Unsupervised Vocabulary Adaptation for Morph-based Language Models

André Mansikkaniemi and **Mikko Kurimo**
Aalto University School of Science
Department of Information and Computer Science
PO BOX 15400, 00076 Aalto, Finland
{andre.mansikkaniemi,mikko.kurimo}@aalto.fi

Abstract

Modeling of foreign entity names is an important unsolved problem in morpheme-based modeling that is common in morphologically rich languages. In this paper we present an unsupervised vocabulary adaptation method for morph-based speech recognition. Foreign word candidates are detected automatically from in-domain text through the use of letter n-gram perplexity. Over-segmented foreign entity names are restored to their base forms in the morph-segmented in-domain text for easier and more reliable modeling and recognition. The adapted pronunciation rules are finally generated with a trainable grapheme-to-phoneme converter. In ASR performance the unsupervised method almost matches the ability of supervised adaptation in correctly recognizing foreign entity names.

1 Introduction

Foreign entity names (FENs) are difficult to recognize correctly in automatic speech recognition (ASR). Pronunciation rules that cover native words usually give incorrect pronunciation for foreign words. More often the foreign entity names encountered in speech are out-of-vocabulary words, previously unseen words not present in neither the lexicon nor background language model (LM).

An in-domain LM trained on a smaller corpus related to the topic of the speech, can be used to adapt the background LM to give more suitable probabilities to rare or unseen foreign words. Proper pronunciation rules for foreign entity names are needed to increase the probability of their correct recognition. These can either be obtained from a hand-made lexicon or by generating pronunciation rules automatically using for example a trainable grapheme-to-phoneme (G2P) converter.

In morph-based speech recognition words are segmented into sub-word units called morphemes. When using statistical morph-segmentation algorithms such as Morfessor (Creutz and Lagus, 2005) new foreign entity names encountered in in-domain text corpora are often over-segmented (e.g. mcdowell ⇒ mc do well). To guarantee reliable pronunciation modeling, it's preferable to keep the lemma intact. Restoring over-segmented foreign entity names back in to their base forms is referred to as morpheme adaptation in this paper.

This work describes an unsupervised approach to language and pronunciation modeling of foreign entity names in morph-based speech recognition. We will study an adaptation framework illustrated below in Figure 1.

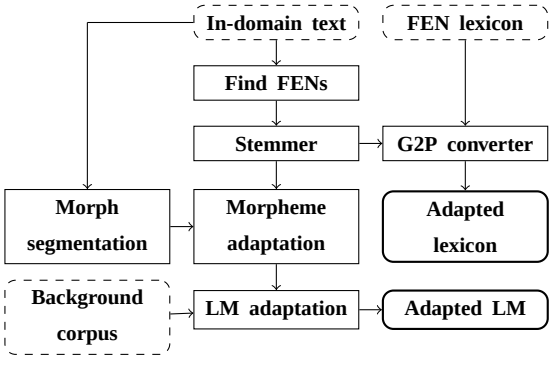

Figure 1: Adaptation framework.

NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, pages 37–40,
Montréal, Canada, June 8, 2012.

The adaptation framework is centered around the following automated steps: 1. Find foreign words in adaptation texts, 2. Convert foreign word candidates into their base forms, 3. Generate pronunciation variants for the retrieved foreign entity name candidates using a G2P converter. Additionally, to facilitate easier and more reliable pronunciation adaptation, the foreign entity names are restored to their base forms in the segmented in-domain text.

The adaptation framework will be compared to a supervised method where the adaptation steps are done manually. The evaluation will be done on Finnish radio news segments.

2 Methods

2.1 Foreign Word Detection

Unsupervised detection of foreign words in text has previously been implemented for English using word n-ngram models (Ahmed, 2005).

Finnish has a rich morphology and using word n-gram models or dictionaries for the detection of foreign words would not be practical. Many of the foreign words occurring in written Finnish texts could be identified from unusual letter sequences that are not common in native words. A letter n-gram model trained on Finnish words could be used to identify foreign words by calculating the average perplexity of the letter sequence in a word normalized by its length.

A two-step algorithm is implemented for the automatic detection of foreign words. First, all words starting in uppercase letters in the unprocessed adaptation text are held out as potential foreign entity names. The perplexity for each foreign word candidate is calculated using a letter-ngram model trained on Finnish words. Words with the highest perplexity values are the most probable foreign entity names. A percentage threshold T for the top perplexity words can be determined from prior information.

The most likely foreign words are finally converted into their base forms using a Finnish stemming algorithm (*Snowball* - `http://snowball.tartarus.org/`).

2.2 Lexicon Adaptation

For Finnish ASR systems the pronunciation dictionary can easily be constructed for arbitrary words by mapping letters directly to phonemes. Foreign names are often pronounced according to their original languages, which can have more complicated pronunciation rules. These pronunciation rules can either be manually added to a lookup dictionary or generated automatically with a grapheme-to-phoneme converter. Constructing a foreign word lexicon through manual input involves a lot of tedious work and it will require a continuous effort to keep it updated.

In this work *Sequitur G2P* is used, a data-driven grapheme-to-phoneme converter based on joint-sequence models (Bisani and Ney, 2008). 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 linguistic origins of the foreign words are mixed but Germanic and Slavic languages are the most common.

The pronunciation model is used to generate the most probable pronunciation variants for the foreign entity name candidates found in the adaptation text.

2.3 Morpheme Adaptation

In current state of the art Finnish language modeling words are segmented into sub-word units (morphemes) (Hirsimäki et. al, 2009). This allows the system to cover a large number of words which result from the highly agglutinative word morphology.

Over-segmentation usually occurs for previously unseen words found in adaptation texts. To ensure reliable pronunciation modeling of foreign entity names it's preferable to keep the lemma intact. Mapping a whole word pronunciation rule onto separate morphemes is a non-trivial task for non-phonetic languages such as English. The morphemes in the in-domain corpus will be adapted such that all foreign words are restored into their base forms and the base forms are added to the morpheme vocabulary. Below is an example. Word boundaries are labeled with the <*w*>-tag.

<*w*> ***oilers*** <*w*> *hävisi* <*w*> ***edmonton*** *in* <*w*> ***com mon we al th*** <*w*> ***sta dium*** *illa* <*w*>

⇒

<*w*> ***oilers*** <*w*> *hävisi* <*w*> ***edmonton*** *in* <*w*> ***commonwealth*** <*w*> ***stadium*** *illa* <*w*>

2.4 Language Model Adaptation

The in-domain adaptation text is segmented differently depending on the foreign entity name candidates that are included. A separate in-domain LM $P_i(w|h)$ is trained for each segmentation of the text. Linear interpolation is used to the adapt the background LM $P_B(w|h)$ with the in-domain LM $P_i(w|h)$.

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

3 Experiments

3.1 Speech Data

Evaluation data consisted of two sets of Finnish radio news segments in 16 kHz audio. All of the recordings were collected in 2011-2012 from YLE Radio Suomi news and sports programs.

The first data set consisted of 32 general news segments. The total transcription length was 8271 words. 4.8% of the words were categorized as foreign entity names (FEN). The second data set consisted of 43 sports news segments. The total transcription length 6466 was words. 7.9% of the words were categorized as foreign entity names.

3.2 System and Models

All speech recognition experiments were run on the Aalto speech recognizer (Hirsimäki et. al, 2009).

The background LM was trained on the Kielipankki corpus (70 million words). A lexicon of 30k morphs and a model of morph segmentation was learnt from the same corpus as the LM using Morfessor (Creutz and Lagus, 2005). The baseline lexicon was adapted with a manually transcribed pronunciation dictionary of 2000 foreign entity names found in Finnish newswire texts. A Kneser-Ney smoothed varigram LM (n=12) was trained on the segmented corpus with the variKN language modeling toolkit (Siivola et al., 2007).

LM adaptation data was manually collected from the Web. On average 2-3 articles were gathered per topic featured in the evaluation data sets. 120 000 words of text were gathered for LM adaptation on the general news set. 60 000 words were gathered for LM adaptation on the sports news set.

The foreign word detection algorithm and a letter trigram model trained on the Kielipankki word list were used to automatically find foreign entity names in the adaptation texts and convert them into their base forms. Different values were used as percentage threshold *T* (30, 60, and 100%).

The adaptation texts were segmented into morphs with the segmentation model learnt from the background corpus. Morpheme adaptation was performed by restoring the foreign entity name candidates into their base forms. Separate in-domain varigram LMs (n=6) were trained for adaptation data segmented into morphs using each choice of *T* in the foreign name detection. The background LM was adapted with each in-domain LM separately using linear interpolation with weight λ = 0.1 chosen based on preliminary experiments.

A pronunciation model was trained with *Sequitur G2P* on the manually constructed foreign word lexicon. The number of the most probable pronunciation variants *m* for one word to be used in lexicon adaptation, was tested with different values (1, 4, and 8).

4 Results

The word error rate (WER), letter error rate (LER), and the foreign entity name error rate (FENER) are reported in the results. All the results are presented in Table 1.

The first experiment was run on the baseline system. The average WER is 21.7% for general news and 34.0% for sports. The average FENER is significantly higher for both (76.6% and 80.7%).

Supervised vocabulary adaptation was implemented by manually retrieving the foreign entity names from the adaptation text and adding their pronunciation rules to the lexicon. Morpheme adaptation was also applied. Compared to only using linear interpolation (λ = 0.1) supervised vocabulary adaptation reduces WER by 4% (general news) and 6% (sports news). Recognition of foreign entity names is also improved with FENER reductions of 18% and 24%.

Unsupervised vocabulary adaptation was implemented through automatic retrieval and pronunciation generation of foreign entity names. The parameters of interest are the foreign name percentage threshold *T*, determining how many foreign word candidates are included for lexicon and morpheme adaptation and *m*, the number of pronunciation vari-

Adaptation method				Results					
LM	Lexicon			General News			Sports News		
	Adaptation	*T*[%]	*m*	WER[%]	LER[%]	FENER[%]	WER[%]	LER[%]	FENER[%]
Background	Baseline			21.7	5.7	76.6	34.0	11.4	80.7
Background + Adaptation	Baseline			20.6	5.3	67.8	32.0	10.7	69.4
	Supervised	-	1	**19.8**	**5.0**	**55.7**	**30.1**	**9.8**	**53.1**
	Unsupervised	30	1	20.4	5.2	64.0	31.5	10.4	64.1
			4	**20.2**	**5.2**	58.7	31.6	10.4	60.4
			8	20.4	5.3	**56.9**	31.5	10.4	56.8
		60	1	20.7	5.3	63.7	32.3	10.4	63.7
			4	20.7	5.3	59.4	31.1	9.9	59.8
			8	21.1	5.5	58.2	**31.0**	**9.9**	**55.6**
		100	1	21.1	5.4	62.7	33.2	10.7	66.1
			4	21.2	5.5	58.2	32.6	10.4	60.7
			8	22.1	5.9	59.2	33.2	10.6	57.0

Table 1: Results of adaptation experiments on the two test sets. Linear interpolation is tested with supervised and unsupervised vocabulary adaptation. *T* is the top percentage of foreign entity name candidates used in unsupervised vocabulary adaptation, and *m* is the number of pronunciation variants for each word.

ants generated for each word. The best performance is reached on the general news set with $T = 30\%$ and $m = 4$ (WER = 20.2%, FENER = 58.7%), and on the sports news set with $T = 60\%$ and $m = 8$ (WER = 31.0%, FENER = 55.6%).

5 Conclusion and Discussion

In this work we presented an unsupervised approach to pronunciation and language modeling of foreign entity names in morph-based speech recognition.

In the context of LM adaptation, foreign entity name candidates were retrieved from in-domain texts using a foreign word detection algorithm. Pronunciation variants were generated for the foreign word candidates using a grapheme-to-phoneme converter. Morpheme adaptation was also applied by restoring the foreign entity names into their base forms in the morph-segmented adaptation texts.

The results indicate that unsupervised pronunciation and language modeling of foreign entity names is feasible. The unsupervised approach almost matches supervised adaptation in correctly recognizing foreign entity names. Average WER is also very close to the supervised adaptation one despite the increased acoustic confusability when introducing more pronunciation variants. The percentage of foreign word candidates included for adaptation affects performance of the algorithm. Including all words starting in uppercase letters significantly degrades ASR results. The optimal threshold value is dependent on the adaptation text and its foreign word frequency and similarity to the evaluation data.

The composition of likely pronunciations of foreign names by Finnish speakers is not a straightforward task. While the native pronunciation of the name is the favored one, the origin of the name is not always clear, nor the definition of the pronunciation. Additionally, the mapping of the native pronunciation to the phoneme set used by the Finnish ASR system can only be an approximation, as well as the pronunciations that the Finnish speakers are able to produce. In future work we will study new methods to model the pronunciation of the foreign names and perform evaluations also in speech retrieval where the recognition of names have particular importance.

References

B. Ahmed. 2005. *Detection of Foreign Words and Names in Written Text*. Doctoral thesis, Pace University.

M. Bisani and H. Ney. 2008. *Joint-Sequence Models for Grapheme-to-Phoneme Conversion*. Speech Communication, vol. 50, Issue 5, pp. 434-451.

M. Creutz and K. Lagus. 2005. *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.

T. Hirsimäki, J. Pylkkönen, and M. Kurimo 2009. *Importance of High-order N-gram Models in Morph-based Speech Recognition*. IEEE Trans. Audio, Speech and Lang., pp. 724-732, vol. 17.

V. Siivola, T. Hirsimäki and S. Virpioja. 2007. *On Growing and Pruning Kneser-Ney Smoothed N-Gram Models*. IEEE Trans. Audio, Speech and Lang., Vol. 15, No. 5.