Semantic cache model driven speech recognition

Benjamin Lecouteux, Pascal Nocera, Georges Linarès

SEMANTIC CACHE MODEL DRIVEN SPEECH RECOGNITION

Benjamin Lecouteux, Pascal Nocera, Georges Linarès

LIA-CERI, university of Avignon (France)

ABSTRACT

This paper proposes an improved semantic based cache model: our method boils down to using the first pass of the ASR system, associated to confidence scores and semantic fields, for driving the second pass. In previous papers, we had introduced a Driven Decoding Algorithm (DDA), which allows us to combine speech recognition systems, by guiding the search algorithm of a primary ASR system by the one-best hypothesis of an auxiliary system. We propose a strategy using DDA to drive a semantic cache, according to the confidence measures. The combination between semantic-cache and DDA optimizes the new decoding process, like an unsupervised language model adaptation. Experiments evaluate the proposed method on 8 hours of speech. Results show that semantic-DDA yields significant improvements to the baseline: we obtain a 4% word error rate relative improvement without acoustic adaptation, and 1.9% after adaptation with a 3xRT ASR system.

Index Terms— speech recognition, driven decoding, Latent Semantic Analysis, cache model

1. INTRODUCTION

Although n-gram models have demonstrated their efficiency in speech recognition systems, they are known to be able to capture, neither long-term dependencies nor semantic level information. Many papers addressed this issue in the past, mainly by modifying n-gram probabilities according to long-term statistics or topic-specific models [1]. The main approaches are:

- The cache model: introduced by [2], it proposes to increase the probability of the words that have occurred recently. The assumption behind this model is that if a word is used in a context, then this word is likely to be used again.
- The trigger model, as explored by [3, 4]: the long-distance dependency problem is addressed by interpolating n-grams with trigger pairs selected according to their mutual information. This approach could be thought of as a generalization of cache models.
- The topic mixture model: [5, 6, 7, 8, 9] propose techniques to use mixtures of language models according to topics. The training data are partitioned into a small set of topic clusters, which are used for adaptation. [10] introduce a framework for constructing language models by exploiting both local and global constraints based on Latent Semantic Analysis (LSA). This approach proposes to compute two models based on LSA and n-grams. The models are combined, which leads to the integration of semantic information in n-gram models, with respect to LSA-found topics.
- Combination between cache and mixtures models: In [11, 12], the authors propose to combine these approaches to capture distant dependencies in language.

This research is supported by the ANR (Agence Nationale de la Recherche), AVISON project.

However, cache models lack robustness, because boosted words are depending on the current hypothesis: an error can easily spread. Furthermore, these models rely only on the past (current hypothesis history) and too many hypotheses can be made with all trigger pairs: they are not easy to tune. Topic-based language models generally estimate weights on a first pass transcription or current hypothesis, without special word confidence selection. Moreover, the main issue of these models is the topic selection/detection for training.

In these papers additional information provided by the ASR system itself (confidence measures etc.) is not used. More generally, confidence measures are used for acoustic models adaptation and unsupervised training, but not for language model adaptation. Some papers propose to use confidence measures for dynamically changing the weighting between the acoustic and the language model during the search process, depending on the confidence of the current language model history. In [13, 14], the authors use word confidence or posteriors directly into the graph exploration, which yields improvements in the system performance.

Our objective is to exploit all the potential of a first decoding pass into the next search process and to provide an unsupervised language model adaptation. We propose to apply a suitable cache model during the decoding process, exploiting both LSA information and previous pass confidence measures.

In this paper, we present an integrated method to drive an ASR system with its confidence measures associated to an LSA-cache model. Our strategy focuses only on words poorly recognized, in order to reduce the noise introduced. In a previous paper [15], we had proposed an algorithm which consisted in integrating the output of an auxiliary ASR system into the search algorithm of a primary system. We present an extension of this algorithm, devoted to the decoding process itself, by using the output of the first pass to drive the second pass according to word confidence scores and to the semantic information associated. The first section presents the entire system. In the second section, we present the experimental framework on 8 hours extracted from the ESTER campaign [16]. The last section presents experiments on our semantic driven decoding. Finally, we conclude and suggest some potential applications and improvements.

2. INTEGRATED APPROACH: DDA DEVOTED TO LSA CACHE

The initial Driven Decoding Algorithm (DDA) consists in integrating the output of an auxiliary system in the search algorithm of a primary system. This integration relies on two steps. Firstly, the current hypothesis of the primary system and the auxiliary transcript are aligned by minimizing the edit distance. Then, linguistic probabilities are combined, according to posteriors and to an hypothesis-to-transcript matching score. We propose to modify the DDA, in order to obtain an improved driven cache language model. The next sub-sections provide details on the system parts.

2.1. A* search algorithm in the Speeral system

The LIA speech recognizer is used as a primary system. It is based on the A* search algorithm operating on a phoneme lattice. The

decoding process relies on the estimate function $F(h_n)$, which evaluates the probability of the hypothesis h_n crossing the node n:

$$F(h_n) = g(h_n) + p(h_n), \tag{1}$$

where $g(h_n)$ is the probability of the current partial hypothesis up to node n, which results from the partial exploration of the search graph, and $p(h_n)$ is the probe that estimates the remaining probability from the current node n to the last node. In order to be able to take into account information resulting from the output of an auxiliary system, the linguistic part of g in (1) is modified according to the auxiliary hypothesis as described below.

2.2. Driven Decoding Algorithm

The Speeral ASR system generates word hypotheses as the phoneme lattice is explored. The best hypotheses at time t are extended according to the current hypothesis probability and the probe results. In order to combine the information provided by the injected transcript H_{aux} and the main search process, a synchronization point has to be found for each word node that the engine evaluates. These points are found by dynamically mapping the provided transcripts to the current hypothesis; this is accomplished by minimizing the edit distance between the provided transcripts and the current hypothesis. This process allows one to identify, in the injected transcript H_{aux}, the best sub-sequence that matches the current hypothesis h_{cur}. This sub-sequence, denoted by h_{aux}, is used for a new estimation of the linguistic score, according to posteriors $\phi(w_i)$:

$$\begin{array}{l} L(w_i|w_{i-2}, w_{i-1}) = P(w_i|w_{i-2}, w_{i-1})^{1-\beta}.\phi(w_i)^{\beta} \\ \beta = 0 \text{ if } w_i \text{ does not belong to } H_{aux} \end{array} \tag{2}$$

where $L(w_i|w_{i-2}, w_{i-1})$ is the resulting linguistic score, $P(w_i|w_{i-2}, w_{i-1})$ the initial probability of the trigram, β an empirical fudge factor and $\phi(w_i)$ is the confidence score of w_i.

2.3. Confidence measures

For LSA-DDA, we use our own confidence measures. They are computed in two stages. The first one extracts low-level features related to the acoustic and search graph topology, and high-level features related to linguistic information. Then, a first *bad word* detection hypothesis is produced by a classifier that is based on the boosting algorithm. Each word from the hypothesis is represented by a feature vector composed of 23 features, that are grouped into 3 classes.

We use **acoustic features** that consist of the acoustic log-likelihood of the word, the average log-likelihood per frame, the difference between the word log-likelihood and the unconstrained acoustic decoding score of the corresponding speech segment. The **linguistic features** are based on probabilities estimated by the 3-gram language model used in the ASR system. We use the 3-gram probability, the perplexity of the word in the window, and the unigram probability. We also add an index that represents the current back-off level of the target word. The **graph features** are based on the analysis of the word confusion networks: the number of alternative paths in the word section and values related to the distribution of posterior probabilities in the word section.

We use a boosting classification algorithm in order to combine word features, as detailed in [17]. The algorithm consists in an exhaustive search for a linear combination of classifiers by overweighting misclassified examples.

The classifier is trained on a specific training corpus, that was not included in the ASR system training. Each word from this corpus is tagged as *not ok* or *ok* word, according to the ASR system reference. The classification results in two classes for each word: *not-ok* words and *ok* words.

Confidence Error Rate is 19.5% on dev and 18.6% on test for a 0.5 score threshold, while Normalized Cross Entropy is respectively 0.373 and 0.282. A 0.85 threshold will be used for high confidence decision-making in the LSA module: 55% of good words are selected, with 2.7% of mis-selected words.

2.4. Latent Semantic Analysis module

Latent Semantic Analysis [10] is a technique that allows one to associate words that tend to co-occur within documents with a semantic relation. The assumption is that co-occurring words within the same document are semantically related. LSA takes the vector space representation of documents based on term frequencies as a starting point and applies a dimension reducing linear projection. The mapping is determined by a given document collection and is based on a Singular Value Decomposition (SVD) of the corresponding term/document matrix.

In our system, a semantically consistent word sequence may be considered as unexpected by the ASR language model due to the limitations of the n-gram language models. Moreover, n-gram models are unable to take advantage of long-range dependencies in natural language.

In our experiments, Latent Semantic Analysis fields are trained on the corpus used for language model estimation. For better coverage, the entire corpus is lemmatized, and its vocabulary is limited to the ASR system lemmatized lexicon (about 33K words). Moreover, a stop-list removes all non relevant words.

When a word is requested, the LSA module returns the top 100 words associated to their LSA confidence score.

2.5. Word selection for LSA

In order to make the ASR system sensitive to good semantic fields, words are selected according to their confidence score. In the next experiments, the threshold is fixed to 0.85. At 0.85, about 55% of the corpus is selected, while the confidence error rate on the detected good words is about 3%. Then, for each selected word, the top 100 words are extracted with the LSA system. For each segment a bag of words is computed, in order to be used by the search algorithm.

2.6. DDA combination with LSA

Driven decoding is relevant in the context of our discussion, because a simple trigger LSA model introduces a lot of noise. Then, the system is driven by its previous hypothesis to avoid deviations for correct words. The LSA trigger is applied when the word confidence score of the previous hypothesis is less than 0.5: correct words are well preserved. Moreover, words associated to low confidence scores are undervalued: this allows the ASR system to explore other paths.

The final system, as detailed in Figure 1, works as an improved cache model. The LSA-DDA becomes:

$$\begin{array}{l} \phi(w_i) \geq 0.5 : \left\{ \begin{array}{l} L(w_i|w_{i-2}..) = P(w_i|w_{i-2}..)^{1-\beta}.\phi(w_i)^{\beta} \\ \beta = 0 \text{ if } \phi(w_i) \text{ belong to to } H_{aux} \end{array} \right. \\ \phi(cw_i) < 0.5 : \left\{ \begin{array}{l} L(w_i|w_{i-2}..) = P(w_i|w_{i-2}..)^{1-\alpha}.\theta(w_i)^{\alpha} \\ \alpha = 0 \text{ if } \theta(w_i) \text{ not found in word bag} \end{array} \right. \end{array} \tag{3}$$

where $L(w_i|w_{i-2}, w_{i-1})$ is the resulting linguistic score, $P(w_i|w_{i-2}, w_{i-1})$ the initial probability of the trigram, β and α are empirical fudges factor, $\phi(w_i)$ is the confidence score of w_i word, cw_i is the aligned word after history (w_{i-2}, w_{i-1}) in auxiliary transcript, and $\theta(w_i)$ is the LSA score of w_i.

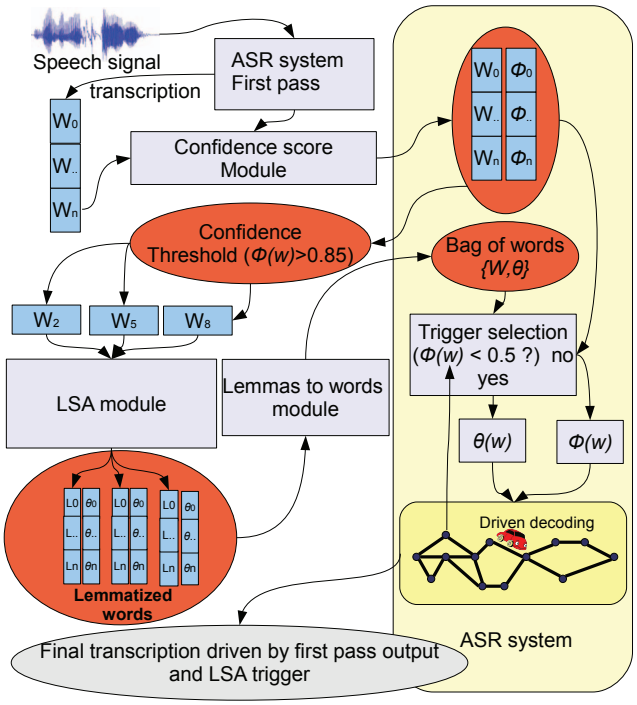

Fig. 1. Principle of the LSA Driven Decoding Algorithm

3. EXPERIMENTAL FRAMEWORK

3.1. Evaluation corpus

Experiments have been carried out in the framework of the French ESTER evaluation campaign [16]. The ESTER corpus contains French and Moroccan radio broadcast news , including some ad-hoc interviews, non-native speakers, on-the-fly translations, etc. Results are reported on a test corpus of 8 hours from six broadcasters, extracted from the official ESTER development set. The training data consists of 80 hours of manually transcribed audio data, corresponding to 1M words, and about 200M words from the French newspaper *Le Monde*.

3.2. The LIA broadcast news transcription system

The LIA broadcast news transcription system relies on the Speeral decoder and the Alize-based segmenter. Cross-word context-dependent acoustic models with 230k Gaussians are used. State tying is achieved by decision trees. The language models are classical 3-grams with a vocabulary of 65K words. The system runs two passes. The first one (3xReal Time) provides intermediate transcripts, which are used for MLLR adaptation. The second one is 3xRT or 10xRT, according to the level of pruning.

4. EXPERIMENTS

4.1. ASR system driven by its first pass

We have tested a second pass driven by the first one associated to confidence measures. The result was not surprising: the ASR system converges to the first pass and there is no significant change. This is because there are no alternatives proposed to the search algorithm.

4.2. LSA-cache without driven decoding

These experiments investigate the use of the LSA trigger without constraints. For each speech segment, a bag of words is associated. We have implemented a simple cache: during the search process, each hypothesis word is searched for in the LSA packet. When a match is found, the word is boosted according to the LSA confidence score. Experimental results are not interesting, because a general degradation is observed, despite of the LSA-specific information.

One reason could explain this disappointing loss: too many words are boosted, and the introduced noise impacts on the final result. This aspect highlights the difficulty to easily tune a cache model.

4.3. LSA trigger with driven decoding

These experiments are carried out without performing any acoustic adaptation, with a 3xRT ASR system. The system uses the previous one-best hypothesis associated to confidence scores. The trigger process is only applied to low-confidence words. We expect to drive the system with correct words of the one-best hypothesis, while wrong words are potentially rescored with the LSA cache. The results reported in Table 1 show that a significant gain can be obtained using the LSA Driven decoding system. The Word Error Rate (WER) is reduced by 4% relatively, compared to the first pass: the DDA allows the LSA-cache model to focus only on previous errors. This strategy avoids the introduction of too much noise. We observe better improvement (5% relative) for the worst show.

Show	#Hours	P1 3RT	P2 3RT	P2-LSA DDA 3RT
Classique	1h	21.4 %	20.8 %	20.9 %
Culture	1h	34.0 %	31.9 %	33.3 %
INTER	1h	22.7 %	22.0 %	21.6 %
INFO	2h	25.8 %	24.6 %	25.0 %
RFI	1h	28.6 %	26.0 %	27.1 %
RTM	2h	35.4 %	32.3 %	33.6 %

Table 1. LSA-driven decoding without acoustic adaptation: Baseline first pass (P1 3RT), baseline second pass with acoustic adaptation (P2 3RT) and second pass with LSA cache-model driven by the first pass (P2-LSA DDA 3RT) without acoustic model adaptation

4.4. LSA-DDA and acoustic adaptation (3xRT)

These experiments combine acoustic adaptation for the second pass (3xReal Time) with LSA-driven decoding. These experiments test the complementarity of LSA-DDA with the acoustic adaptation. Results are reported in Table 2: the WER is reduced by 1.9% relative, but less than without acoustic adaptation. LSA driven complements the Maximum Likelihood Linear Regression (MLLR) process, as a language model adaptation. We observe again better improvement (4.6% relative) for the worst show.

4.5. LSA-DDA and acoustic adaptation (10xRT)

These last experiments test the system with full graph exploration in second pass (10xRT). Results are reported in Table 3: the WER is reduced by only 1.1% relative: the complementarity between acoustic model adaptation and language model adaptation is small. However, the worst show present a 3.8% relative improvement. These results show that LSA-DDA strategy is more interesting in the context of a 3xRT ASR system where results are very closed to the 10xRT ASR system.

Show	#Hours	P2 3RT	P2-LSA DDA 3RT
Classique	1h	20.8 %	20,5 %
Culture	1h	31.9 %	31.8 %
INTER	1h	22.0 %	21.6 %
INFO	2h	24.6 %	24.5 %
RFI	1h	26.0 %	25.5 %
RTM	2h	32.3 %	30.8 %

Table 2. Baseline second pass 3xRT (P2 3RT), LSA-driven decoding with acoustic adaptation 3xRT (P2-LSA DDA 3RT)

Show	#Hours	P2 10RT	P2-LSA DDA 3RT	P2-LSA DDA 10RT
Classique	1h	20.2 %	20,5 %	20.0 %
Culture	1h	31.7 %	31.8 %	31.5 %
INTER	1h	21.6 %	21.6 %	21.6 %
INFO	2h	24.0 %	24.5 %	23.9 %
RFI	1h	25.4 %	25.5 %	25.3 %
RTM	2h	31.7 %	30.8 %	30.5 %

Table 3. Baseline second pass with acoustic model adaptation 10xRT (P2 10RT), LSA-driven decoding with acoustic adaptation 3xRT (P2-LSA DDA 3RT), LSA-driven decoding with acoustic adaptation (P2-LSA DDA 10RT)

The LSA-DDA allows for significant improvements on the 3xRT system, especially on the RTM show. With the 10xRT system, the WER improvement is slight, except on RTM. Unlike others shows, RTM is a Moroccan broadcast news: the training data for the language model are derived from a French newspaper. Hence, the model coverage is lower for RTM. This aspect explains the most significant gains and shows the contribution of the LSA model.

5. CONCLUSION

We proposed a semantic-driven decoding that allows for the integration of the first pass of the ASR system into the search algorithm in order to apply a semantic cache model. Our strategy focuses a Latent Semantic Analysis (LSA) cache model only on previous errors: according to the confidence scores, the LSA cache is applied or not. Experiments show that this integrated LSA-driven decoding improves the initial system and complements the acoustic adaptation. The enrichment of the first pass with confidence scores is useful for fine-tuning the search algorithm, while LSA enrichment allows one to select alternative paths when confidence scores are low: the strategy can be likened to an unsupervised language model adaptation. Moreover, due to the driven decoding, the tuning for a well working system is light. Finally, this method allows for a gain of 4% of relative WER on the first pass without acoustic adaptation, 1.9% on the 3xRT system, after acoustic adaptation and 1.1% on the 10xRT system. Nevertheless, these results show that the strategy is interesting for the 3xRT system where the WER is close to the 10xRT baseline system, resulting in a highly reduced computational cost. Moreover, the most constant gains are observed with the show farthest from training datas.

Presently, these experiments are limited to the segment granularity. We plan to extend this approach to sets of segments (discussion between two speakers, etc). We also plan to integrate external semantic data, such as abstracts of spoken documents, or meta-data (titles, speaker, etc.). Furthermore, we would also like to apply this method to more particular domains (e.g. medical surgeries)

6. REFERENCES

[1] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here ?," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, Aug. 2000.

[2] R. Kuhn and R. De Mori, "A cache-based natural language model for speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 6, pp. 570–583, 1990.

[3] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer Speech and Language*, vol. 10, no. 3, pp. 187–228, 1996.

[4] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modeling," in *Computer Speech and Language, vol. 8, pp. 1-38,*, 1994.

[5] Yoshihiko Gotoh and Steve Renals, "Topic-based mixture language modelling," *Natural Language Engineering*, vol. 5, pp. 355–375, 1999.

[6] Daniel Gildea and Thomas Hofmann, "Topic-based language models using em," in *Eurospeech*, 1999.

[7] N. Singh-Miller and C. Collins, "Trigger-based language modeling using a loss-sensitive perceptron algorithm," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, 15–20 April 2007, vol. 4, pp. 25–28.

[8] Hsuan-Sheng Chin and B. Chen, "Word topical mixture models for dynamic language model adaptation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, 15–20 April 2007, vol. 4, pp. IV–169–IV–172.

[9] C. Martins, A. Teixeira, and J. Neto, "Dynamic language modeling for a daily broadcast news transcription system," in *Proc. ASRU Automatic Speech Recognition & Understanding IEEE Workshop on*, 2007, pp. 165–170.

[10] J.R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, 2000.

[11] P.R. Clarkson and A.J. Robinson, "Language model adaptation using mixtures and an exponentially decaying cache," in *IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, 1997, vol. 2, pp. 799–802 vol.2.

[12] R.M. Iyer and M. Ostendorf, "Modeling long distance dependence in language: topic mixtures versus dynamic cache models," *IEEE transactions on speech and audio processing*, vol. 7, no. 1, pp. 30–39, Jan. 1999.

[13] P. Fetter, F. Dandurand, and P. Regel-Brietzmann, "Word graph rescoring using confidence measures," in *Fourth International Conference on Spoken Language ICSLP*, 1996, vol. 1, pp. 10–13 vol.1.

[14] F. Wessel, F. Wessel, K. Macherey, and R. Schluter, "Using word probabilities as confidence measures," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, K. Macherey, Ed., 1998, vol. 1, pp. 225–228 vol.1.

[15] B. Lecouteux, G. Linares, Y. Esteve, and G. Gravier, "Generalized driven decoding for speech recognition system combination," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008*, 2008, pp. 1549–1552.

[16] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase 2 evaluation campaign for the rich transcription of french broadcast news," in *Proc. of the European Conf. on Speech Communication and Technology*, 2005.

[17] P. Moreno, B. Logan, and B. Raj, "A boosting approach for confidence scoring," in *Interspeech, Aalborg, Denmark*, 2001, pp. 2109–2112.