LEMMATIZED LATENT SEMANTIC MODEL FOR LANGUAGE MODEL ADAPTATION OF HIGHLY INFLECTED LANGUAGES

Tanel Alumäe, Toomas Kirt
Institute of Cybernetics at Tallinn University of Technology, Estonia

Abstract

We present a method to adapt statistical N-gram models for large vocabulary continuous speech recognition of highly inflected languages. The method combines morphological analysis, latent semantic analysis (LSA) and fast marginal adaptation for building topic-adapted trigram models, based on a background language model and very short adaptation texts. We compare words, lemmas and morphemes as basic units for language model adaptation. Experiments on a set of Estonian test texts and broadcast news speech data show that lemmas and morphemes give better performance than words in all cases. In speech recognition experiments, morpheme-based adaptation is found to perform significantly better than lemma-based adaptation.

Keywords: speech recognition, language model adaptation, morphological analysis, LSA

1. Introduction

Language model (LM) adaptation is a task of building a LM for large vocabulary continuous speech recognition (LVCSR) that is better suited for the given domain than a general background model, given a small adaptation corpus. In recent years, *latent semantic analysis* (LSA) has been successfully used for integrating long-term semantic dependencies into statistical LMs (Bellegarda 1998). The LSA-based approach gradually adapts the background LM based on the recognized words by boosting the unigram probabilities of semantically related words, using co-occurrence analysis of words and documents.

However, this approach cannot be efficiently directly used for highly inflective and/or agglutinative languages, such as Estonian, Finnish, Turkish, Korean and many others. In such languages, each word-phrase can occur in a large number of inflected forms, depending on its syntactic and semantic role in the sentence. In addition, many such languages are also so-called compounding languages, i.e., compound words can be formed from shorter particles to express complex concepts as single words. The compound words can again occur in different inflections. As a result, the lexical variety of such languages is very high and it is not possible to achieve a good vocabulary coverage when using words as basic units for language modeling. In order to increase coverage, subword units, such as morphemes, are used as basic units in language modeling. This approach has also been used for Estonian LVCSR (Alumäe 2004).

When we look at the morpheme decomposition of the word as an unordered *bag of morphemes*, it gives us much less information about the semantics of the phrase than the original inflected compound word. On the other hand, the high variety of different word inflections and the resulting sparseness of word-document occurrences makes the performance of word-based term-document analysis questionable. One way to decrease the lexical variety of the language would be to lemmatize all words, i.e., to use a morphological analyzer to find canonical forms or *lemmas* for all words before mapping them into the LSA space. Lemmatization increases the coverage of the LSA model vocabulary, and should act as an additional smoothing measure, since all different inflections of the same words are mapped to the same vocabulary item.

Based on these observations, this paper investigates a LSA-based LM adaptation approach for highly inflected languages. We compare three different sets of units – words, lemmas and morphemes – as basic units for modeling document similarities using the LSA technique. Performance of different basic unit sets is measured on Estonian language modelling and speech recognition experiments.

Many similar LSA-based adaptation methods have been investigated recently. A somewhat similar method was proposed by Klakow (2006) to select a subset of training corpus for fast marginal adaptation, however, training set perplexity minimization was used as a measure for selecting the documents from the corpus. Another related work (Turunen and Kurimo 2006) uses morpheme-like units and lemmas for LSI-based Finnish spoken document retrieval.

The presented approach differs from previous ones in dividing the adaptation process into three steps: LSA-based similar document retrieval, topic-specific unigram estimation from the retrieved documents and adaptation of the background LM using the estimated topic unigram. This enables us to use a different vocabulary for document similarity modelling and retrieval than is used for language modelling.

2. Latent Semantic Analysis

LSA is a mathematical technique for extracting and representing the semantic similarity of words and documents by analysis of large document corpora. The task of LSA is to define a mapping between the vocabulary $\mathcal{V}$ of M words, the document set $\mathcal{T}$, comprising N articles, and a vector space so that each word in $\mathcal{V}$ and each document in $\mathcal{T}$ is represented by a vector in this space. This is done by first constructing a word-document matrix W, where each element W_{ij} is a weighted count of word w_i in document d_j. The weighted count expresses both the word's importance in the particular document as well as the degree to which the word carries information in the domain of discourse in general. A suitable expression for W_{ij} as proposed by Bellegarda (1998) is

$$W_{ij} = G_i L_{ij} \tag{1}$$

where G_i is the global weight, indicating the overall importance of w_i in the corpus, and L_{ij} is a local value, describing the normalized frequency of w_i in d_j.

The global weight G_i is calculated via normalized term entropy E_i as $G_i = 1 - E_i$. Term entropy reflects the indexing value of the word w_i and can be calculated as

$$E_i = -\frac{1}{log(N)} \sum_{j=1}^{N} \frac{c_{ij}}{t_j} \log \frac{c_{ij}}{t_j} \tag{2}$$

where c_{ij} is the number of times w_i occurs in d_j, $t_i = \sum_j c_{ij}$ is the total number of times w_i occurs in $\mathcal{T}$. Thus, words distributed across many documents in the corpus receive a high term entropy value, while words present in relatively few specific documents receive a low entropy value.

The local value L_{ij} is a transformed version of c_{ij}, normalized for document length and dampened by applying a logarithm in order to reduce the effect of large differences in document length:

$$L_{ij} = log_2(1 + \frac{c_{ij}}{n_j}) \tag{3}$$

where n_j is the length of document d_j.

Next, LSA applies rank-R singular value decomposition (SVD) to the word-document matrix W:

$$W \approx \hat{W} = USV^T \tag{4}$$

where U is the $(M \times R)$ matrix of left singular vectors u_i, S is the diagonal matrix of R singular values and V is the $(N \times R)$ matrix of right singular vectors v_j. Matrix $\hat{W}$ is the best rank-R approximation to the original W. Rank R is the order of decomposition, $R \ll M(\ll N)$. The vectors u_i represent the word w_i in the new LSA space and the vectors v_j represent the document d_j in the same space.

The main benefit of SVD for our work is that it eliminates the sparseness issue, by reducing the dimensions of word and document vectors which isolates the most characteristic components of W and ignores the higher order effects that are unreliable and can be considered noise. This means that two words that do not necessarily co-occur in the original space $\mathcal{T}$ could still be close in the LSA space if they consistently tend to co-occur with a common set of words.

3. Language model adaptation

To find the closest documents to the given adaptation data in the LSA space, we first convert the adaptation data to pseudo-document representation $\tilde{d_p}$ by using the weighted counts (1) with $j = p$. Then, the representation of the adaptation data in the LSA space can be given as

$$\tilde{v}_p = \tilde{d}_p^T US^{-1} \tag{5}$$

Next, we calculate the "distance" between pseudo-document representation $\tilde{v}_p$ and every training document representation v_i by finding the the angle between $\tilde{v}_p S$ and $v_i S$:

$$K(\tilde{v}_p, v_i) = \angle(\tilde{v}_p S, v_i S) = \arccos \frac{\tilde{v}_p S^2 v_i^T}{\|\tilde{v}_p S\| \, \|v_j S\|} \tag{6}$$

In this way, the training documents are ranked by their distance measure and the top L documents can be selected for use as adaptation data.

Before constructing the unigram models, we apply count weighting, depending on the closeness of the training document to the adaptation data. Given a set to document identities found in the previous document selection step $\mathcal{T}_{Adap}$, we calculate the total weighted count $c_{Adap}(i)$ of the ith word as follows:

$$c_{Adapt}(i) = \sum_{j \in \mathcal{T}_{Adap}} \left(1 - \frac{K(\tilde{v}_p, v_j)}{\pi}\right) * c_{ij} \tag{7}$$

Given a domain-specific unigram model, we apply Fast Marginal Adaptation (Kneser et al. 1997) to quickly adapt a given LM to in-domain text characteristics. It uses the trigram trained on the background corpus as the initial LM. The background model is adapted so that its marginal is the unigram trained on the adaptation data. It turns out that this can be reformulated as a scaling of the background LM:

$$P_{Adap}(w|h) = \frac{\alpha(w) P_{BG}(w|h)}{Z(h)} \tag{8}$$

where $P_{Adap}(w|h)$ is the adapted word probability, given the history h, $P_{BG}(w|h)$ the word probability according to the background model and $Z(h)$ a normalization factor that guarantees that the probability sums to unity. The scaling factor $\alpha(w)$ is usually approximated as follows:

$$\alpha(w) \approx \left(\frac{P_{Adap}(w)}{P_{BG}(w)} \right)^{\beta} \tag{9}$$

where $P_{Adap}(w)$ is the unigram probability of w based on in-domain corpus, $P_{BG}(w)$ the background unigram probability and β a tuning factor between 0 and 1. The task of $\alpha(w)$ is to scale certain words up or down, depending on their relative frequency in the adaptation corpus with respect to the background corpus. The normalization factor $Z(h)$ can be efficiently calculated using the approximated scaling factor:

$$Z(h) = \sum_{w} \alpha(w) P_{BG}(w|h) \tag{10}$$

4. Experimental evaluation

4.1. Language modelling

For training the background LM, we used a subset of the Mixed Corpus of Estonian (Kaalep and Muischnek 2005) (70M words). In addition, we compiled a corpus of online newspaper articles from a website of a daily ”Eesti Päevaleht” (93M words) and a corpus of news stories from an online news site *etv24.ee* (4.8M words).

The LM was constructed by first processing the text corpora using the Estonian morphological analyzer and disambiguator (Kaalep and Vaino 2001). Using the information from morphological analysis, it’s possible to split compound words into particles and separate morphological suffixes from preceding stems. The LM vocabulary was created by selecting the most likely 60K units from the mixture of the corpora. A trigram LM was estimated using modified Kneser-Ney smoothing. The morpheme-level (unnormalized) perplexity of the LM against the a language modelling test set 0.8%.

4.2. Acoustic modeling

The acoustic models for recognition experiments were trained on the Estonian SpeechDat-like phonetic database (Meister et al. 2003) which has about 241.1 hours of audio data.

The open source SphinxTrain toolkit was used for training the acoustic models. Models were created for 25 phonemes, the five filler/noise types and silence. Our acoustic models are fairly conventional MFCC-based tied triphone HMM models with 8000 shared states in total. Each state is modeled by 8 Gaussian mixture components. The final models were adapted via MLLR, using around 30 minutes of hand-transcribed broadcast news data from various speakers.

The pronunciation dictionary was created from word orthography using a set of context sensitive grapheme-to-phoneme rewrite rules and a small and incomplete set of foreign name pronunciation rules.

4.3. LSA estimation

The word-based, lemma-based and morpheme-based LSA models were created based on the word statistics in a corpus of approximately 0.5M documents (mainly newspaper articles). For creating the lemma-based LSA model, words in the training data were first replaced with their respective lemmas using the morphological analyzer/disambiguator. The same tool was used for splitting the words into morphemes for creating the morpheme-based model.

The LSA models were constructed using a vocabulary of 60 000 most frequent units. The vocabulary of the morpheme-based model was taken from the N-gram LM. The OOV-rate of the word-based LSA model vocabulary is 13.7%. The corresponding value for the lemma-based model was 5.9%. The OOV-rate of the morpheme-based model is the same as that of the N-gram LM – 0.8%. We used rank-200 SVD.

4.4. Results

First, we looked how adaptation reflects in LM perplexity. We took 18 articles from a newspaper that were not present in the training corpus. The average length of the articles was 28 sentences. The first 10 sentences were used for adaptation and the rest were used for perplexity calculations. The perplexities of the background trigram model and the adapted models, together with mean relative improvements across test articles are given in Table 1. Lemma and morpheme-based adaptation performed statistically significantly better than word-based adaptation.

Next, adaptation was applied to an Estonian broadcast news transcription task. The material consists of studio-recorded hourly short radio news broadcasts. The broadcasts were manually segmented into stories and sentences. For testing, we used around 21 minutes of audio (193 sentences). For tuning the parameters, we used around 11 minutes of audio (101 sentences). The average number of sentences per story was 4.6. Recognition was performed in two passes: in the 1st pass, the background LM was used. Output from the 1st pass was used for constructing adapted LMs for each story which were then used in the 2nd pass of recognition. The recognition results of the baseline system and after applying adapted LMs are listed in Table 2. All systems using adapted models performed significantly better than the baseline system. The morpheme-based adapted system performed significantly better than the other two adapted systems. There was no significant difference between word-based and lemma-based adaptation.

Table 1: LM perplexities before and after adaptation.

System	PPL
Baseline	192
Word-based adapt.	154 (-20%)
Lemma-based adapt.	**151 (-22%)**
Morpheme-based adapt.	**152 (-21%)**

Table 2: Letter error rates before and after adaptation.

System	LER
Baseline	7.1
Word-based adapt.	6.7 (-6%)
Lemma-based adapt.	6.6 (-7%)
Morpheme-based adapt.	**6.4 (-9%)**

5. Conclusion

We presented a statistical LM adaptation method that is especially suitable for agglutinative and highly inflected languages that use sub-word units as basic units for N-gram language modelling. The method relies on a separate independent LSA model to retrieve documents from a large corpus that are semantically similar to a small ”adaptation seed” and applies fast marginal adaptation of background N-gram models based on the resulting adaptation corpus. We compared three different basic units – words, lemmas and morphemes – for modelling document similarity.

Experiments showed that morpheme-based adaptation provided consistently good results compared to other adaptation models. The results were somewhat surprising, since we initially doubted in the ability of morphemes to carry semantic content.

References

Alumäe, Tanel 2004. Large vocabulary continuous speech recognition for Estonian using morpheme classes. In: *Proceedings of ICSLP 2004 - Interspeech*, Jeju, Korea. 389–392

Bellegarda, Jerome R. 1998. A multispan language modeling framework for large vocabulary speech recognition. In: *IEEE Transactions on Speech and Audio Processing* **6(5)**, 456–467

Kaalep, Heiki-Jaan; Muischnek, Kadri 2005. The corpora of Estonian at the University of Tartu: the current situation. In: *The Second Baltic Conference on Human Language Technologies : Proceedings*, Tallinn, Estonia. 267–272

Kaalep, Heiki-Jaan; Vaino, Tarmo 2001. Complete morphological analysis in the linguist’s toolbox. In: *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, Tartu, Estonia. 9–16

Klakow, Dietrich 2006. Language model adaptation for tiny adaptation corpora. In: *Proceedings of Interspeech 2006 - ICSLP*, Pittsburgh, PA, USA. 2214–2217

Kneser, R.; Peters, J.; Klakow, D. 1997. Language model adaptation using dynamic marginals. In: *Proceedings of Eurospeech*: Vol. 4, Rhodes, Greece. 1971–1974

Meister, Einar; Lasn, Jürgen; Meister, Lya 2003. Development of the Estonian SpeechDat-like database. In: *Proceedings of Eurospeech*: Vol. 2, Geneva, Switzerland. 1601–1604

Turunen, Ville; Kurimo, Mikko 2006. Using latent semantic indexing for morph-based spoken document retrieval. In: *Proceedings of Interspeech*, Pittsburgh, USA. 341–344

TANEL ALUMÄE is a researcher at the Institute of Cybernetics at Tallinn University of Technology. In 2006 he received a PhD from the same university. Research interests include speech and language processing, statistical methods, machine learning. E-mail: tanel.alumae@phon.ioc.ee

TOOMAS KIRT works for Institute of Cybernetics at TUT as a researcher. In 1999 he received a M.Sc. degree in Information Processing at Tallinn University of Technology. Currently he is a doctoral student at the same university and his research interests are data analysis, pattern recognition and artificial intelligence. E-mail: Toomas.Kirt@mail.ee