Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition in Finnish

Vesa Siivola, Mikko Kurimo, Krista Lagus

Neural Networks Research Centre
Helsinki University of Technology
Vesa.Siivola@hut.fi, Mikko.Kurimo@hut.fi, Krista.Lagus@hut.fi

Abstract

Statistical language modeling (SLM) is an essential part in any large-vocabulary continuous speech recognition (LVCSR) system. The development of the standard SLM methods has been strongly affected by the goals of LVCSR in English. The structure of Finnish is substantially different from English, so if the standard SLMs are directly applied, the success is by no means granted. In this paper we describe our first attempts of building a LVCSR for Finnish and the new SLMs that we have tried. One of our objective has been the indexing and recognition of broadcast news, so special issues of our interest are topic detection, word stemming and modeling words that are poorly covered in the training data. Our new methods are based on neural computing using the self-organizing map (SOM) which has recently been shown to successfully extract and approximate latent semantic structures from massive text collections.

1. Introduction

Speech recognition technology is rapidly advancing into practical applications. However, this development is still restricted in very few, although widely spread, languages. One of the main restrictive factors is the lack of data resources such as electronic texts, speech recordings, transcriptions, pronunciation lexicons, and morphological analyzators. Another important factor is that many other languages have some rather fundamental differences [20] compared to English, in which the main development work is concentrated. In Finnish, for example, the word order in sentences is much more flexible (in fact, discourse-conditioned [19]). Another fundamental difference is the huge variety of inflected forms for most of the words, so that the use of out-of-vocabulary words (OOVs) and other rare words complicate significantly the statistical language modeling (SLM). For example, a Finnish noun can have more than 2,000 different inflectional forms and in our news wire corpus we found 13 % OOV rate for a normal 60000-word vocabulary whereas the corresponding rate for English is less than 1% and for Serbo-Croatian 8 % [20]. All these properties are likely to have the effect that N-gram models are not very successful.

The history of speech recognition in Finnish is rather long. For example, more than two decades ago in Helsinki University of Technology (HUT) there was already a system that performed speaker dependent automatic speech recognition (ASR) for a closed thousand-word vocabulary in a 94 % word accuracy [5]. Until very recently, however, the large-vocabulary ASR in Finnish has been concentrated in vocabulary independent phonemic ASR. The main reasons for this rather different approach than for English were the inflected forms that made the lexicon building very difficult and the fact that word pronunciations can be adequately approximated by automatic rules. Among the best Finnish open-vocabulary ASR has been the systems at HUT [7, 18, 12], which have reported up to 5 % speaker dependent phoneme error rates. All these systems, however, suffered from the small size of the available speech resources, so that changes in the ASR task like speech conditions, speakers and vocabulary, deteriorated the performance easily. Today, for small vocabulary tasks there already exist commercial applications, for example, to control mobile phones. Very recently the interest in large-vocabulary applications has again increased along with the marketing of the first commercial Finnish system (by Philips).

One of our research objectives is the indexing and retrieval of Finnish broadcast news and developing large-vocabulary continuous speech recognition (LVCSR) methods that will be needed to successfully decode these spoken news. In broadcast news the vocabulary typically changes in time, so that there will be important words for which there are not enough training data. This problem is not only in the acoustic models and pronunciations, but in SLM, as well. Efficient methods are needed that can cluster the words and use the clusters to derive properties for the words that are poorly represented in the training data. Broadcast news are also characterized by quick and sharp turns between distinct topics when the active vocabulary and the frequency of some typical language structures can rapidly change. Usually the news topics are covered in other recent or contemporary data, but the problem is the topic classification and the detection of topic changes. In audio indexing the index terms usually are base forms or stems of the words, not the words itself. For Finnish this difference is quite significant, because the stemming is a non-trivial task, so we integrated it to the SLM, as well.

The language models introduced in this paper try to model the language with a limited amount of statistical information of the preceding words. The most commonly used language model, n-gram, belongs to this group of models. The new models we introduced can use a much larger vocabulary by compressing the language model, so that it will deal with groups of words instead of single words. To cope with the word inflections, a model based on grouping the words by their base form is introduced in section 2.3. A model, where words are grouped according to their statistical context, is presented in section 2.4. It is based on the assumption that it is not necessary to know the preceding words exactly, but the knowledge of their category would be sufficient for SLM [6, 17]. In the last model (2.5) the semantic classification of the document context is utilized to find a SLM for the topic that the document belongs to. The semantic classification is made by WEBSOM algorithm [10] that applies the self-organizing map (SOM) [8] paradigm to unsupervised semantic clustering document collections.

In this paper we present SLM experiments using a Finnish news wire corpus. The recording and preparation of a Finnish broadcast news corpus is still under development, but for training the SLMs for it this news wire material will probably still continue to be the best that is available. For LVCSR we have a novel Finnish corpus that consists of read speech that is not related to news, but has a lot of rare foreign names and a large vocabulary. We have tested our SLMs for this data as well to see how they can be exported to a different LVCSR task.

2. Statistical language models

2.1. Experiments

The data used for the SLM experiments was taken from a Finnish news wire corpus. The training set consisted of 63k articles (12M words) coming from the 8 largest news groups of the corpus. The test data consisted of 400 new articles (77k words) coming from the 4 largest news groups. All the data is from years 1999 to 2001.

The data was preprocessed by tagging the punctuation marks and converting all the numbers into one tag. For some experiments, the inflected forms were returned to base forms with TWOL, a morphological analyzer for Finnish by Lingsoft [11].

We used the perplexity to measure the performance of the language model relative to the test data. Perplexity P_p of the text $W = \{w_1, w_2, \ldots, w_L\}$ for model M is $P_p = P(W|M)^{-\frac{1}{L}}$. For trigrams, for example, it is

$$P_p = \left(\left[\prod_{i=3}^{L} P(w_i|w_{i-1}, w_{i-2}) \right] P(w_2|w_1)P(w_1) \right)^{-\frac{1}{L}} \quad (1)$$

2.2. N-gram model

This is the baseline model. The probability of word w_k is estimated based on n previous words $w_{k-1}, \ldots, w_{k-n+1}$, that is $P(w_k|w_{k-1}, w_{k-2}, \ldots, w_1) = P(w_k|w_{k-1}, \ldots, w_{k-n+1})$. Since the data is sparse, some kind of smoothing and/or backing off to lower order n-grams is usually needed.

The results prefixed with CMU are computed using the CMU/Cambridge SLM toolkit with Good-Turing smoothing and back-offs. Other results in Table 1 are calculated with no smoothing and if the word sequence does not match the given n-gram, it is considered as an out of language model (OOLM) sequence. Note, that the models without backoffs and smoothing do not cover the data space very well, thus low perplexity does not necessarily mean a good model.

Table 1: *The perplexity results with standard n-grams.*

vocab	1gram/OOLM	2gram/OOLM	3gram/OOLM
CMU 20k	1733 / 21%	372 / 21%	285 / 21%
CMU 60k	2867 / 13%	582 / 13%	460 / 13%
20k	1733 / 21%	97 / 51%	13 / 79%
60k	2867 / 13%	96 / 44%	11 / 75%

2.3. Base form n-gram model with inflections

First we considered n-gram models where we replaced all the inflected form by the corresponding base forms. Because this significantly reduces the scope of the problem these results are not directly comparable to conventional perplexities.

To create a more general and comparable model, we grouped all the possible inflections of each base form into one group, i.e. the inflections $w_{b1}, \ldots, w_{bN}$ belong to group G_b. If we assume, that the relevant statistical information of the sentence is contained in the knowledge of the base forms, it can be shown that the probability of word w_k given the history $\{w_{k-1}, \ldots, w_1\}$ is given by

$$P(w_k|w_{k-1}, \ldots, w_1) = P(G_{w_k}|G_{w_{k-1}}, \ldots, G_{w_1})P(w_k|G_{w_k}) \quad (2)$$

where $P(w_k|G_{w_k})$ is an estimate for the relative frequency of inflection w_k with respect to all other inflections of the same base form.

The results for base form n-grams with inflections are in Table 2. The perplexity measure P_p for the group trigram model is

$$P_p = \left(\left[\prod_{i=3}^{L} P(G_{w_i}|G_{w_{i-1}}, G_{w_{i-2}})P(w_i|G_{w_i}) \right] P(w_2|w_1)P(w_1) \right)^{-\frac{1}{L}} \quad (3)$$

Table 2: *Base form n-grams with inflections.*

base forms	vocab	2gram/OOLM	3gram/OOLM
20k	219k	443 / 44%	40 / 77%
60k	358k	424 / 42%	38 / 75%

2.4. Context clustered n-grams

We can extend the idea of grouping (section 2.3) to unsupervised clustering of all words based on the statistical properties of the context they appear. If C_k is the cluster of words of similar context, we can derive as in (2)

$$P(w_k|w_{k-1}, \ldots, w_1) = P(C_{w_k}|C_{w_{k-1}}, \ldots, C_{w_1})P(w_k|C_{w_k}) \quad (4)$$

The advantage of this is that we can reduce the size of our language model significantly or model longer contexts (larger n). In addition, the problem of data sparseness can be reduced. Naturally, this method is not limited to Finnish [17].

For massive text data collections the word clustering can be conveniently performed as follows:

1) To reduce the dimensionality of the problem, generate a low dimensional random vector r_i for each word w_i, that is $r_i \in \mathcal{R}^l$, where $l << N$. N is the size of the vocabulary.

2) Collect and average the n-grams around every occurrence of w_i. For example, for trigrams find the average of random vectors for words in each position of $W = \{w_k, w_i, w_j\}$ and let the context vector $c_i = [\bar{r}_k \bar{r}_j]$ for the word w_i [15].

3) Find the clusters in the context vector space. SOM-algorithm is used here, but other fast clustering algorithms could probably do as well.

4) Go through the text corpus again. For each word find $P(w_i|c_i)$, the probability being generated from cluster c_i. This is estimated by the proportion of the frequency of w_i out of the frequency of the whole cluster c_i.

2.5. Using WEBSOM document maps for topic models

A new way to generate topic based SLMs was developed by using the WEBSOM [10] for the generation of the topics and for mapping documents to them. The WEBSOM is a neural computation method based on SOM that extracts latent semantic structures from a text collection and visualizes the collection in an ordered low dimensional display. In brief, the method

Table 3: *This test was conducted by clustering the base forms of the words and applying the inflection grouping (section 2.3.). The size of the language model is greatly diminished, but the perplexity goes up.*

clusters	base forms	vocab	2gram/OOLM	3gram/OOLM
300	20k	219k	3454 / 29%	3608 / 43%
300	60k	358k	4676 / 23%	3696 / 35%
800	20k	219k	3659 / 29%	3932 / 43%
800	60k	358k	4713 / 23%	4062 / 35%

consists of mapping words into a (random) vector space, then building document vectors using the histograms of these words weighted by inverse document frequency (idf) or entropy across document classes , and finally using SOM to construct the topological mapping of the whole document space. Previously it has been shown that the method obtains a topical ordering of the documents, which can be utilized for text exploration [10] and retrieval [13]. The ordered map can thus be considered to represent the latent topical aspects of the texts.

The purpose of using WEBSOM for SLM is to separate a subset of training data, where the documents are semantically similar to the current test document. By semantic similarity we mean here that the documents lie close together in the semantic document space spanned by the whole training data. In news wire material this normally means that those documents will concern similar news topics [10]. A separate language model is then formed for each subset ("a topic"). The statistical accuracy of a model restricted to a subset of the training data is probably quite limited, but we hope that it is more accurate for the current active vocabulary and sentence structures than an overall model.

Generating topic based SLMs by partitioning the training data and defining the topics through the differences in weighted word histograms has also been proposed in several other works, e.g. [4, 1]. The novelty of our approach is in the unsupervised way in which the WEBSOM creates a semantically organized mapping of the documents space of massive text collections, in which fast approximative methods can be used to determine the best-matching map areas [10]. Furthermore, utilization of the ordering of the map allows for efficient smoothing between models, which is a key issue in modeling with partitioned data.

Table 4: *The perplexity measurements for SLM with WEBSOM document maps, reported for a separate test document set of 400 documents. Results are shown for five different models (numbers 1–5) and interpolated versions of some of them. Models 1 and 2 are regular tri- and bigram models over the whole training data. Model 3 utilizes WEBSOM with each topic consisting of documents in the 10 closest map units. In model 4 the three closest map units were utilized. For comparison, model 5 is based on a prior manual document classification.*

	Model	Average PP
1	General trigram	607
2	General bigram	571
3	Unsupervised topics; 10-topic bigram	334
4	Unsupervised topics; 3-topic bigram	224
	Interpolated 2 & 3	275
	Interpolated 2 & 4	181
	Interpolated 1 & 3	255
	Interpolated 1 & 4	**168**
5	Supervised topics; bigram	324
	Interpolated 5 & 2	328
	Interpolated 5 & 3	222

For studying the document map approach we created a document map of 200 units for the training material (section 2.1.). For each test document, the model was formed based on documents in N best-matching map units, of which we tried N=10 and N=3. For comparison, a "supervised topical model" was created based on a prior categorization that uses the news group label embedded into the training and testing documents.

As in [4, 1], the topic based SLMs can be further improved by interpolating them with the overall SLMs using interpolation co-efficients determined by held-out data sets (for simplicity, we just applied equal interpolation weights). The interpolated model benefits from the overall model with words that are poorly represented in the smaller data set and, conversely, from the topic model with words that are characteristic of the topic.

It can be seen in Table 4 that the SLM models based on unsupervised topic generation and detection (models 3 and 4) perform better on the test documents than the corresponding overall SLM (model 2). Moreover, the results are comparable to those obtained with the supervised topical categorization (model 5). Because these perplexities were measured using the base forms, these results are not directly comparable to Tables 1–3. The vocabulary in the general n-grams (models 1 and 2) was 40k words, but in the topic models (3–5) the vocabulary varied depending on the topic.

3. LVCSR system

The audio data for LVCSR consisted of a book read by one reader. The first 5 hour were separated for training and the remaining 30 minutes were used for testing only. No transcription of the training data was available, only a rough script of the book. Some manual corrections were made for the script, but otherwise the accurate transcription was created in a fully automatic manner as follows:

First the script was transformed closer to the spoken format by writing the numbers and some common abbreviations like they are normally uttered. Then a rough phonetic transcription was obtained using a very simple version of common Finnish pronunciation rules. The first rough alignment between the speech and the transcription was carried out using forced Viterbi alignment with a simple speech recognizer trained on an earlier isolated word speech database [16]. This database consisted of 21000 isolated words from news text uttered by 60 different speakers [16]. To make the alignment of such long speech file as the book possible we divided it into 30s long overlapping windows. To avoid border effects and to align the text correctly, a new window was started at the center point of the previous one. Finally, this initial segmentation was used to retrain the phoneme models and to create the final forced alignment which was then used for training the acoustic models of the LVCSR system.

The front end of the recognizer used MFCCs [18] and their Δ's with cepstral means subtraction and codebook adaptation [16]. The emission probabilities for the HMM of each phoneme were estimated by Gaussian mixtures, initialized by SOM [12].

For the lexical model, the base form model with inflections listed as alternative pronunciations was used as explained in section 2.3. The pronunciations themselves were approximated by automatic rules. The start-synchronous "Noway" decoder [14] is used for computing the final most probable word hypothesis that connects the acoustic models, lexicon and language models.

It is clear that our SLMs for broadcast news are not optimal for the current LVCSR task, but we still wanted to test how they can be exported to this different task. It was expected as well, that the training part of the book would be too short (23k words) alone to train generalizable SLMs. The results in Table 5 confirms and shows that the combination "comb" of "book" and "news" models was better than the other 2 models alone. Note the high OOLM rate in the "book" models due to the small vocabulary.

Table 5: *Test with an excerpt from a book. In column t, n stands for n-gram model, i stands for inflection n-gram model.*

model	t	base forms	vocab	2gram/OOLM
book	n	-	5k	19 / 84%
book	i	5k	9k	67 / 83%
news	n	-	60k	249 / 62%
news	i	60k	358k	1374 / 56%
comb	n	-	60k	234 / 60%
comb	i	60k	360k	1206 / 54%

4. Discussion

The system presented here is still at very preliminary stages. It is the first academic LVCSR project of this scale, that has been made for Finnish. Databases are still under construction and the models presented are the first takes on how to handle this kind of problem.

Better smoothing and back-off methods have to be implemented to make the models presented comparable to previous works and even to each other. Since the handling of out of vocabulary sequences is not very graceful, the reader is advised to pay attention to the coverage of the language model as well.

The best results were obtained by the topic SLMs based on WEBSOM. Actually, it performed even better than the manual topic classification. One reason could be that it allowed to specify more accurate topics, because there were only 4 manually tagged groups of news in the test material. It should be noted that in this experiment we estimated the topic for a document based on the same document. In LVCSR the topics can only be specified based on the hypothesis of the current word sequence, so the topic classification will be more prune to errors.

Furthermore, additional experiments are necessary to see the effects of the new SLMs on the actual speech recognition accuracy. It is a well-known fact that all the perplexity improvements are not necessarily reflected in the LVCSR error rate (see e.g. [3]).

5. Conclusions

In this paper we described our new SLMs for LVCSR. The new developments aimed at enhancing the conventional SLMs to better suit for broadcast news data in Finnish. The main ideas were related to statistical modeling of word inflections, word contexts and document contexts. Although we think that some our of perplexity results are very good, additional experiments and speech data are still needed to see the effects on the actual LVCSR task.

6. Acknowledgements

We thank the Finnish News Agency STT for the news wire material and Näkövammaisten Keskusliitto NKL for the speech material and all the people who helped us in the preprocessing of the text materials. We wish to thank also the WEBSOM project for letting us use their software and Steve Renals for letting us use Noway decoder [14]. This research was supported by USIX research program of Technology Development Centre of Finland TEKES.

7. References

[1] P. Clarkson and A. Robinson. Language model adaptation using mixtures and an exponentially decaying cache. In ICASSP-97, pages 799–802, 1997.

[2] P. Clarkson and R. Rosenfeld. Statistical language modeling using CMU-Cambridge toolkit. In EUROSPEECH-97, pages 2707–2710, 1997.

[3] P. Clarkson and A. Robinson. Improved language modelling through better language model evaluation measures. *Computer Speech and Language*, 15(1):39–53, 2001.

[4] R. Iyer and M. Ostendorf.. Modeling long range dependencies in language. In ICSLP-96, pages 236–239, 1996.

[5] M. Jalanko. *Studies of Learning Projective Methods in Automatic Speech Recognition*. PhD thesis, Helsinki University of Technology, Espoo, Finland, 1980.

[6] R. Kneser and H. Ney. Improved clustering techniques for class-based statistical language modelling. In EUROSPEECH-93, pages 973–976, 1993.

[7] T. Kohonen. The 'neural' phonetic typewriter. *Computer*, 21(3):11–22, 1988.

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

[9] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen. SOM_PAK: the self-organizing map programming package. Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996. Program package available via WWW at URL http://nucleus.hut.fi/nnrc/som_pak.

[10] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, V. Paatero, and A. Saarela. Organization of a massive document collection. *IEEE Trans. Neural Networks*, 11(3):574–585, 2000.

[11] K. Koskenniemi. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD thesis, University of Helsinki, Finland, 1983.

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

[13] K. Lagus. *Text Mining with the WEBSOM*. PhD thesis, Helsinki University of Technology, Espoo, Finland, 2000.

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

[15] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biol. Cyb.*, 61(4):241–254, 1989.

[16] V. Siivola. An adaptive method to achieve speaker independence in a speech recognition system. Master's thesis, Helsinki University of Technology, Espoo, Finland, 1999.

[17] V. Siivola. Language modeling based on neural clustering of words. IDIAP-Com 02, Martigny, Switzerland, 2000.

[18] K. Torkkola, J. Kangas, P. Utela, S. Kaski, M. Kokkonen, M. Kurimo, and T. Kohonen. Status report of the Finnish phonetic typewriter project. In *Artificial Neural Networks*, pages 771–776, North-Holland, Amsterdam, 1991.

[19] M. Vilkuna. *Free Word Order in Finnish. Its Syntax and discourse functions*. SKS, Helsinki, Finland, 1989.

[20] A. Waibel, P. Geutner, L.M. Tomokiyo, T. Schultz, and M. Woszczyna. Multilinguality in speech and spoken language systems. *Proc. IEEE*, 88(8):1297–1313, 2000.