A LANGUAGE MODEL COMBINING N-GRAMS AND STOCHASTIC FINITE STATE AUTOMATA*

Alexis Nasr, Yannick Estève, Frédéric Béchet, Thierry Spriet, Renato de Mori
LIA - University of Avignon
BP 1228 - 84911 Avignon Cedex 9 - France
alexis.nasr@lia.univ-avignon.fr

1 ABSTRACT

This paper describes a new kind of language models composed of several local models and a general model linking the local models together. Local models describe more finely subparts of the textual data than a conventional n-gram trained on the complete corpus. They are built on lexical and syntactic criteria. Both local and global models are integrated in a single hidden Markov model. Experiments showed a 14% decrease in perplexity compared to a bigram model on a small corpus of telephonic communications.

2 INTRODUCTION

Language models described in this paper are based on the intuitive idea that different parts of sentences in a given corpus could be described more finely by local models rather than by a conventional n-gram model trained on the entire corpus, especially when not enough data is provided to train a trigram. We propose to implement this idea by bringing together some word sequences appearing in a corpus. Every grouping will constitute a subcorpus, we will call a *class*, on which a local model is trained. The structure of the word sequences gathered together in a class corresponds to well formed syntactic units or phrases, as noun phrases, prepositional phrases or verb phrases, of variable length. Phrases are gathered together according to the similarity of their distibution in a training corpus. Two advantages are expected from this kind of models. The first is better reaction to data sparseness. The reason here is the same put forward with word class language models: the probability of occurrence of some word w in a given context could be very low although the occurrence of the class of w is quite common in the same context. The same is true of variable length phrases and classes of phrases. The second advantage is a better modelling of long distance dependencies, which are not modelled in a bigram or trigram model. Gathering into classes sequences of words for language modelling is already discussed in [2], [5] and [4]. The distinguishing feature of our approach is the use of syntactic constraints in the definition of word sequences. Other criteria for gathering words into phrases could be used, as in [3].

3 STRUCTURE OF THE LANGUAGE MODEL

The language model is composed of a variable number of local models and a general model, a bigram in our case, which predicts the occurrence probability of a class knowing the preceding class. Each local model gives the probability of some word sequence knowing a class. Formally a local model is represented as a tree-structured weighted finite state automaton (WFSA) which edges are labelled with words. A successful path in the WFSA represents a phrase of the class. Both the bigram and the WFSAs can be put together and regarded as a hidden Markov model[1]. Figure 1 shows a subpart of it. In the figure, the probabilities of the external model are denoted by P while the probabilities of the single local model are denoted by $PA1$.
The general model being an hidden Markov Model, the probability of a word sequence according to such a model is the sum of the probabilities of all the paths corresponding to the word sequence. There might be several paths since a given word sequence forming a phrase can appear in different classes. Taking into account in the linguistic score of a sentence all its possible segmentations in word sequences of variable length is discussed in [1].

The construction of the model takes place in two stages: the construction of phrase classes and their representation as WFSA, followed by the parameter estimation of the external model. These two stages are described below. The overall process is graphically represented as a block diagram in figure 3.

*This research is supported by France Telecom's center for research and development (CNET) under the contract 971B427

[1] The sum of the weights of the transitions exiting a state in each WFSA equals 1.

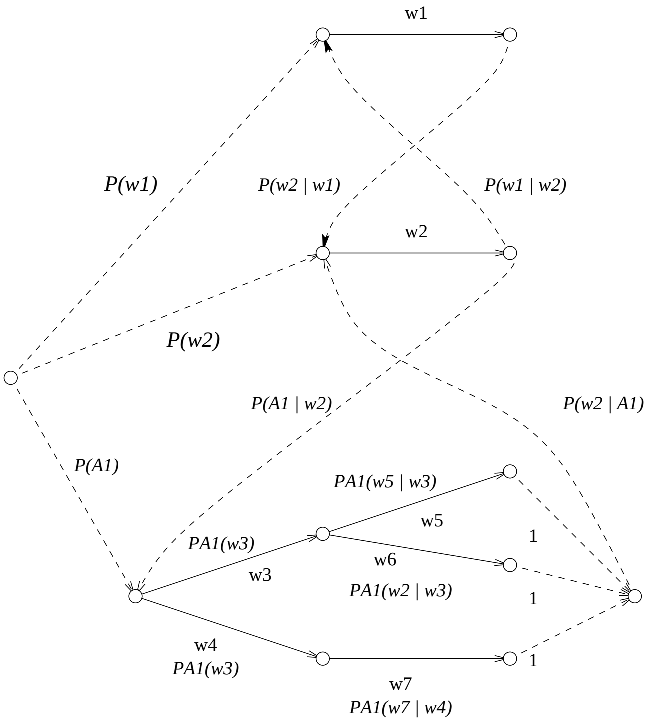

Figure 1: The language model as a hidden Markov model

4 CONSTRUCTION OF THE MODEL

The first step of the class construction consists in partially parsing the training corpus in order to recognize sequences of words as phrases. The training corpus is first annotated with part of speech tags. The tagging is performed with a stochastic part of speech tagger [6] using a tagset of 103 different tags. At the end of this process, each word of the corpus is associated to its most probable part of speech. The annotated corpus is then partially parsed using a greedy finite state partial parser. The parser gathers together adjacent words composing a phrase of a given type (noun phrase, verb phrase . . .). The grammar used by the parser is partial in the sense that it is not design to parse whole sentences. Different grammars will recognize phrases of different natures and lengths. The coverage of the grammar is therefore an important parameter in the phrase classes construction process, as will be made clearer in section 5. The parser is greedy in the sense that, when parsing an ambiguous structure, it will chose the first analysis according to the ordering of the rules in the grammar. At the end of this stage, the corpus is composed of phrases possibly reduced to single words.

The second step is the construction of the initial phrase classes. It consist in grouping together into classes phrases of the same category appearing in identical contexts. The context of a phrase is reduced to the L phrases appearing to its left and the R phrases appearing to its right. Every class is therefore associated to a context. A counter is associated to each phrase P of the class C, which is incremented every time an occurrence of P appears in the context associated to C in the training corpus. The sum of the counters of the different phrases appearing in a class defines the *weight* of the class. After the classes have been built those having a weight inferior to a threshold, called the *minimun weight threshold*, are discarded.

V being the vocabulary of phrases recognized in the corpus, a class can be regarded as a point in the real $| V |$-dimentional space (the P-th coordinate of the point being the frequency of the phrase P in the class). An example of verb group class is represented below. It is originally in French, the translation in English was added for readability purpose.

0.69	je voudrais	I want
0.03	je cherche	I am looking for
0.03	je voudrais avoir	I would like to have
0.07	je voudrais connai3tre	I would like to know
0.05	je veux	I want
0.02	je souhaite avoir	I wish to have
0.04	je recherche	I am looking for
0.02	je voudrais savoir	I would like to know
0.02	j' aimerais connai3tre	I wish to know

The third step consists in merging together classes having a close internal distribution (frequency of the phrases in the class). The merging is realized by an iterative algorithm which merges together the two closest classes, according to the following distance.

$$d(C_1, C_2) = \sum_{P \in V} | C_1(P) - C_2(P) |$$

Where $C_1(P)$ and $C_2(P)$ denote the frequency of the phrase P in the classes C_1 and C_2, or the P-th coordinate of the points C_1 and C_2. The algorithm stops when the distance between the two closest classes in superior to a threshold, called the *merging threshold*. When two classes are merged together, their respective contexts are also merged: if the two classes C_1 and C_2 are merged to form the class C_3, the contexts associated to the class C_3 is the union of the contexts of C_1 and C_2. After the class merging, each class is represented as a weighted finite state automaton which constitutes the local model corresponding to this class. The automaton corresponding to the class above is shown in figure 2. The weights on the transitions have been omitted for the sake of readability.

After the classes have been built, an identifier is associated to every class. Every occurrence of a

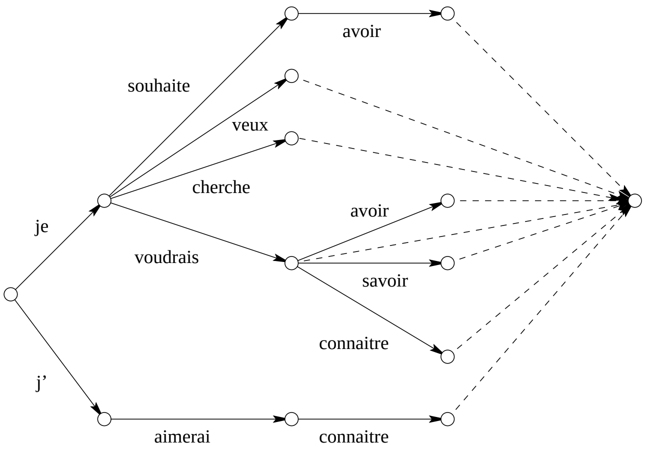

Figure 2: A phrase class as a WFSA

phrase P appearing in the partially parsed corpus and belonging to a class C is replaced by the identifier of C when the context of P matches one of the contexts associated to C. At the end of this process, a hybrid corpus has been created, composed of words and class identifiers[2]. Eventually, a bigram is trained on this corpus using the CMU-Cambridge Statistical Language Modeling Toolkit (`www.speech.cs.cmu.edu/speech/SLM_info`). This bigram represents the external model.

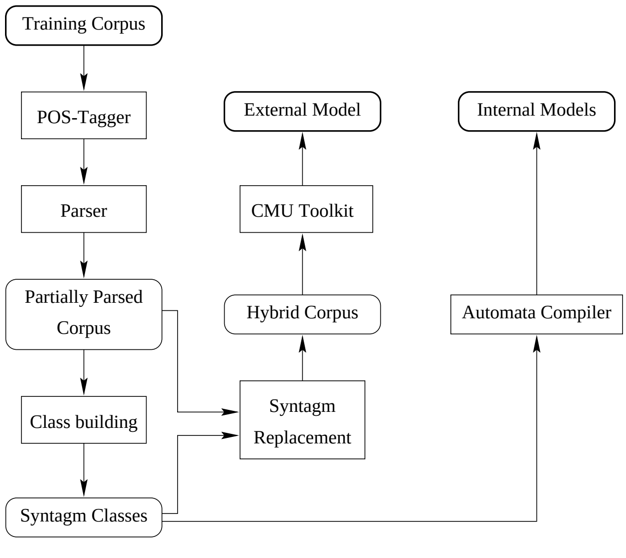

Figure 3: A block diagram of the model construction

5 EXPERIMENTS AND RESULTS

The experiments have been conducted CNET's (France Telecom center for research and development) AGS corpus of telephonic communications composed of 64K words and having a vocabulary of 822 words. The test corpus is composed of 7K words. The partial parsing was conducted using a grammar of low coverage, composed of 69 rules. The minimun weight threshold was set to 10 and the merging thershold to 1.5. Perplexity results computed on train set and test set appear in Table 1. The first two columns concern the parts of the corpora composed by sentence of length greater or equal to eight words which are the sentences primarily concerned by this study.

	Test≥8	Train≥8	Test	Train
LM-WFSA	7.21	5.75	7.20	5.92
2-gram	8.38	6.66	7.86	6.54
gain	14.2%	13.6%	8%	9.4%

Table 1: Perplexity results

We have represented in tables 2 and 3 the probabilities given by the model language to a sentence which is quite representative of our corpus. The sentence is *je recherche un emploi à l'étranger* which translates as *I am looking for a job abroad*. The probability given to the sentence by a bigram trained on the complete corpus (refered to as the conventional bigram) is $5.043e-07$ while the probability given by our language model is $3.603e-05$. This increase of the probability is due to the recognition of three phrases in the sentence: the verb phrase *je recherche*, the noun phrase *un emploi* and the prepositional phrase *à l'étranger*. The best path associated to the sentence in the hidden Markov model goes through three atomata (`C_18 C_33 C_24`). The probability of the path is the product of the inter class probabilities (given by the external model) and the intra class probabilites (given by each of the local models). It is higher than the conventional bigram probability because inter class probabilities are globally better than their conventional bigram counterparts, so are the class internal probabilities. We have represented in table 2 the external probabilities (`P(C_18 | <s>), P(C_33 | C_18) ...`). The last column shows the corresponding conventional bigram probabilities, namely, the probability of the first word of the phrase given the last word of the previous phrase (`P(je | <s>), P(un | recherche) ...`).

phrase	class	ext. P	bigram P
je recherche	C_18	2.29e-2	1.41e-1
un emploi	C_33	4.57e-1	3.98e-1
a2 l' e1tranger	C_24	1.38e-1	4.79e-2
</s>		9.8e-1	7.41e-1

Table 2: Inter class probabilities

[2] The phrases which have not been replaced by a class identifier are broken down into words.

Table 3 shows the probabilities of each of the three phrases (in bold) given by their corresponding local model and by the conventional bigram, in the last column. The other phrases of each class have been added for reader's curiosity.

plass	phrase	local P	bigram P
C_18	je voudrais	9.46e-3	5.41e-1
	je cherche	9.05e-1	2.09e-1
	je recherche	**8.54e-2**	**4.00e-2**
C_33	un travail	3.14e-2	1.11e-2
	un emploi	**9.43e-1**	**2.89e-1**
	un petit-boulot	2.52e-2	6.84e-2
C_24	pour la re1gion Midi-Pyre1ne1es	1.02e-1	1.74e-3
	dans la re1gion parisienne	1.02e-1	3.63e-2
	a2 l' e1tranger	**3.27e-1**	**2.34e-2**
	en France	4.69e-1	7.01e-1

Table 3: Intra class probabilities

Conversely, some sentences which have good bigram probabilities will be given lower probabilities by our language model, as for example the sentence *je recherche un petit-boulot dans la re1gion parisienne* (*I am looking for a job in the region of Paris*). Although the class internal probabilities of the phrases composing this sentence are better than their conventional bigram equivalent, the inter class probabilities showed to be lower explaining the decrease from $3.14e - 06$, for conventional bigram probability, to $8.04e - 07$ for our language model.

As described in section 4, several parameters influence the class extraction stage. Some are numerical: the context span, the minimum class weight threshold and the merging threshold, while other are not: the grammar coverage. Different combinations of these parameters give birth to different local models and therefore different language models. Several combinations of the parameters values have been tried that resulted in different perplexity figures, some were lower than the bigram model, other were higher. We did not try yet to optimize the parameters.

The experiments conducted until now showed the importance of two complementary features of the classes. The first is the average length of the phrases in the classes, the second is the class weight after merging. Long phrases will generally be given better probability by local models than by the bigram model, they tend therefore to lower the perplexity. On the other hand long phrases will have low occurrence in the training corpus, they will therefore give rise to light weight classes which will tend to have low probabilities in the external model and therefore increase perplexity. There is hence an important tradeoff between average phrase length of a class and its weight. The parameters mentioned above will have different influences on these two variables. Grammar coverage will tend to increase phrase length. Increasing the context span will tend to decrease class weight. Conversely, increasing merging threshold will tend to increase class weight.

6 CONCLUSION

The perplexity decrease, although moderate, shows the potential benefits of combining local language models in the proposed way. Future work will concern optimization of the class construction parameter set, as well as some refinements to the class construction algorithm. These include an iterative algorithm which will iteratively extract classes based on the hybrid corpus (the corpus composed of words and class identifiers) produced by the previous class extraction stage. The result of this algorithm will be to increase the class weight. Other refinements concern smoothing of the local models. Eventually, the integration of this language model in an automatic speech recognition system is planned.

References

[1] Frederic Bimbot and Sabine Deligne. Language modelling by variable length sequences: theoretical formulation and evaluation of multigrams. In *ICASSP*, pages 169–172, 1995.

[2] Sabine Deligne and Yoshinori Sagisaka. Learning a syntagmatic and paradigmatic structure from language data with a bi-multigram model. In *A-CL*, pages 300–306, 1998.

[3] Peter Heeman and Geraldine Damnati. Deriving phrase-based language models. In *Proc. IEEE Workshop on Automatic Speech Recognition*, pages 41–47, Santa Barbara, CA, 1997.

[4] T. R. Niesler and P. C. Woodland. Variable-length category n-gram language models. *Computer Speech and Language*, 13:99–124, 1999.

[5] Klaus Ries, Finn Dag Buø, and Alex Waibel. Class phrase models for language modeling. In *ICSLP*, volume 1, pages 398–401, Philadelphia, PA, 1996.

[6] Thierry Spriet and Marc El-bèze. Etiquetage probabiliste et contraintes syntaxiques. In *TALN*, 1995.