Appeared in Proc. NNSP'02, Sept.4-6, Martigny, Switzerland, 2002

# LANGUAGE MODEL ADAPTATION IN SPEECH RECOGNITION USING DOCUMENT MAPS

Krista Lagus and Mikko Kurimo
Helsinki University of Technology
Neural Networks Research Centre
P.O.Box 9800, 02015 HUT, Finland
E-mail: krista.lagus,mikko.kurimo@hut.fi

**Abstract. In this paper we present speech experiments that were carried out to evaluate a topically focusing language model in large vocabulary speech recognition. An ordered topical clustering is first computed as a self-organized mapping of a large document collection. Language models are then trained for each text cluster or for several neighboring clusters. The obtained organized collection of language models is efficiently utilized in continuous speech recognition to concentrate on the model that corresponds closest to the current topic of discussion. The speech recognition experiments are carried out on a novel Finnish speech database. A property of Finnish that is particularly challenging for speech recognition is the extremely fast vocabulary growth that makes many of the standard word-based language modeling methods impractical for large vocabulary tasks.**

## INTRODUCTION

Statistical language models assign probability estimates to sequences of words. Such estimates are essential, e.g., for speech or optical character recognition, spelling correction, handwriting recognition and statistical machine translation. For example, in continuous speech recognition with large vocabulary and several possible speakers it becomes exceedingly difficult to interpret the acoustic signal into the correct word sequence, and misinterpretations are common. Some form of knowledge about the possible structures within a language can greatly aid this interpretation. Such knowledge can be provided as estimates of probability of each word sequence considered likely based on the acoustic signal.

The estimation of large statistical language models has recently become possible due to the large data sets now available. However, a language model needed to provide estimates for every 5-word sequence from a vocabulary of 60,000 words, for example, would require the estimation and storage of as many as $7.7 * 10^{23}$ probabilities. Due to the sparsity of language, most of the word sequences in fact never even appear in practice.

The $n$-gram model with its variants has turned out to be very hard to consistently improve upon for varying corpora, especially in English. However, the model has several well-known drawbacks: (1) an observation of a word sequence does not affect the prediction of the same words in a different order, (2) long-term dependencies between words do not affect predictions, and (3) very large vocabularies pose a computational challenge. In languages with syntactically less strict word order and a rich inflectional morphology, such as Finnish, these problems are particularly severe.

We propose a practical and efficient method for benefiting from the sparsity that yields savings in processing time and memory usage during the operation of the model. The presented topically focusing language model utilizes the discussion history to concentrate onto the relevant topic-specific sub-model. The topic-specific sub-models are obtained by unsupervised organization of a large text corpus onto a topical document map, and then by creating a local language model for each map area. The ensuing model dynamically adapts to the current topic and uses a small, focused vocabulary.

### Related work

Information regarding long-term dependencies in language can be incorporated into language models in several ways. For example, in word caches [2] the probabilities of words seen recently are increased. In word trigger models [14] probabilities of word pairs are modeled regardless of their exact relative positions. Mixtures of sentence-level topic-specific models have been applied together with dynamic $n$-gram cache models with some perplexity reductions [8]. In [5] and [1] EM and SVD algorithms are employed to define topic mixtures, but there the topic models only provide good estimates for the content word unigrams which are not very powerful language models as such.

Perplexity improvements have been achieved when these methods are applied together with the general trigram models. However, it is generally much more difficult to obtain reductions in word error rate (WER) during speech recognition.

## ADAPTATION USING A TOPICALLY FOCUSING LM

### Construction of the document map

The document map is created using the WEBSOM method [7, 11], which applies the Self-Organizing Map (SOM) [10] algorithm for clustering short documents onto an ordered map. Due to the document encoding and the SOM process, each document obtains a location on the map near other documents that are similar in content.

In short, the method is as follows: Encode a large collection of short documents using the bag-of-words model with word weighting. Rare words and a stoplist of common words are excluded. The document vectors are condensed for computational reasons by applying the random projection method [11]. Finally, the document vectors are automatically ordered on a self-organizing map. Speedup methods may be applied in the construction of large maps.

When the corpus consists of longer documents (say, over 500 words on the average), such as books or novels, they can be segmented into short documents prior to creation of the map. It is preferable that the segments are cut at likely points of topical change, say at sentence or paragraph borders—bearing this in mind one may use a windowing of about 100 words, for example.

The WEBSOM document maps created for text exploration have usually contained about 10 documents per map unit. For language models, however, we have found a much larger ratio 100-400 documents per map unit to be suitable (200 was used in these experiments).

**Creation of sub-models for each map region**

For each map region, create an ordinary $n$-gram language model based on the documents falling in that region. A map region may consist of a single SOM unit or of several neighboring units. In the latter case, the regions may overlap so that there is still one sub-model centered at each map unit.

The benefits of this latter approach are obvious: there is sufficient amount of text for reliable estimation of the sub-model while high precision is still attained with the focusing operation. Moreover, when the focus of the discussion slowly changes, the language model only shifts slightly due to the overlap with the neighboring regions.

Finally, the submodels are stored by the index of the corresponding map unit.

**Operation of the model during speech recognition**

The text segment used for focusing the model is encoded as a document vector, which is then compared to the model vectors associated with each SOM unit to locate the best-matching map unit, i.e., the one with the most similar model vector. The corresponding local, topically focused language model is retrieved from an index. As the text segment one can utilize the recognition history or the candidate transcriptions in the n-best-list provided by the acoustic speech recognizer.

The local model may then be combined with an overall model to obtain probability estimates for words and sequences that did not appear in the local topic region. An illustration of the combined model is shown in Figure 1.

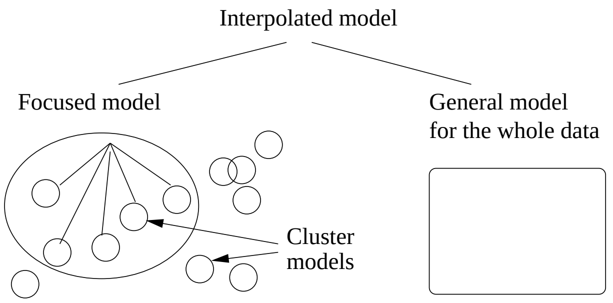


Figure 1: *A focusing language model obtained as an interpolation between a local topical model specific to a document map region and a general model.*

## EXPERIMENTS

### Baseline speech recognizer and baseline LM

The first speech recognition task where the focusing language model adaptation was evaluated was read text in Finnish. There were plenty of reasons for choosing this task, the primary ones being the difficulty of the language modeling, the relatively simple acoustic modeling involved, and the importance of the task as a part of a future benchmark database in Finnish LVCSR (large vocabulary continuous speech recognition). The training data for the acoustic models (AM data) consists of approximately 12 hours speech (one book) read by one person in acoustically rather stable and clean conditions. However, for language modeling this material is insufficient and it is not obvious how to choose (and obtain) a relevant and sufficiently large database.

For creation of document maps and language models we used a collection of various electronic books, journals, and newspapers from the CSC [4] and a collection of short news articles [1]. The corpus (LM data), consisting of totally about 40 M words and 1 M different word forms, contains stylistically very nonhomogeneous material.

We implemented a baseline speech recognizer for the task first using rather conventional LVCSR methods:

1. *Feature extraction:* Short time MFCCs and their $\Delta$ coefficients.

2. *Acoustic modeling (AM):* Context independent 3-state phoneme HMMs with emission probabilities modeled by 21 Gaussian mixtures per state. The mixtures were initialized and trained as in the baseline recognizer in [12] for unlimited vocabulary using SOM and segmental K-means algorithms. The training data was segmented iteratively using forced alignment by the speech recognizer.

3. *Lexical modeling:* A pronunciation dictionary was constructed using the general pronunciation rules of Finnish words. For the vocabulary we chose 64,000 most common word forms from the LM training material.

[1]The news articles were provided by the Finnish News Agency STT.

4. *Language modeling (LM):* Trigram models with basic back-offs and Good-Turing discounting by the CMU-Cambridge Toolkit [3] for the full lexicon.

In experiments we ran the Noway stack decoder [15] for automatically detected speech windows of approximately 10 sentences at a time from the evaluation data (the last unused 40 mins of the AM data). The baseline results (first column of Table 1) were produced in a single recognition pass, but the obtained word lattices were saved for further experiments.

**"Optimal" adaptation using parallel text data**

| LM Mode | Baseline 1-pass | Local 1-pass | Local 2-pass | Interpolated 2-pass | Local* 1-pass |
|---|---|---|---|---|---|
| OOV % | 21 | 25 | 25 | *16* | 0.3 |
| PP | 4190 | *1780* | *1780* | 2070 | 86 |
| WER % | 83 | 81 | 81 | *80* | 51 |
| CBER % | 69 | 68 | *67* | 68 | 45 |
| TER % | 77 | *71* | 73 | 72 | 52 |

TABLE 1: CONTINUOUS SPEECH RECOGNITION RESULTS FOR ADAPTED LMS. THE LOCAL LM ('LOCAL') WAS TRAINED USING ADAPTATION DATA (TRANSCRIPTIONS OF AM DATA) ONLY. THE LATTICE RESCORING EXPERIMENTS ('2-PASS') USE THE FIRST-PASS DECODING LATTICE PRODUCED BY THE BASELINE RECOGNIZER. 'INTEPOLATED' RESCORES THE LATTICE BY AN INTERPOLATED MODEL BETWEEN THE BASELINE AND THE LOCAL LM. 'LOCAL*' MODEL IS TRAINED BY THE ACTUAL TEST DATA AS WELL AS THE ADAPTATION DATA AND IS THUS NOT A REALISTIC LM, BUT MERELY JUST A THEORETICAL PLAY. THE EVALUATION MEASURES OOV, PP, WER, CBER, AND TER ARE EXPLAINED IN THE TEXT.

The text corresponding to the AM data (50 K words and 17 K different word forms) that was not included in the actual LM data, was used to construct a small trigram LM for adaptation purposes. Since the text is from the same author and from the same book as the held-out evaluation set, the style should match as closely as possible to the recognition task. Ideally the focusing LM should automatically produce something similar from the LM data, i.e. a LM concentrated on the most relevant material to the task. In Table 1 we give the results for the evaluation data decoded using this "optimal" adaptation LM (local LM) in a second recognition pass to rescore the lattices produced by the baseline recognizer. Other experiments include the local LM in a single-pass recognition and a rescoring experiment with an interpolated model of the baseline and the local LM.

**The document map**

The document map for the results reported here was constructed as follows: A corpus of 14 M words from [4] was segmented to obtain documents with an average length of 90 words, resulting in 152,845 documents.

The documents were organized on a document map of size 24x32 (768 map units) using the methods described in [11]. Words appearing fewer than 50 times were discarded, and the remaining words were weighted by tf.idf weighting. The dimensionality of the document vectors was reduced from 13807 to 500 by random projection of the document-by-word-matrix, and the map magnification and other speedup methods for the map organization were applied.

**Focusing LM experiments**

| | Focus by text | | | Focus by speech | | |
|---|---|---|---|---|---|---|
| | A - 1 | A - 2 | A - 3 | B - 1 | B - 2 | B - 3 |
| OOV % | 68 | *21* | 43 | 67 | *21* | 42 |
| PP | *415* | 4790 | 2360 | 590 | 4730 | 2450 |
| WER % | 103 | 103 | *96* | 103 | 103 | *96* |
| CBER % | 85 | 85 | *81* | 85 | 85 | *81* |
| TER % | 97 | 97 | *89* | 97 | 97 | 90 |

Table 2: Continuous speech recognition results for focused LMs of words in inflected forms. Focusing modes A and B; focused LM constructions 1, 2, and 3; and result evaluation measures OOV, PP, WER, CBER, and TER are explained in the text.

For the focusing LM we evaluated two different focusing modes (see Table 2):

A. Focusing based on available text data that is assumed to be relevant to the speech.

B. Focusing based on first-pass speech recognition results.

As relevant text data we used the same adaptation data (AM data) as in Table 1. In the setup B, the first-pass speech recognition result was used in each recognition window to select and focus on the most accurate LM for rescoring the lattice obtained for that ten-sentence window. Since the window size was relatively short, the previous windows were allowed to affect the focusing using gradual forgetting in which the weight of the latest window was 0.5.

Three different types of LMs were assigned to the map units:

1. A bigram LM for words appearing more than twice in the Voronoi region of a single map unit.
2. An interpolation of 1 and the monolithic baseline LM.
3. As 1, but including the Voronoi region of several neighboring map units. The neighborhood in these experiments was determined as a bubble of radius 2 in a hexagonal grid of units [10].

**Base form LM experiments**

As a way to both successfully compress the data for language modeling and to concentrate on base forms, we experimented LMs where the units were base forms instead of inflected word forms (see Table 3). In this model the inflections only appear as pronunciation variants of the base forms associated with their relative probabilities [16]. For the LM we chose 64,000 most common *base forms* extending the effective vocabulary by an order of magnitude to cover totally 650,000 different word forms. However, this huge lexicon turned out to be difficult to handle in the decoding, so as the effective lexicon in the first pass recognition (Table 3) we used only the base forms seen in the adaptation data (17,000) and their inflections (170,000).

| | Single pass | Focus by text A - 1 | A - 2 | A - 3 | Focus by speech B - 1 | B - 2 | B - 3 |
|---|---|---|---|---|---|---|---|
| Base OOV % | 5 | 47 | *4* | 20 | 40 | *4* | 18 |
| Base PP | 2600 | *560* | 2900 | 1660 | 600 | 2650 | 1650 |
| WER % | *94* | 103 | 106 | 99 | 103 | 106 | 99 |
| CBER % | 77 | 81 | 80 | 76 | 81 | 81 | *75* |
| TER % | 76 | 85 | 86 | *75* | 85 | 88 | *75* |

TABLE 3: CONTINUOUS SPEECH RECOGNITION RESULTS FOR FOCUSED LMS OF WORDS IN BASE FORMS. FOCUSING MODES A AND B; FOCUSED LM CONSTRUCTIONS 1, 2, AND 3; AND RESULT EVALUATION MEASURES OOV, PP, WER, CBER, AND TER ARE EXPLAINED IN THE TEXT. SINGLE PASS RECOGNITION RESULT WITH THE EXTENDED VERSION OF THE BASELINE LM IS USED TO PREPARE THE BASE FORM LATTICES. BASE OOV AND PP ARE OOV RATE AND PERPLEXITY FOR THE BASE FORM DATA.

**Evaluation measures**

Because the obtained WER is very high in the results, it is important to separate some of the word recognition errors that we believe to have significantly less effect in speech understanding for humans, machines, and both; e.g. with Finnish compound words also people frequently make spelling mistakes by writing a combination of two words as a single compound word, or by writing an actual compound word as two separate words. Other such errors in Finnish are the word inflections which may have hundreds of easily confusable forms that often have only a subtle effect for the main idea of the sentence. Some of the endings may also be left out in speech. In IR, then, the words are normally returned to base forms (stemming) and the analysis of the content of a paragraph often excludes the common function words (stop words) and even disregards the actual order of appearance of the words.

As the evaluation measures we choose the following ones: OOV, PP, WER, CBER, and TER. OOV is the out-of-vocabulary rate in the text corresponding to the speech test data and PP its perplexity [2]. WER (word error

[2]Perplexity is the inverse predictive probability for all the words in test set.

rate) is the number of errors (substitutions, deletions and insertions) between the recognized and the correct text sequences, divided by the total number of words in the correct text. BER (base form error rate) is computed as WER, too, except after returning all words into base forms. Instead of just BER we used CBER where additionally all the compound words are further split into single words. The term error rate (TER) takes a few steps further towards extracting the content of the speech by neglecting some very frequent words (stoplist) and the exact ordering of the recognized words in the paragraphs. TER is most relevant for speech used in information retrieval tasks, such as discourse topic identification and speech indexing. In [9] TER is defined as the difference of two word histograms (the recognition result $H$ and the correct transcription $R$) after stopping and stemming both word sequences (summation $t$ is over all resulting terms):

$$\mathrm{TER} = \sum_t |R(t) - H(t)| / \sum_t R(t) * 100\% \tag{1}$$

Because TER is most meaningful in relatively short paragraphs, it is here computed as an average over all recognition windows, whereas for the other measures the total number of errors is divided by the total number of words.

## DISCUSSION

There are probably three general remarks from the results that are striking for people familiar with LVCSR applications: High out-of-vocabulary rate, high perplexity, and high word error rate. The first two show directly how difficult the language modeling task is. Even the monolithic baseline LM leaves out 20 % of the word forms in the evaluation data and its extended version still 6 % (5 % of the base forms). This indicates not only that the LM training data does not match particularly well to the speech data used for evaluation, but also that the vocabulary growth in Finnish is a particularly tough problem for LMs. For another large vocabulary task in Finnish, where the match between the LM evaluation and training data is good [16], we have measured an OOV rate 13 % for a trigram LM that corresponded to the baseline model here. And as seen also from the results here, the increase of vocabulary usually increases perplexity, which roughly means that there are more rival candidates from which to choose the right one.

High word error rate indicates several things for the system and the task. The difficulties in the baseline LM lead to suboptimal lattices and poor first-pass recognition results. The success in focusing to a relevant specialized LM is affected by the first-pass results, and the success in using that model for lattice rescoring is further affected also by the lattice quality. Table 1 shows that rescoring the lattice by the local model, which represents an optimal focusing result, does improve the decoding results slightly. Since the lattice rescoring by a small LM is a very fast operation compared to the first recognition pass by a large LM, this corresponds to a fast light-weight way of LM adaptation for speech recognition. The results are also close to both those obtained by rescoring with the interpolated LM and by the single pass recognition with the local LM.

In addition to the need for improvements in LMs and LM training data, a further remark about WER is that it can be a slightly misleading recognition performance measure for languages like Finnish where words constitute of several morphemes. Because the individual recognition errors can be anything between completely wrong words and the right words in slightly different inflected forms, we suggest the compound base form error rate (CBER) and term error rate (TER) as potentially more realistic and useful error measures. Whereas WER actually increased in lattice rescoring by focusing LM due to the reported problems with the LM data, the utilization of the bubble neighborhood for focusing LMs could still slightly decrease CBER and WER in the experiments with base form LMs.

The suggested directions to improve the recognition are to gather more relevant LM data regarding the current task and to improve LM accuracy by suitably splitting the words into morphemes. In acoustic models there are probably less room for improvements, except maybe applying context-dependent phoneme models and manual inspection of the automatic segmentation of the training data.

In this work we did not yet perform comparison experiments using other recent developments in large vocabulary language modeling (for a good review of those, see [6]). In [13] the focusing LMs compared favorably against a few of its closest rivals in average perplexity measurements on two text databases (one in Finnish and one in English) and the future work is to extend these experiments on the speech recognition task, as well.

## CONCLUSIONS

In this paper we presented experiments on using focusing language models in large vocabulary continuous speech recognition. The focus can be based either on relevant text data, if available, or on first-pass decoding results in multi-pass speech recognition. By utilizing a collection of LMs spanned by an organized document mapping, the focusing can be carried out very efficiently. The second recognition pass including focusing and lattice rescoring process is an order of magnitude faster than the first pass. For evaluation of continuous speech recognition results we proposed the use of more sophisticated performance measures than word error rate for languages like Finnish where some of the misrecognitions are not fatal for speech understanding and related applications. The speech recognition results do not indicate clear improvements over the baseline, but by using a local LM of small but relevant text material, we see, however, that lattice rescoring can decrease the error rates despite the higher OOV rate of the LM. This indicates that lattice rescoring by an automatically focusing LM is likely to improve speech recognition given that relevant text material is represented in the LM training data.

## REFERENCES


[1] J. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling," **Proc. of the IEEE**, vol. 88, no. 8, pp. 1279–1296, 2000.

[2] P. Clarkson and A. Robinson, "Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache," in **Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing**, 1997, pp. 799–802.

[3] P. Clarkson and R. Rosenfeld, "Statistical language Modeling Using CMU-Cambridge Toolkit," in **Proc. of 5th European Conference on Speech Communication and Technology**, Rhodes, Greece, 1997, pp. 2707–2710.

[4] CSC—Tieteellinen laskenta Oy. Finnish Language Text Bank: Corpora Books, Newspapers, Magazines and Other. http://www.csc.fi/ kielipankki/, 2001.

[5] D. Gildea and T. Hofmann, "Topic-based Language Modeling Using EM," in **Proc. of 6th European Conference on Speech Communication and Technology**, Budapest, Hungary, 1999, pp. 2167–2170.

[6] J. T. Goodman, "A bit of Progress in Language Modeling," **Computer Speech and Language**, vol. 15, no. 1, pp. 403–434, 2001.

[7] T. Honkela, S. Kaski, K. Lagus and T. Kohonen, "Newsgroup exploration with WEBSOM method and browsing interface," TR A32, **Helsinki University of Technology, Laboratory of Computer and Information Science**, Espoo, Finland, 1996.

[8] R. Iyer and M. Ostendorf, "Modelling Long Distance Dependencies in Language: Topic Mixtures versus Dynamic Cache Model," **IEEE Trans. Speech and Audio Processing**, vol. 7, 1999.

[9] S. Johnson, P. Jourlin, G. Moore, K. S. Jones and P. Woodland, "The Cambridge University Spoken Document Retrieval System," in **Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing**, 1999, pp. 49–52.

[10] T. Kohonen, **Self-Organizing Maps**, Berlin: Springer, 2001, 3rd ed.

[11] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, V. Paatero and A. Saarela, "Organization of a Massive Document Collection," **IEEE Trans. on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery**, vol. 11, no. 3, pp. 574–585, May 2000.

[12] M. Kurimo, **Using Self-Organizing Maps and Learning Vector Quantization for Mixture Density Hidden Markov Models**, Ph.D. thesis, Helsinki University of Technology, Espoo, Finland, 1997.

[13] M. Kurimo and K. Lagus, "An Efficiently Focusing Large Vocabulary Language Model," in **International Conference on Artificial Neural Networks, ICANN'02**, 2002.

[14] R. Lau, R. Rosenfeld and S. Roukos, "Trigger-based Language Models: A Maximum Entropy Approach," in **Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing**, 1993, pp. 45–48.

[15] S. Renals and M. Hochberg, "Start-synchronous search for large vocabulary continuous speech recognition," **IEEE Trans. Speech and Audio Processing**, vol. 7, pp. 542–553, 1999.

[16] V. Siivola, M. Kurimo and K. Lagus, "Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition," in **Proc. of 7th European Conference on Speech Communication and Technology**, Aalborg, Copenhagen, 2001.