Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 60 (2014) 78–87

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Lexicon optimization based on discriminative learning for automatic speech recognition of agglutinative language [☆], [☆☆]

Mijit Ablimit [a], Tatsuya Kawahara [a,*], Askar Hamdulla [b]

[a] *School of Informatics, Kyoto University, Kyoto, Japan*
[b] *Institute of Information Engineering, Xinjiang University, Urumqi, China*

Available online 6 October 2013

Abstract

For automatic speech recognition (ASR) of agglutinative languages, selection of a lexical unit is not obvious. The morpheme unit is usually adopted to ensure sufficient coverage, but many morphemes are short, resulting in weak constraints and possible confusion. We propose a discriminative approach for lexicon optimization that directly contributes to ASR error reduction by taking into account not only linguistic constraints but also acoustic–phonetic confusability. It is based on an evaluation function for each word defined by a set of features and their weights, which are optimized by the difference in word error rates (WERs) between ASR hypotheses obtained by the morpheme-based model and those by the word-based model. Then, word or sub-word entries with higher evaluation scores are selected to be added to the lexicon. We investigate several discriminative models to realize this approach. Specifically, we implement it with support vector machines (SVM), logistic regression (LR) model as well as the simple perceptron algorithm. This approach was successfully applied to an Uyghur large-vocabulary continuous speech recognition system, resulting in a significant reduction of WER with a modest lexicon size and a small out-of-vocabulary rate. The use of SVM for a sub-word lexicon results in the best performance, outperforming the word-based model as well as conventional statistical concatenation approaches. The proposed learning approach is realized in an unsupervised manner because it does not require correct transcription for training data.

Keywords: Speech recognition; Language model; Lexicon; Morpheme; Discriminative learning; Uyghur

1. Introduction

Automatic speech recognition (ASR) systems have been successfully developed for many languages based on the statistical framework of acoustic and language models. With greater accumulation of training data, system performance has been improved so much that they are now used in many practical applications. However, there are still a number of languages for which ASR systems have not been developed due to a small population and limited data availability. Another problem is the agglutinative nature of certain languages. If we simply apply the word-based *N*-gram language model to these languages, the vocabulary size would greatly increase and a reliable language model will not be trained. Uyghur, which we deal with in this paper, is one of the under-resourced agglutinative languages. To our knowledge, there was no ASR system for Uyghur prior to our work.

In agglutinative languages, selection of a lexical unit is an important issue in designing language models for ASR. There is a trade-off between word and morpheme units. Generally the word unit provides better linguistic constraint but greatly increases the vocabulary size, causing out-of-vocabulary (OOV) and data sparseness problems in language modeling. Therefore, the morpheme unit is conventionally adopted in many agglutinative or inflectional languages, including Japanese (Kawahara et al., 2000), Korean (Kiecza et al., 1999; Kwon and Park,

[☆] This work is a substantial extension from Ablimit et al. (2012).

[☆☆] The paper was accepted for publication under the special issue 'Processing Under-Resourced Languages'.

[*] Corresponding author. Address: Sakyo-ku, Kyoto 606-8501, Japan. Tel.: +81 75 753 7484.

E-mail address: kawahara@i.kyoto-u.ac.jp (T. Kawahara).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.09.011

2003; Kwon, 2000), Turkish (Çarkı et al., 2000; Hacioglu et al., 2003; Arisoy et al., 2009, 2012; Sak et al., 2012), German (Geutner, 1995; Berton et al., 1996; Jeff Kuo and Reichl, 1999; Larson et al., 2000; Nußbaum-Thom et al., 2011), Arabic (Xiang et al., 2006; Afify et al., 2006; El-Desoky et al., 2009; Sarikaya et al., 2008), and other languages (Ircing et al., 2001; Jongtaveesataporn et al., 2009; Creutz, 2006; Creutz et al., 2007; Puurula and Kurimo, 2007; Mihajlik et al., 2007). However, the predefined morphemes are short, often consisting of one or two phonemes; thus, they are more likely to be confused in ASR than the word unit. The goal of this study is to incorporate effective word or sub-word entries selectively while maintaining high coverage of the morpheme unit.

A number of studies have addressed this problem, and many are based on statistical measures, such as co-occurrence frequency, mutual information, and likelihood (Saon and Padmanabhan, 2001; Whittaker and Woodland, 2003; Goldsmith, 2001). However, these criteria are not directly related to the word error rate (WER). They take into account statistical characteristics of text but not phonetic similarity and unit length, which are potentially related with confusability in ASR. Moreover, the statistical approaches require a large amount of training data, which are not available in under-resourced languages.

To address this problem, we propose a novel discriminative approach for selecting word or sub-word entries that are likely to reduce the WER (Ablimit et al., 2012). This is realized by aligning and comparing the ASR results obtained by the morpheme-based model with those by the word-based model. We describe each word by a set of features and define an evaluation function with their weights. Then, the weights are learned to select "critical" word entries, which generate different (probably correct) hypotheses from the morpheme-based model. This learning mechanism is applicable to any unseen words or even subwords. In this paper, we investigate several discriminative models including support vector machines (SVM) and logistic regression (LR) model.

The proposed approach is applied to and evaluated in a large-vocabulary Uyghur ASR system. We investigate several features and compared them in terms of WER and lexicon size. Although there have been a number of studies on discriminative learning for language models, such as *N*-gram, (Roark et al., 2007; Collins et al., 2005; Collins, 2002), there have been no prior studies on the use of discriminative learning for lexicon optimization.

The remainder of the paper is organized as follows: We first review the conventional approaches to the lexicon design problem in agglutinative languages in Section 2 and describe the corpus and baseline ASR models for the Uyghur language in Section 3. Next, we present the formulation of the proposed discriminative approach in Section 4. Then, we describe the lexical features and lexicon design in Sections 5 and 6, respectively. We explain the experimental evaluations in Section 7 before concluding the paper in Section 8.

2. Review of lexicon design for agglutinative languages

There have been a number of studies on lexicon design for agglutinative languages (Kawahara et al., 2000; Kiecza et al., 1999; Kwon and Park, 2003; Kwon, 2000; Çarkı et al., 2000; Hacioglu et al., 2003; Arisoy et al., 2009, 2012; Sak et al., 2012; Geutner, 1995; Berton et al., 1996; Jeff Kuo and Reichl, 1999; Larson et al., 2000; Nußbaum-Thom et al., 2011; Xiang et al., 2006; Afify et al., 2006; El-Desoky et al., 2009; Sarikaya et al., 2008; Ircing et al., 2001; Jongtaveesataporn et al., 2009; Creutz, 2006; Creutz et al., 2007; Puurula and Kurimo, 2007; Mihajlik et al., 2007). The majority started with the linguistic morpheme unit, and attempted to concatenate morphemes based on some criteria, which can be classified into linguistic considerations and statistical measures (Saon and Padmanabhan, 2001; Whittaker and Woodland, 2003; Goldsmith, 2001). A naïve statistical method is to concatenate frequent morpheme sequences (FMS). This method counts the morpheme bigram co-occurrence frequency $\boldsymbol{C}(\boldsymbol{m}_i\boldsymbol{m}_j)$ of consecutive morphemes $(\boldsymbol{m}_i\boldsymbol{m}_j)$ and concatenates them if the frequency is higher than a threshold.

Many other statistical measures have been investigated including mutual information and mutual bigram (MBI). Saon and Padmanabhan (2001) compared these measures for Arabic, and found that the MBI performs best. The MBI is calculated as a geometrical mean of forward and reverse bigram probabilities, as below. Here $\mathbf{C}(-)$ represents the occurrence counts of unigram and bigram patterns.

$$MBI(m_im_j) = \sqrt{\mathrm{P_f}(m_i|m_j)\mathrm{P_r}(m_j|m_i)} = \frac{\mathrm{C}(m_im_j)}{\sqrt{\mathrm{C}(m_i)\mathrm{C}(m_j)}} \quad (1)$$

There have also been studies to optimize a lexicon in terms of language model likelihood or perplexity of the training data set (Deligne and Bimbot, 1995; Masataki and Sagisaka, 1996).

Extensive work on lexicon optimization has been conducted for Turkish, which is close to Uyghur by Arisoy et al. (2009, 2012) and Sak et al. (2012). They incorporated a discriminative language model (Roark et al., 2007; Collins et al., 2005; Collins, 2002), which makes optimization in the *N*-gram level. Our research is inspired by the discriminative language model, but we construct a formulation for the problem of lexicon optimization.

There have been investigations on lexicon optimization for other agglutinative or inflectional languages such as Korean (Kiecza et al., 1999), German (Larson et al., 2000; Nußbaum-Thom et al., 2011), Arabic (Xiang et al., 2006; Afify et al., 2006; El-Desoky et al., 2009; Sarikaya et al., 2008), Czech (Ircing et al., 2001), and Thai (Jongtaveesataporn et al., 2009). They are primarily based on the morpheme concatenation approach using the statistical criteria mentioned above or their variations.

The other approach of defining a lexicon for agglutinative languages is unsupervised sub-word or pseudo-