraints which induce a number of linguistically correct distance constraints. Nevertheless, some of the induced constraints result to be linguistically incorrect, while some other essential constraints are not established. This leads to the generation of maximum likelihood sentence hypotheses which are syntactically incorrect or semantically unacceptable.

The ability of an LM to correctly remove ambiguities due to the presence of various candidates in homophone classes can be evaluated without an Automatic Speech Recognition (ASR) system. In fact, a Text-To-Speech (TTS) system can generate the phonetic transcription of a text. This phonetic transcription can then be translated into a sequence of word hypotheses by a search process.

The knowledge used by such a search is a statistical LM and a collection of homophone classes obtained from the phonetic representation of the lexicon. The sequence of words obtained by this search process is then compared by an elastic matching algorithm with the original sequence from which the phonetic transcription has been generated. Substitution errors are used to compute Word Error Rates (WER).

Section 2 introduces homophone graphs which are the content of the search space for homophone disambiguation. Section 3 introduces new LMs. Section 4 provides experimental results discussed in Section 5.

2. PROCESSING HOMOPHONE GRAPHS

The analysis of a French lexicon containing 40.7K lemma has shown that in 72% of the cases, a flexion from singular to plural results in an homophone. On the other hand, there is evidence

that homophone words are a major source of error in ASR systems for dictation in French and most of these errors correspond to number flexion.
For the above reason, the research discussed in this paper focuses on the correct transcription of singular/plural homophones. The transcription is obtained by a stack-decoding search process which uses a single Knowledge Source (KS) consisting of a Language Model (LM). The search space is a word-graph where each singular/plural homophone has been replaced by both its inflected forms obtained with a grapheme-to-phoneme transcription system described in [1].

The maximum likelihood sentence hypothesis is then aligned and compared with the correct orthographic transcription in order to obtain counts for insertion (I), deletion (D) and substitution (S) errors.
A mixture of probabilities obtained with four LMs is used to score hypotheses. The LMs are:

- M1 : 3-gram LM,
- M2 : 3-class LM with 105 Part-Of-Speech (POS),
- M3 : phrase-based model,
- M4 : cache-based model.

The last two models are large-span LMs and will be described in the following section.

3. LANGUAGE MODELS

3.1. Phrase-based model

In order to represent long distance constraints avoiding data sparseness, sequences of words are grouped into units called phrases.
These units can be obtained by an approach based on stochastic grammars [5] or based on purely statistic criteria [3] [4]. Our approach uses both knowledge-based and stochastic methods.
In a first step, the phrases to be used in the third LM are selected, using a 40M word training corpus consisting of articles from the French newspaper *Le Monde* which was tagged with a statistic tagger described in [7]. Examples of the 105 POS used are 'Masculin Singular Name' (NMS), 'Verb 3rd Plural' (V3P), etc.

The tagged corpus is then parsed with a finite-state parser to recognise syntactic phrases like nominal, verbal or prepositional syntagms. Information about number and gender of the syntagms is kept when relevant.

The grammar used contains about 80 rules and its coverage is voluntarily low for two main reasons : first, the aim of this grammar is to compute a shallow parsing of sentences in order to detect basic syntagms between which exist number or gender agreements. Second, this prevents possibly ambiguous attachments like prepositional phrase attachment. A label is then assigned to each phrase according to its syntactic structure, e.g., GVS for singular verbal syntagm. The set of phrase labels contains 70 items.

Each phrase-label is associated to a set of phrase-patterns which is a sequence of POSs. A total of 6000 most frequent phrase-patterns were selected from the result of the parsing process.
These patterns represent a new grammar, subset of the original grammar, which is used by the decoding process in a deterministic way : when a sentence-hypothesis is evaluated, POSs are assigned to words according to the 3-class model. The sentence is further parsed using the POS and the phrase grammar by systematically choosing the longest phrase-patterns which match the sentence. The likelihood of a sentence-hypothesis is computed as the linear combination of the probabilities of the 3-gram LM on the words, the 3-class LM on the POS and the 3-class LM on the phrases. The example in Table 1 shows the result of the correct analysis of a sentence.

Table 1 - parsing example

word	POS	phrase	1	2	3
quand	COSUB	COSUB			
d'	DETFP				
authentiques	AFP	NFP	*	*	*
valeurs	NFP			*	*
de	PREPADE				
justice	NMS	GP	*	*	*
ne	ADVNE				
constituent	V3P	VP			*
plus	ADV				
le	DETMS				
fondement	NMS	NMS	*	*	*
des	PREPDES				
lois	NFP	GP	*	*	*
c'	PPER3MS	PPER3MS			
est	VE3S	VS			
souvent	ADV				
l'	DETMS				
arbitraire	NMS	NMS	*	*	*
qui	PRELMS	PRELMS			
les	PPOBJMP				
remplace	V3S	VS		*	*

The words in bold are singular/plural homophones and the last three columns correspond to the decoding results produced by using only one of the three LMs : the 3-gram model on words in column 1, the 3-class model on POS in column 2 and the 3-class phrase-based model in column 3. The correctly disambiguated homophones are marked with a star '*'.
In table I, the 3-class POS model realises the agreement between the verb "*constituent*" and the noun "*justice*" instead of its subject "*valeurs*". By grouping the words "*de justice*" into a prepositional syntagm and the words "*d'authentiques valeurs*" into a nominal syntagm, the phrase-based model brings the verb closer to its subject and fulfils the number agreement.

However, there are some cases which are difficult to process with phrase-based models and more generally syntactic-based models. These "difficult" cases can be classified in two different sets :
- syntactic constraints not captured by simple grammars (overlapping prepositional syntagms or relative clause, co-ordinate clauses, etc.) ;
- syntactically undecidable or really ambiguous cases;

A solution to the first problem would be a full syntactic parsing. But, such a parsing is very difficult to integrate in a speech decoding process due to coverage and complexity..

The second problem refers to number agreement when lexical or semantic information is essential to remove ambiguities. For example, in the sentence :
Le président Boris Eltsine dans un message de vœux ***diffusé*** *à la télévision russe*
The number agreement between *'diffusé'* and *'message'* (singular) rather than *'voeux'* (plural) can't be predicted by a syntactic model.

There is another kind of problems which is specific to LMs used in Speech Recognition Systems. Due to the impossibility to have a 0% WER, substitutions, insertions and deletions errors which occur during the decoding phase make a full syntactic parsing nearly impossible. Moreover, adding strong syntactic constraints to the decoding process of a sentence which can have some errors can lead to increase dramatically the WER of the system.

All these reasons lead us to propose a decision-model, robust to speech recognition errors, which can take a decision on the number of a homophone word without strong syntactic constraints. This is the cache-based model presented in the following section.

3.2. Cache-based model

This model consists in storing, for each singular/plural homophone word, its left contexts as seen in the training corpus. These contexts are word histories made of the last ten words stored in a cache memory [6]. Each cache content C(w) is a vector whose components are the syntactic POSs assigned to the words by the tagger. The size of the vectors corresponds to the number of POS which is 105.
The training of this model consists in using the training corpus for updating two cache memory vectors for each homophone. The first vector CP(w) corresponds to the contexts where the inflected form of lemma *w* was plural and the second one, CS(w), corresponds to the contexts of the singular flexion of the same lemma *w*. During the decoding process, when two singular/plural homophones of the same lemma *w* are in competition, two distances are computed : one between CP(w) and the current cache and the other between CS(w) and the cache.
The distance used is a symmetric Kullback-Leibler divergence measure [2]. When the difference between these two distances is higher than a threshold estimated on the training corpus, the system chooses the flexion corresponding to the smaller distance.

In Table 2, an example is provided on the use of the cache model. The word *'diffusé'* is a singular/plural homophone and represents the value of *w* in the discussion above. In the context of sentence 1, it can either be singular or plural, depending on the agreement with *'message'* or *'voeux'*. A cache memory vector, called A(w), is obtained based on the 9 words preceding the word *'diffusé'*. This vector is then compared with the two vectors, CP(w) and CS(w), associated to the homophone *w='diffusé'* . By using a threshold *th* the following condition is considered:

$$|\mathrm{dist}(A,CS)-\mathrm{dist}(A,CP)| > th$$

If it is satisfied, then the flexion is selected whose vector is the closest to the current vector A(w). In this example, the vector CS(w) results to have the minimum distance. It represents the singular flexion of the word *'diffusé'*.

Table 2 - POSs for sentence 1

1	Le	DETMS	6	un	DETMS
2	président	NMS	7	message	NMS
3	Boris	XPRE	8	de	PREP
4	Eltsine	XFAM	9	vœux	NMP
5	dans	PREP	10	diffusé *or* diffusés	VPPMS *or* VPPMP

Figure 1 shows the 2 cache memory vectors CS(w) and CP(w) for the homophone word "*diffusé/diffusés*" as well as the current cache vector A(w) calculated on the sentence of table 2. The 9 POS of the words preceding the homophone in the sentence are stored in the corresponding components of the vector.

A(w) : current cache memory vector

1	2	3	4	...	....	102	103	104	105
1	0	1	2	0	0	2	1	2	0

CS(w) : Cache memory vector for : *'diffusé'*

1	2	3	4	...	....	102	103	104	105
0.8	1.2	2.2	1.6	0.6	1.0	2.8	0.6	0.6	0.2

CP(w) : Cache memory vector for : *'diffusés'*

1	2	3	4	...	....	102	103	104	105
0.2	0.8	1.2	0.6	2.2	1.8	1.0	0.8	0.1	0.1

Figure 1 - Cache memory vectors used for solving the ambiguity of the example of table 2.

4. EVALUATION

A 80K words corpus of articles from the newspaper *Le Monde Diplomatique* has been used for testing the proposed methodology for homophone disambiguation. For each sentence a phonemic transcription has been generated using a TTS component and a graph of possible graphemic transcriptions has been generated in order to take into account all ambiguities arising from singular/plural homophones. The best sequence of words has been obtained with a stack decoding search using scores obtained with various combinations of LMs.

Tables 3 shows results in terms of word accuracy (WA) for 17.4K singular/plural homophones in the test corpus. Each column corresponds to the use of a single LM.

LM	M1	M2	M3	M4
WA	90.95	95.36	89.02	84.59

Table 3 – WA obtained with each model separately

LM	M1+M2	M1+M3	M2+M3
WA	96.89	96.14	96.22

LM	M1+M2+M3	M1+M2+M3+M4
WA	96.98	97.36

Table 4 – Results with model combination

5. DISCUSSION AND CONCLUSIONS

The results show the benefits of model combination for homophone disambiguation and confirm that different models capture properties which are in some cases complementary. Models M3 and M4 are less precise than n-grams. In fact M3 uses only 70 classes and M4 does not capture all the syntactic constraints because the size of the cache is small and the cache is applied only to homophones. Nevertheless, they cover some useful cases not covered by the other models. The four models can be used for refining word hypotheses generated in a first pass of a progressive search performed by an ASR system.

REFERENCES

[1] Béchet F., Derderian S., El-Bèze M. (1995) *Conversion graphème-phonème automatique : le système GRIPHON*. IA 95, Montpellier.

[2] Bigi B., De Mori R., El-Bèze M., Spriet T. (1998) *Detecting topic shifts using a cache memory*, ICSLP'1998, pp. 2331-2334.

[3] Deligne S., Bimbot F.(1995). *Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams.* ICASSP 95

[4] Deligne S., Sagisaka Y. (1998), *Learning a syntagmatic and paradigmatic structure from language data with a bi-multigram model* Coling-ACL'98, Montreal, pp 300-306

[5] Gillett J., Wayne W. (1998) *A Language Model Combining Trigrams and Stochastic Context-free Grammars.* ICSLP 1998, pp. 2319-2322

[6] Kuhn R., De Mori R (1990) *A Cache Based Natural Language Model for speech Recognition* IEEE Trans. Pattern anal. Machine Intell., PAMI-12(6):570-582.

[7] Spriet T., El-bèze M. (1995) *Etiquetage probabiliste et contraintes syntaxiques. TALN 95.*