Latent Semantic Analysis based Language Models for Meetings

Michael Pucher[1,2] and Yan Huang[1,3]

[1] International Computer Science Institute, Berkeley
[2] ftw. Telecommunications Research Center, Vienna
[3] University of California, Berkeley
{pucher,yan}@icsi.berkeley.edu

1 Introduction

Language models that combine N-gram models with Latent Semantic Analysis (LSA) based models have been successfully applied for conversational speech recognition [3] and for the Wall Street Journal recognition task [1]. LSA defines a semantic similarity space using a training corpus. This semantic similarity can be used for dealing with long distance dependencies, which are an inherent problem for traditional word-based N-gram models. Since LSA models adapt dynamically to topics, and meetings have clear topics, we think that these models can improve speech recognition accuracy on meetings. This poster presents first perplexity results for our experiments on using LSA models in combination with N-gram models for language modeling for meetings.

2 Initial Work

For the training and testing of our models we used the ICSI meeting corpus. The training set contains about 730K words, the test set contains 37K words. The vocabulary size is 10K. We used the meeting boundaries as document boundaries, which are needed for the training of the LSA model.

In LSA first the training corpus is encoded as word-document co-ocurrance matrix (using weighted term-frequency), which has high dimension and is very sparse. Then a semantic space with much lower dimension is constructed using Singular Value Decomposition (SVD). In this semantic space the similarity between words and documents is defined.

Since we need a probability for the integration with the N-gram models, the similarity is converted into a probability by normalizing it. Because this conversion is seen as a weakness of LSA models, other methods have been proposed that derive the probabilities directly [4]. We extended the small dynamic range of the similarity function by introducing a temperature parameter. Additionally we defined the concept of a pseudo-document, which is needed because the model is used to compare words with documents that have not been seen so far.

In the definitions presented in Table 1, below, λ_w is a word-dependent parameter defined using the word entropy [3], λ is a fixed (trained) constant interpolation weight, K_{sim} is the similarity between a word and a document, $K_{\min}$ is the

minimum similarity for a document, and $\propto$ denotes that the result is normalized by the sum over the whole vocabulary.

For the interpolation of N-gram models and LSA models, we used three different interpolation methods. The *information weighted geometric mean*, the *similarity modulated N-gram* and simple *linear* interpolation. The *information weighted geometric mean* interpolation represents a loglinear interpolation of normalized LSA probabilities and the standard N-gram, weighted by λ_w. The *similarity modulated N-gram* interpolation uses K_{sim} and K_{min} directly, without converting it into a probability first. K_{min} is substracted, so that the resulting similarity is nonnegative. Table 1 shows the perplexity results for the different methods. While the *linear* interpolation and the *similarity modulated N-gram* do not bring any improvements over the baseline trigram model, the *information weighted geometric mean* interpolation reduces perplexity. The improvement of the *information weighted geometric mean* interpolation over the trigram model is consistent with findings in [2] and [3].

Model	Definition	Perplexity
Trigram (baseline)	$P_{N-\text{gram}}$	84.3
Information weighted geometric mean interpolation	$\propto P_{\text{LSA}}^{\lambda_w} * P_{N-\text{gram}}^{1-\lambda_w}$	81.7
Similarity modulated N-gram interpolation	$\propto (K_{\text{sim}} - K_{\text{min}}) * P_{N-\text{gram}}$	85.1
Linear interpolation	$\lambda P_{\text{LSA}} + (1-\lambda) P_{N-\text{gram}}$	88.2

Table 1. Perplexity results for the ICSI corpus

In further experiments we plan to test larger models on a variety of meeting corpora (AMI, NIST, CMU). We also want to use topic boundaries as document boundaries, instead of meetings boundaries. Furthermore a different methodology for the conversion from similarities to probabilities will be investigated.

We believe that the results achieved so far make it promising to continue to work on these kinds of models for the meeting domain.

References

1. Bellegarda, J.R.: Large Vocabulary Speech Recognition with Multispan Statistical Language Models. IEEE Transactions on Speech and Audio Processing, Vol. 8, Nr. 1, January 2000
2. Coccaro, N., Jurafsky, D.: Towards better Integration of Semantic Predictors in Statistical Language Modeling. Proc. of ICSLP, Sydney, 1998
3. Deng, Y., Khudanpur, S.: Latent Semantic Information in Maximum Entropy Language Models for Conversational Speech Recognition. Proc. of HLT-NAACL, pp. 56-63, Edmonton, 2003
4. Gildea, D., Hofmann, T.: Topic-Based Language Models using EM. Proc. of Eurospeech99, Budapest, 1999