Unsupervised topic adaptation for morph-based speech recognition

André Mansikkaniemi[1], *Mikko Kurimo*[2]

[1]Aalto University, School of Science, Department of Information and Computer Science
[2]Aalto University, School of Electrical Engineering, Department of Signal Processing and Acoustics
andre.mansikkaniemi@aalto.fi, mikko.kurimo@aalto.fi

Abstract

Topic adaptation in automatic speech recognition (ASR) refers to the adaptation of language model and vocabulary for improved recognition of in-domain speech data. In this work we implement unsupervised topic adaptation for morph-based ASR, to improve recognition of foreign entity names. Based on first-pass ASR hypothesis similar texts are selected from a collection of articles, which are used to adapt the background language model. Latent semantic indexing is used to index the adaptation corpus and ASR output. We evaluate three different types of index terms and their usefulness in unsupervised LM adaptation: statistical morphs, words, and a combination of morphs and words. Furthermore, we implement vocabulary adaptation alongside unsupervised LM adaptation. Foreign word candidates are selected from the in-domain texts, based on how likely they are topic-related foreign entity names. Adapted pronunciation rules are generated for the selected foreign words. Morpheme adaptation is also performed by restoring over-segmented foreign words back into their base forms, to ensure more reliable pronunciation modeling.

Index Terms: unsupervised language model adaptation, vocabulary adaptation, morph-based speech recognition

1. Introduction

Statistical morph-based automatic speech recognition (ASR) enables the recognition of an unlimited number of words. For a morphologically rich language such as Finnish, morph-based ASR has improved recognition accuracy significantly [1]. A large error source that still remains an unsolved problem for Finnish ASR is foreign entity names (FENs) [2].

Foreign words are difficult to recognize because they are usually topic-specific words, unreliably estimated by the background language model. Additionally, the pronunciation rules of Finnish that cover native words usually give incorrect pronunciation for foreign words. A problem which is specific to statistical morph-based ASR is the over-segmentation of foreign words. When words are built from morph units that are statistically discovered from Finnish texts, the foreign words which do not contain Finnish morphemes will be built from very short units. Using many short morphs and untypical morph sequences is problematic for both language and pronunciation modeling.

In this work we adapt both the language model, pronunciation rules and morph segmentation for Finnish morph-based ASR to improve the recognition of foreign entity names. We call these different but connected methods LM adaptation, pronunciation adaptation and morpheme adaptation. The main novelty in this paper is the implementation of these methods as fully unsupervised processes and combining them to improve recognition of foreign words.

In LM adaptation the background LM is adapted to model topic-specific foreign words and their contexts more accurately. Because the topic and these words are not known in advance, the first-pass ASR output is used to select topically related texts from an information retrieval (IR) source, which consists of documents retrieved from the Web. To select most suitable texts despite the ASR errors, both the recognized words and statistical morphs are used as index terms and latent semantic indexing (LSI) [3] to match the documents. Then an in-domain LM is trained on the selected texts and utilized to adapt the background LM.

Vocabulary adaptation is performed to improve pronunciation modeling of foreign entity names. Because the topic-specific foreign words are not known in advance, they need first to be identified from the in-domain text. A score is calculated based on perplexity using a letter n-gram trained from Finnish texts to measure how foreign a word appears to be. A second score based on either the tf-idf measure or document cosine similarity is calculated to measure the word's relatedness to the topic. The two scores are combined into a total score to select the most likely and useful topic-specific foreign words for vocabulary adaptation. Pronunciation rules are then adapted by generating new pronunciations for the selected foreign word candidates. Morpheme adaptation is performed by restoring over-segmented foreign words to include their unsegmented base forms as morphs.

The architecture of the adaptation framework used is illustrated in Figure 1. Evaluation in this work is done on Finnish broadcast news.

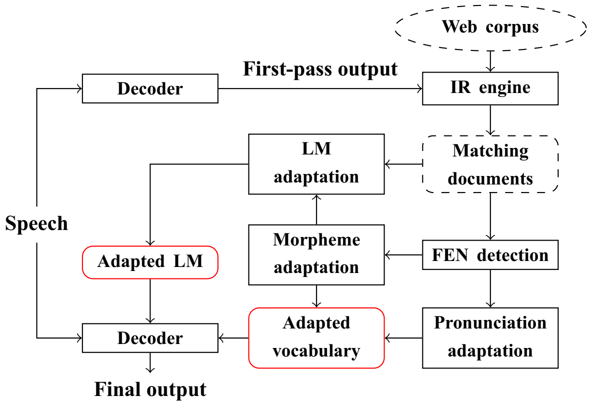

Figure 1: First-pass ASR output is used to find topically related articles from a Web corpus. The retrieved texts are used to adapt the language model and vocabulary. The adapted models are used in second-pass recognition.

2. Unsupervised LM adaptation

Unsupervised LM adaptation methods have been extensively studied in recent years. A common approach is to identify the topic from the speech data using first-pass ASR output and select matching articles from a text collection to train an in-domain language model [4]. This approach is based on indexing the text in order to calculate similarity measures between the ASR output and articles.

Another approach is to extract keywords from the ASR hypothesis and send them to a Web search engine to retrieve in-domain articles from the Internet [5] [6].

Many past works about unsupervised LM adaptation have dealt with comparing different indexing and LM adaptation methods. We concentrate our efforts in this work to highlight specific issues concerned with unsupervised LM adaptation for a morph-based ASR system.

The focus point is to study the effect on performance for different indexing terms. In morph-based speech recognition words are segmented into sub-word units or morphemes. We use the statistical morph-segmentation algorithm Morfessor to split words into smaller units [7]. We want to compare statistical morphs as indexing terms against base forms of the words.

2.1. Latent semantic indexing

In information retrieval documents are represented by vectors. The vector components are word weights, indicating the importance of an individual word to a document. Word weights are typically calculated using the term frequency-inverse document frequency (tf-idf) measure.

Latent semantic indexing (LSI) transforms the document-word matrix into lower space using singular value decomposition [8]. The benefit of LSI is dimensionality reduction and tying together of words with similar meaning.

The relatedness of two documents (a,b) can be calculated using a vector similarity measure, such as the cosine distance (1).

$$similarity(a,b) = \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n}(a_i)^2} \times \sqrt{\sum_{i=1}^{n}(b_i)^2}} \quad (1)$$

In this work LSI is implemented using the *Gensim* software package [9]. Articles in the background corpus are transformed into an LSI space with 100 latent variables. The ASR output is transformed into the same LSI space, and the cosine distance is used to find the best matching articles.

2.2. Index terms

In information retrieval, index terms refer to the vocabulary units. Words as such can be used as index terms or their base forms. In statistical morph-based language modeling there's also the option of using the morphs as index terms.

In spoken document retrieval tasks for Finnish, best performance has been found when using a combination of both statistical morphs and words as index terms [10].

In this work statistical morphs, words, and a combination of both morphs and words are evaluated as indexing terms in the context of retrieving topic-related articles based on first-pass ASR output.

2.3. Language model adaptation

An in-domain LM $P_i(w|h)$ is trained on the matching articles retrieved from the Web corpus. Linear interpolation is used to the adapt the background LM $P_B(w|h)$ with the in-domain LM $P_i(w|h)$. The adaptation weight λ is set to a pre-defined value in all experiments.

$$P_{adap_i}(w|h) = \lambda P_i(w|h) + (1-\lambda)P_B(w|h) \quad (2)$$

3. Vocabulary adaptation

Vocabulary adaptation is performed by first detecting topic-specific foreign word candidates in in-domain texts. Thereafter adapted pronunciation rules are generated for the words. Finally foreign words that are over-segmented are restored back into their base forms.

3.1. Foreign word detection

To automatically detect foreign word candidates in in-domain text, we build on our previous algorithm used in [11]. The aim is to produce a sorted list of the most likely foreign entity names (FENs). First step is to select all words starting in uppercase letter as potential FEN candidates. Sorting is done by calculating two selection scores for each word.

First selection score is based on how *foreign* the word appears to be. For this we calculate the letter perplexity (*ppl*) for each FEN candidate using a letter-ngram model trained on Finnish words. Words with higher letter perplexity are more likely to be foreign words.

Second selection score is based on how important the word is to the *topic*. To measure this the term frequency-inverse document frequency (*tf-idf*) is calculated for each word, in relation to the background corpus and the retrieved article that it appears in. An alternative to measure *topic* relatedness is to use the cosine similarity score (*sim*) between the ASR output and the document in which the word appears in. The selection scores are normalized using max-min normalization and a final score is calculated by multiplying the two normalized scores (3).

$$score(word) = foreign(word) * topic(word) \quad (3)$$

When selection scores for all FEN candidates have been calculated they are sorted in descending order. Words at the top of the list can then be considered to be more likely of foreign origin and related to the topic. A final set of FEN candidates are chosen for adaptation using a cut-off value. In this work a maximum of 30% of the FEN candidates are included for adaptation. Because restoring over-segmented FENs (morpheme adaptation) results in new vocabulary units an additional constraint is set that the vocabulary size can only increase with 5%.

3.2. Pronunciation adaptation

For Finnish ASR systems pronunciation of arbitrary words is simply modeled by mapping letters directly to phonemes. Foreign names are often pronounced according to their original languages, which can have more complicated pronunciation rules.

In this work *Sequitur G2P* is used, a data-driven grapheme-to-phoneme converter based on joint-sequence models [12]. A pronunciation model is trained on a manually constructed foreign word lexicon consisting of 2000 foreign entity names with a manually given pronunciation hand-picked from a Finnish newswire text collection. The pronunciation model is used to generate the most probable pronunciation variants for the foreign entity name candidates found in the adaptation text. For each FEN candidate four alternative pronunciations are generated and added to the pronunciation dictionary.

3.3. Morpheme adaptation

Current state of the art Finnish language modeling is morpheme-based, where words are segmented into sub-word units (morphemes) [1]. This allows the system to cover a large number of words which result from the highly agglutinative word morphology.

Over-segmentation often occurs for rare names or previously unseen words found in adaptation texts (***mcafee* ⇒ *m cafe e***). To ensure reliable pronunciation modeling of foreign entity names it's preferable to keep the base form intact. In our previous work we found that recognition accuracy is improved when pronunciations of foreign words are modeled as single units instead of trying to map whole word pronunciation rules onto separate morphs [13]. This is the main reason for performing morpheme adaptation, which is done by restoring over-segmented FENs back into their base forms (***m cafe e* ⇒ *mcafee***).

4. Experiments

4.1. Speech data

Speech data consists of Finnish radio and TV news segments in 16 kHz audio. The recordings were collected in 2011-2012 from YLE news and sports programs. The data is divided into a 1 hour development set and a 10 hour evaluation set. In the development set 8.4% of the words were categorized as foreign words compared to 4.1% in the evaluation set.

4.2. System and models

Speech recognition experiments in this work were run on the Aalto speech recognizer. The background language model was trained on the Kielipankki corpus (140 million words) [14]. A segmentation model and a 45k morph lexicon were trained on the Kielipankki corpus using Morfessor [7]. A Kneser-Ney smoothed varigram LM (n=12) was trained on the morph-segmented background corpus with the variKN language modeling toolkit [15]. A letter trigram model was trained on the Kielipankki word list for the FEN detection algorithm.

4.3. Adaptation data

Language model adaptation data was collected from the Web. Articles were retrieved automatically from the Web from a set of pre-defined Finnish online news sources. In total over 44 000 Finnish news articles were retrieved (over 7 million words). The Web corpus and the Kielipankki background corpus are both mainly made up of news articles. The Web corpus is more current however and better matched in terms of topics and vocabulary to the speech data. Therefore it can be considered to be an in-domain adaptation set.

4.4. Experimental setup

Speech recognition experiments are conducted in this work using the unsupervised LM adaptation framework. Based on first-pass ASR output *N* best matching articles are selected from the Web corpus. A Kneser-Ney smoothed varigram LM (n=6) is trained on the selected articles. The in-domain LM is adapted with the background LM using linear interpolation (λ = 0.1). Vocabulary adaptation is performed on the retrieved in-domain text. Different vocabulary selection scores (*ppl*, *ppl*tfidf*, and *ppl*sim*) are evaluated. The adapted LM and vocabulary are used in second-pass recognition.

5. Results

5.1. Development set

First experiments were conducted on the 1 hour development set. We tested different parameter values and indexing terms for unsupervised LM adaptation and different selection scores for vocabulary adaptation. The word error rate (WER) and foreign entity name error rate (FENER) are reported in the results.

5.1.1. Indexing terms

Results of first set of experiments are in Table 1. Performance of unsupervised LM adaptation was evaluated using morph-based indexing, word-based indexing and a combination of both morphs and words. Also different values for *N*, the number of articles retrieved, were tested.

Index term	Articles retrieved	WER	FENER
-		**33.0**	**80.6**
Morphs	1000	31.1	73.2
	5000	**30.4**	**72.1**
	10000	30.6	72.5
Words	1000	30.8	72.3
	5000	**30.0**	71.9
	10000	30.1	**71.5**
Morphs+Words	1000	30.8	73.2
	5000	**30.1**	**72.1**
	10000	30.1	72.3

Table 1: Results of unsupervised LM adaptation experiments on development set. Different indexing terms (morphs, words, morphs+words) and number of retrieved articles are evaluated.

There is a significant performance improvement for all system settings compared to the baseline result (33.0%). There is a relative WER reduction between 8% (morph-based) and 9% (word-based and combined) when *N*=5000. Error rate of foreign words (FENER) drops with around 10% for all adaptation settings. Performance doesn't improve anymore at *N*=10000.

Word-based indexing is slightly better than morph-based indexing, in terms of both lower WER and FENER. Combining morph-based and word-based indexing doesn't give any additional improvement.

5.1.2. Vocabulary adaptation

Results of vocabulary adaptation experiments are in Table 2. Word-based indexing was used. At best, vocabulary adaptation lowers WER with 2% (*N*=10000, *ppl*sim*). It seems that vocabulary adaptation is more efficient when using larger adaptation sets ($N \geq 5000$). Including *topic* scores (*tfidf*, *sim*) for FEN selection also have more effect with larger adaptation sets. The selection score based on letter-ngram perplexity (*ppl*) and document cosine similarity (*sim*) produces the lowest FENER values, with around 10% relative improvement compared to an unadapted vocabulary.

In a second set of experiments, we studied the performance of the individual parts of the vocabulary adaptation algorithm (morpheme and pronunciation adaptation). Results are in table 3. Morpheme adaptation has a detrimental effect on performance when applied separately. Recognition of foreign words peaks when using a combination of both morpheme and pronunciation adaptation. On the other hand, applying only pronunciation adaptation achieves lower average WER.

Index term	Articles retrieved	Vocabulary adaptation	WER	FENER
-			**33.0**	**80.6**
Words	1000	-	30.8	72.3
		ppl	**30.7**	66.4
		*ppl*tfidf*	31.5	69.4
		*ppl*sim*	31.3	**64.5**
	5000	-	30.0	71.9
		ppl	**29.7**	68.7
		*ppl*tfidf*	29.9	69.4
		*ppl*sim*	29.8	**64.7**
	10000	-	30.1	71.5
		ppl	29.8	70.0
		*ppl*tfidf*	29.9	70.6
		*ppl*sim*	**29.5**	**65.2**

Table 2: Results of unsupervised vocabulary adaptation experiments on development set. Different vocabulary selection scores (*ppl*, *ppl*tfidf*, *ppl*sim*) are evaluated.

Vocabulary adaptation (*ppl*sim*)		WER	FENER
Morpheme	Pronunciation		
x		30.5	74.3
	x	**29.3**	67.0
x	x	29.5	**65.2**

Table 3: Results of adaptation experiment (word-based indexing, N=10000, *ppl*sim*) on development set. Vocabulary adaptation components, morpheme and pronunciation adaptation, are evaluated separately.

5.2. Evaluation set

Results of experiments on the 10 hour evaluation set are in Table 4. Word-based indexing was used for unsupervised LM adaptation. Letter ngram-perplexity (*ppl*) and document cosine similarity (*sim*) were used in vocabulary adaptation.

Index term	Articles retrieved	Vocabulary adaptation	WER	FENER
-			**30.0**	**71.7**
Words	1000	-	**29.7**	68.0
		*ppl*sim*	31.0	**63.9**
	5000	-	**29.5**	66.4
		*ppl*sim*	30.0	**61.1**
	10000	-	**29.4**	66.5
		*ppl*sim*	29.7	**61.1**

Table 4: Results of experiments on evaluation set.

In terms of performance increase, unsupervised LM adaptation has a more modest effect on the evaluation set compared to the development set. Relative WER reduction is around 2%, and relative FENER reduction is about 5-7%. Vocabulary adaptation further lowers FENER with around 7%. Vocabulary adaptation seems however to have an adverse effect on average word recognition, with slightly higher WER values compared to when only applying LM adaptation.

6. Conclusions

The focus in this work was to improve recognition of foreign entity names which make up a large error source in Finnish ASR. Based on first-pass ASR output topically related articles were selected from a Web corpus, which were used to adapt the background LM. Latent semantic indexing was used to index the Web corpus and ASR output. We evaluated three different types of index terms: statistical morphs, words, and a combination of words and morphs. Results of experiments on the development set indicate that there are only slight differences between the three indexing terms. Word-based indexing performs a little better than morph-based indexing, in terms of both lowering the WER and FENER. No significant performance increase is gained from combined morph-based and word-based indexing, contrary to results in Finnish spoken document retrieval (SDR) [10]. Unsupervised LM adaptation and SDR tasks are probably too different by nature to expect similar performance gains.

Vocabulary adaptation was implemented to improve pronunciation modeling of topic-specific foreign words. Adapted pronunciation rules were generated for the foreign words that were detected in the in-domain documents. Foreign word candidates were selected from the in-domain documents based on a score which measures how likely the word is a topic-specific foreign word. To measure how likely the word is foreign, letter-ngram perplexity (*ppl*) for each word starting in uppercase letter was calculated. Topic-relatedness was calculated using either the *tf-idf* measure or the document cosine similarity (*sim*). A total score was calculated by multiplying the normalized values of *ppl* and *tf-idf* or *sim*. Vocabulary adaptation experiments on the development set give indication that a selection score based on *ppl* and *sim* produces the lowest FENER. Using *tf-idf* measure as a topic score seems to even have a small detrimental effect on FENER values compared to just using *ppl*. It's possible that the *tf-idf* measure elevates topic-specific native words too much. In this scenario it would seem that document cosine similarity functions better as a score to measure topic-relatedness.

Vocabulary adaptation is successful in efficiently lowering error rates for foreign words, but even in the best case average WER is only lowered slightly. This is most evident in experiments on the evaluation set where vocabulary adaptation has a detrimental effect on average WER. The trade-off of improving the recognition of foreign words seems to be poorer recognition of native words. This is probably due to increased acoustic confusability when introducing additional morphs and pronunciation variants. It seems also that combining morpheme and pronunciation adaptation didn't have as much effect on performance as in our previous work where we used a supervised adaptation framework [13]. It's likely that unsupervised FEN detection isn't able to retrieve that many foreign words that are both over-segmented and relevant to the spoken document.

The main aim of this work was to improve recognition of foreign words for Finnish morph-based ASR. The implemented unsupervised topic adaptation framework was successful in significantly lowering error rates for foreign words. Based on results from this work, it's well motivated to focus future research in this topic on lessening the effect of acoustic confusability from vocabulary adaptation. Pronunciation variants need to be pruned more efficiently, so that harmful variants that conflict with correctly recognized native words are removed. Deeper analysis of the FEN detection algorithm is also needed, to find the relation between recall rate and improved recognition of foreign words, and precision rate and average WER.

7. Acknowledgements

This work has been funded by Aalto MIDE programme (project UI-ART) of the Aalto University and by the grant 251170 of the Academy of Finland.

8. References

[1] T. Hirsimäki, J. Pylkkönen, and M. Kurimo, "Importance of high-order n-gram models in morph-based speech recognition", IEEE Trans. Audio, Speech and Lang., pp. 724-732, vol. 17, 2009.

[2] T. Hirsimäki and M. Kurimo, "Analysing recognition errors in unlimited-vocabulary speech recognition", In Proc. of *NAACL*, Boulder, USA, May 2009.

[3] V. Turunen and M. Kurimo, "Using latent semantic indexing for morph-based spoken document retrieval", In Proc. of *INTERSPEECH*, Pittsburgh, USA, September 2006.

[4] L. Chen, J. Gauvain, L. Lamel, and G. Adda, "Unsupervised language model adaptation for broadcast news", In Proc. of *ICASSP*, Hong Kong, April 2003.

[5] G. Lecorve, J. Dines, T. Hain, and P. Motlicek, "Supervised and unsupervised Web-based language model domain adaptation", In Proc. of *INTERSPEECH*, Portland, USA, September 2012.

[6] S. Meng, K. Thambiratnam, Y. Lin, L. Wang, G. Li and F. Seide, "Vocabulary and language model adaptation using just one speech file", In Proc. of *ICASSP*, Dallas, USA, March 2010.

[7] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0", Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology.

[8] S. Deerwater, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis". Journal of the American Society of Information Science, 41(6):391407, 1990.

[9] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora", In Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45-50, May, 2010.

[10] V. Turunen and M. Kurimo, "To recover from speech recognition errors in spoken document retrieval", In Proc. of *EUROSPEECH*, Lisbon, Portugal, September 2005.

[11] A. Mansikkaniemi and M. Kurimo, "Unsupervised vocabulary adaptation for morph-based language models", In Proc. of the *NAACL* 2012 Workshop on the Future of Language Modeling for HLT. ACL, June 2012.

[12] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion", Speech Communication, vol. 50, Issue 5, pp. 434-451, 2008.

[13] A. Mansikkaniemi and M. Kurimo, "Adaptation of morpheme-based speech recognition for foreign entity names", In Proc. of the Fifth International Conference Human Language Technologies - The Baltic Perspective. IOS Press, October 2012.

[14] Department of General Linguistics, University of Helsinki, Linguistics and Language Technology Department, University of Joensuu, Research Institute for the Languages of Finland, and CSC., "Finnish Text Collection - collection of Finnish text documents from years 1990-2000", Available online: http://www.csc.fi/kielipankki/.

[15] V. Siivola, T. Hirsimäki and S. Virpioja., "On growing and pruning Kneser-Ney smoothed n-gram models", IEEE Trans. Audio, Speech and Lang., Vol. 15, No. 5, 2007.