LSA-based Language Model Adaptation for Highly Inflected Languages

Tanel Alumäe, Toomas Kirt

Institute of Cybernetics at Tallinn University of Technology, Estonia
{tanel.alumae, toomas.kirt}@phon.ioc.ee

Abstract

This paper presents a language model topic adaptation framework for highly inflected languages. In such languages, subword units are used as basic units for language modeling. Since such units carry little semantic information, they are not very suitable for topic adaptation. We propose to lemmatize the corpus of training documents before constructing a latent topic model. To adapt language model, we use few lemmatized training sentences to find a set of documents that are semantically close to the current document. Fast marginal adaptation of subword trigram language model is used for adapting the background model. Experiments on a set of Estonian test texts show that the proposed approach gives a 19% decrease in language model perplexity. A statistically significant decrease in perplexity is observed already when using just two sentences for adaptation. We also show that the model employing lemmatization gives consistently better results than the unlemmatized model.

Index Terms: speech recognition, language model adaptation, LSA, inflected languages

1. Introduction

Language model adaptation is a task of building a language model (LM) for speech recognition 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 language models [1]. The LSA-based approach gradually adapts the background language model 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 effectively 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 additions, 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 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 are used as basic units in language modeling. Subword units may be found using morphological analysis, as has been proposed for Korean [2], Estonian [3] or discovered automatically in a data-driven manner [4]. Since the subword units carry much less semantic information than the full words, it makes them inappropriate for topic adaptation. A straightforward workaround would be to reconstruct words from recognized subword units, and use the words for language model adaptation. However, the high variety of different word inflections and the resulting sparseness of word-document occurrences make this approach problematic.

Based on these observations, this paper investigates a LSA-based language model adaptation approach for highly inflected languages. We propose to use a morphological analyzer to find canonical forms or *lemmas* for all words before mapping them into the LSA space. Lemmatization aims to reduce the lexical variety of the language and should act as an additional smoothing measure. In our approach, a very small adaptation corpus (we experiment with two to ten sentences) is needed to find a mapping of the current topic in the LSA space. Next, we find N training documents in the LSA space that lie closest to the pseudo-document consisting of the adaptation sentences. We use the resulting corpus to estimate sub-word unigram language model and apply fast marginal adaptation (FMA) as proposed in [5] to combine the unigram with the background trigram model. Similar LSA-based adaptation methods have also been investigated recently, e.g. [6, 7]. In [8], a somewhat similar method was proposed 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. There have been previous attempts in language model adaptation for inflected languages. In [9], inflected words are clustered according to fuzzy string comparison rules, and language model is adapted by only using the texts from the closest topic from a predefined topic set for training. Naive Bayes classifier and TDIDF classifier is used for topic detection. The presented approach differs from the previous ones in the use of lemmatization to deal with the inflected words problem and in the use of LSA-space document closeness measure for selecting in-domain training corpus.

The paper is organized as follows: in section 2, we provide an overview of Latent Semantic Analysis. In section 3, we describe how in-domain documents are selected from the training corpus and how FMA is used for language model adaptation. In section 4, we describe the experiments and the results, followed by a conclusion in section 5.

2. Review of Latent Semantic Analysis

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 in [1] is

$$W_{ij} = (1 - \varepsilon_i)\log_2\left(1 + \frac{c_{ij}}{n_j}\right) \tag{1}$$

where c_{ij} is the number of times w_i occurs in d_j, n_j is the length of document d_j and ε_i is the normalized term entropy of w_i in the training corpus $\mathcal{T}$. Term entropy reflects the indexing value of the word w_i and can be calculated as

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

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

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

$$W \approx \hat{W} = USV^T \tag{3}$$

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.

In this task, we also apply lemmatization before constructing the original term-document matrix W. Since LSA uses the bag-of-words paradigm, syntactical information that is carried with the word inflections can be considered redundant and the lemmatization acts as the first step in the dimensionality reduction.

3. Language model adaptation approach

For language model adaptation, we apply fast marginal adaptation using a unigram model trained on weighted counts from a set of in-domain documents. The in-domain documents are retrieved by selecting the 500 documents that are closest to the adaptation data in the LSA space.

3.1. Selecting adaptation documents

To find the closest documents to the given adaptation data in the LSA space, we first convert the lemmatized 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

$$\tilde{v}_p = \tilde{d}_p^T US^{-1} \tag{4}$$

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

$$K(\tilde{v}_p, v_i) = \cos(\tilde{v}_p S, v_i S) = \frac{\tilde{v}_p S^2 v_i^T}{\|\tilde{v}_p S\|\,\|v_j S\|} \tag{5}$$

In this way, the training documents are ranked by their closeness measure and the top documents can be selected for use as adaptation data. Of course, if we only need the closest documents and are not interested in the actual closeness values, the cosine calculation can be discarded and the ranking be inverted. However, in the next section we explain how we use the closeness values for improving the unigram compilation.

3.2. Weighted unigram counts

Before constructing the unigram models, we apply count weighting, depending on the closeness of the training document to the adaptation data. The total count $c_{Adap}(i)$ of the ith word for building unigrams is calculated as follows:

$$c_{Adapt}(i) = \sum_{j=1}^{L}\left(1 - \frac{K(\tilde{v}_p, v_j)}{\pi}\right) * c_{ij} \tag{6}$$

where L is the number of closest adaptation document to use. In our experiments, $L = 500$ worked well. The distance measure $K(\tilde{v}_p, v_j)$ is calculated as given in (5) and lies in $[0, \pi]$. This approach enables to emphasize the counts in documents that lie closer to the adaptation data. Using the resulting fractional counts, we apply absolute discounting to estimate unigram models. The described count weighting gave slightly better mean improvements in perplexity than when using unweighted counts, however, there were no statistical significance between the results.

3.3. Fast marginal adaptation

Fast Marginal Adaptation [5] is a method to quickly adapt a given language model to in-domain text characteristics. It uses the trigram trained on the background corpus as the initial language model. 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:

$$P_{Adap}(w|h) = \frac{\alpha(w)P_{BG}(w|h)}{Z(h)} \tag{7}$$

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:

$$\alpha(w) \approx \left(\frac{P_{Adap}(w)}{P_{BG}(w)}\right)^{\beta} \tag{8}$$

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. In our reported experiments, we set β equal to 0.5. 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:

$$Z(h) = \sum_{w}\alpha(w)P_{BG}(w|h) \tag{9}$$

For efficiency, $\alpha(w)$ can be precalculated for the whole vocabulary.

We also experimented with adjusted FMA, as proposed in [8]. It uses unigram on the adaptation corpus as the starting distribution and adjusts it with a scaling factor that takes into account the history, using the background trigram probabilities:

$$P_{Adap}(w|h) = \frac{1}{Z(h)} \left(\frac{P_{BG}(w|h)}{P_{BG}(w)} \right)^{\beta} P_{Adap}(w) \qquad (10)$$

However, the adjusted FMA didn't give consistent improvement over the original FMA in our experiments, in particular because it severely degraded the perplexity of a few articles.

4. Experiments

For training the background language model, we used the following subset of the Mixed Corpus of Estonian [10], compiled by the Working Group of Computational Linguistics at the University of Tartu: daily newspaper ”Postimees” (33 million words), weekly newspaper ”Eesti Ekspress” (7.5 million words), weekly newspaper ”Maaleht” (4.3 million words), Estonian original prose from 1995 onwards (4.2 million words), academic journal ”Akadeemia” (7 million words), transcripts of Estonian Parliament (13 million words), weekly magazine ”Kroonika” (600 000 words). The SRILM toolkit [11] was used for selecting language model vocabulary and compiling the language model. The language model was constructed by first processing the text corpora using the Estonian morphological analyzer and disambiguator [12]. The analyzer uses a rule-based approach, a statistical method based on hidden Markov models is used for disambiguation. Using the information from morphological analysis, it's possible to split compound words into particles and separate morphological suffixes from preceding stems. Language model vocabulary was created by selecting the most likely 60 000 units from the mixture of the corpora, using a set of dictated sentences from various domains as held-out text for optimization. Using the vocabulary of 60 000 particles, a trigram language model was estimated for each training corpus subset. The cutoff value was 1 for both bigrams and trigrams, i.e., singleton n-grams were included in the models. A modified version of Kneser-Ney smoothing as implemented in SRILM was applied. Finally, a single LM was built by merging the seven models, using interpolation coefficients optimized on a set of dictated sentences from various domains.

The word-based and lemma-based LSA models were created based on the word statistics in the three above-mentioned newspaper corpora. The corpora were divided into documents according to the markers as provided by the corpus creators. The number of documents was 111 833. For creating the lemma-based LSA model, words in the training data were first replaced with their respective lemmas using the morphological analyzer. The number of different words in the LSA training data was 1 595 478, as opposed to 787 177 lemmas. The coverage of word and lemma vocabularies selected on the maximum likelihood principle for varying vocabulary sizes are described in Figure 1.

The LSA models were constructed using a vocabulary of 60 000 most frequent units. The initial word-document matrix contained about 25.5 million entries, and the lemma-document matrix about 23.8 million entries. The rank-200 SVD was calculated using PROPACK[1].

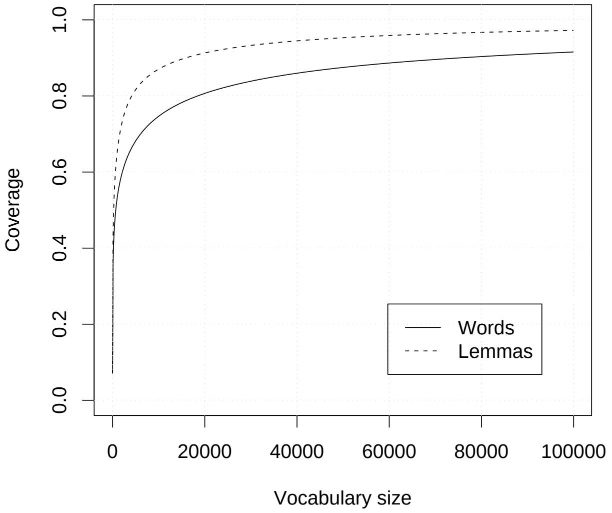

Figure 1: Coverage of word and lemma vocabulary with increasing vocabulary size, using maximum likelihood vocabulary.

For testing language model adaptation, we took 18 articles from the daily newspaper ”Postimees” 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. In the experiments were adaptation data was restricted to less than 10 sentences, the remaining of the 10 sentences was discarded, i.e., the perplexity was always calculated from the same set of sentences. Perplexities were calculated on the morpheme level, that is, the sentences were processed by the morphological analyzer and words were split into particles, in the same way as the training corpus was preprocessed.

The word-level out-of-vocabulary (OOV) rate of the test sentences against the LSA model was 14.1%. The lemma-level OOV-rate was 6.3%.

Language model adaptation was tested with three different adaptation data: first two sentences of every article (first of which was the article heading), first five and first ten sentences. The trigram perplexities of the background trigram model and the adapted models, together with mean relative improvements across test auricles are given in Table 1.

	Lemma-based LSA	Word-based LSA
Background trigram	215	215
Adapted on 2 sentences	187 (-15%)	192 (-12%)
Adapted on 5 sentences	181 (-17%)	187 (-14%)
Adapted on 10 sentences	176 (-19%)	182 (-16%)

Table 1: Perplexities before and after adaptation using the word-based *vs.* the lemma-based LSA model, together with mean improvement in perplexity across the articles.

The Wilcoxon test of statistical significance shows that all adapted models improve the perplexity results significantly over the baseline background trigram. The table also shows that the

[1] http://sun.stanford.edu/~rmunk/PROPACK/

lemma-based models result in better mean improvement than the corresponding word-based models, however, the difference is only significant with the model that is adapted on ten sentences. This can be explained by the fact that when using two or five sentences for adaptation, the adaptation pseudo-document does not always settle correctly in the LSA space, and the variation of the perplexity improvement across articles is higher. The distribution of perplexity improvement across articles when using different adapted models is graphically depicted in Figure 2. The box plots show the median, lower and upper quartiles, lowest and highest values and any outliers of the perplexity improvements.

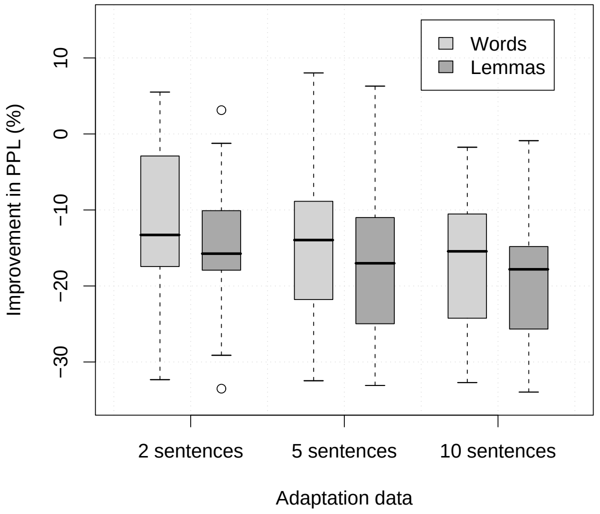

Figure 2: Box plot of perplexity improvement across test articles when using different adaptation data.

5. Conclusion and future work

We investigated the use of lemmatization for adapting language models for highly inflected languages. Lemmatization decreases the lexical variety of the language and removes inflectional word features that should not be semantically relevant. We proposed to use a LSA-based in-domain document selection method and fast unigram adaptation of background language models that use subword units as basic modeling units. Using a set of Estonian test articles, we obtained a significant perplexity improvement when using only up to ten sentences for adaptation. Lemma-based adaptation showed to provide consistently better results than corresponding word-based adaptation.

The proposed method can be applied for other highly inflected languages, regardless of the selection of language modeling units. However, a language specific morphological analyzer is needed for lemmatization.

Future work will investigate unsupervised adaptation, that is, using the speech recognizer output for gradually adapting the language model. Also, we are interested in the impact of the adaptation method on speech recognition word error rate.

LSA-based language modeling techniques have been shown to benefit from word and document clustering that enable to apply a variety of smoothing algorithms. Thus, further improvement can be expected when such methods are implemented into our framework.

Finally, the LSA-based method for selecting in-domain training documents should be compared with the perplexity-minimization based document selection method that was proposed in [8], using sub-word language modeling units.

6. Acknowledgements

This work has been partially funded by the Estonian National Programme for Language Technology.

7. References

[1] J. R. Bellegarda, “A multispan language modeling framework for large vocabulary speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 456–467, September 1998.

[2] O.-W. Kwon and J. Park, “Korean large vocabulary continuous speech recognition with morpheme-based recognition units,” *Speech Communication*, vol. 39, pp. 287–300, 2003.

[3] T. Alumäe, “Large vocabulary continuous speech recognition for Estonian using morpheme classes,” in *Proceedings of ICSLP 2004 - Interspeech*, Jeju, Korea, 2004, pp. 389–392.

[4] V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo, “Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner,” in *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, pp. 2293–2296.

[5] R. Kneser, J. Peters, and D. Klakow, “Language model adaptation using dynamic marginals,” in *Proceedings of Eurospeech*, vol. 4, Rhodes, Greece, 1997, pp. 1971–1974.

[6] B. Chen, “Dynamic language model adaptation using latent topical information and automatic transcripts,” in *IEEE International Conference on Multimedia and Expo*, Amsterdam, The Netherlands, 2005, pp. 97–100.

[7] Y.-C. Tam and T. Schultz, “Unsupervised language model adaptation using latent semantic marginals,” in *Proceedings of Interspeech 2006 - ICSLP*, Pittsburgh, PA, USA, 2006, pp. 2206–2209.

[8] D. Klakow, “Language model adaptation for tiny adaptation corpora,” in *Proceedings of Interspeech 2006 - ICSLP*, Pittsburgh, PA, USA, 2006, pp. 2214–2217.

[9] M. S. Maučec, Z. Kačič, and B. Horvat, “A framework for language model adaptation for highly-inflected Slovenian,” in *Proceedings of ITRW on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, 2001, pp. 211–214.

[10] H.-J. Kaalep and K. Muischnek, “The corpora of Estonian at the University of Tartu: the current situation,” in *The Second Baltic Conference on Human Language Technologies : Proceedings*, Tallinn, Estonia, 2005, pp. 267–272.

[11] A. Stolcke, “SRILM – an extensible language modeling toolkit,” in *Proceedings of ICSLP*, vol. 2, Denver, USA, 2002, pp. 901–904.

[12] H.-J. Kaalep and T. Vaino, “Complete morphological analysis in the linguist’s toolbox,” in *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, Tartu, Estonia, 2001, pp. 9–16.