IEICE TRANS. INF. & SYST., VOL.E96–D, NO.11 NOVEMBER 2013

LETTER

Discriminative Approach to Build Hybrid Vocabulary for Conversational Telephone Speech Recognition of Agglutinative Languages

Xin LI[†a)], **Jielin PAN**[†], *Nonmembers*, **Qingwei ZHAO**[†], *Member*, *and* **Yonghong YAN**[†], *Nonmember*

SUMMARY Morphemes, which are obtained from morphological parsing, and statistical sub-words, which are derived from data-driven splitting, are commonly used as the recognition units for speech recognition of agglutinative languages. In this letter, we propose a discriminative approach to select the splitting result, which is more likely to improve the recognizer's performance, for each distinct word type. An objective function which involves the unigram language model (LM) probability and the count of misrecognized phones on the acoustic training data is defined and minimized. After determining the splitting result for each word in the text corpus, we select the frequent units to build a hybrid vocabulary including morphemes and statistical sub-words. Compared to a statistical sub-word based system, the hybrid system achieves 0.8% letter error rates (LERs) reduction on the test set.

key words: *agglutinative languages, speech recognition, sub-words, discriminative learning, hybrid system*

1. Introduction

In agglutinative languages, the word can be formed by concatenating suffixes to the stem successively. Theoretically, the number of distinct word types is unlimited in these languages. Due to this agglutinative nature, the word-based automatic speech recognition (ASR) system with a moderate size vocabulary suffers from the problem of a high out-of-vocabulary (OOV) rate. To increase the vocabulary's coverage, sub-words are utilized as the recognition units in the automatic transcription task for agglutinative languages such as Estonian [1], Hungarian [2], Finnish [3], and Turkish [4]. Morphemes, which are obtained from the morphological parsing, and statistical sub-words, which are derived from the data-driven splitting, are two commonly used sub-lexical units.

For a word type, the result of morphological parsing is usually different from that of data-driven splitting. Since there are many differences between the morpheme vocabulary and the statistical sub-word vocabulary, a hybrid vocabulary can be built to take advantage of individual benefits from them.

Complementarity between words and sub-words has been exploited to improve the recognition performance previously. In [5], words exceeding a certain count in the corpus are left unparsed when building a sub-word vocabulary to increase the LM context. In [6], a discriminative approach is used to select critical words, which may help to reduce the word error rate of a morpheme-based recognizer. These works focus on the combination of words and sub-words, but complementarity of different kinds of sub-words has not been studied.

In this letter, we propose a discriminative framework to build a hybrid vocabulary with morphemes and statistical sub-words. An objective function, which includes the LM probability of the sub-word sequence and the misrecognized phone count of a word on the acoustic training data, is defined and minimized. After the minimization procedure, the optimal splitting result is determined for each word token in the corpus. We collect frequent morphemes and statistical sub-words from the sub-lexical representation of the corpus to generate the hybrid vocabulary.

We evaluate this approach in the Uyghur conversational telephone speech (CTS) transcription task. Uyghur is a typical agglutinative language which is similar to Turkish. Our method is general and ought to be applicable for other agglutinative languages.

This paper is organized as follows. In Sect. 2, we present the details of morphological parsing and data-driven splitting in our work. In Sect. 3, we give the objective function and the optimized algorithm. We describe the acoustic and text data in Sect. 4. Section 5 presents the acoustic modeling and language modeling of the Uyghur ASR system. Experimental results and discussion are given in Sect. 6. We conclude our work in Sect. 7.

2. Word Splitting Methods

2.1 Morphological Parsing

Xerox Finite State Tool (XFST) [7] is used to build a finite state transducer (FST) based morphological parser for Uyghur. There are 97,934 stems and 225 inflectional suffixes in the lexicon of the morphological parser. Morphotactics for nominal, verbal and adjectival inflection is described with "lexc", a high-level language provided by XFST. Orthographic rules including vowel harmony, constant harmony and vowel weakening are described with "twolc" formalism. These two source files are compiled to the lexicon FST and the rule FST. The Uyghur morphological parser is

Manuscript received March 29, 2013.
Manuscript revised July 14, 2013.

[†]The authors are with the Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, China.
a) E-mail: lixin@hccl.ioa.ac.cn.

DOI: 10.1587/transinf.E96.D.2478

Words: kitablirim kitabimning

Morphemes:
Lexical level: kitab -lAr -Hm kitab -Hm -ning
Surface level: kitab -lir -im kitab -im -ning

Statistical sub-words:
kitab -lirim kitab -imning

Fig. 1 Examples of morphological parsing and data-driven splitting.

implemented through the composition operation upon them. Figure 1 gives the morphological parser's output for words "kitablirim" (my books) and "kitabimning" (my book's). In our work, the morpheme sequence in the surface level is selected as the parsing result.

2.2 Data-driven Word Splitting

Baseline–Morfessor algorithm [8] is widely used to split a word into the statistical sub-word sequence. The segmentation algorithm is based on the minimum description length (MDL) principle and requires no expert knowledge of a language. A list of word types with their occurrence count in the text corpus is needed for the data-driven word splitting. For the words "kitablirim" and "kitabimning" mentioned above, the data-driven splitting results are also shown in Fig. 1.

3. Discriminative Framework for Splitting Selection

3.1 Objective Function Definition

For a word token W in the text corpus, the morphological paring result, $SEG_M(W)$, and the data-driven splitting result, $SEG_S(W)$, can be derived through the splitting methods mentioned in Sect. 2. The acoustic training data is decoded with the morpheme based and the statistical sub-word based recognizers. Then the output hypothesis for each utterance is aligned to the corresponding manual label with their phone level edit distance minimized. The number of erroneous phones can be counted for each word W in the reference transcripts. It is written as $PE_M(W)$ for the morpheme system and $PE_S(W)$ for the statistical sub-word system. For a word W in the reference transcripts, the expected number of misrecognized phones is defined as follows,

$$E(W) = \frac{\Pr(SEG_M(W))PE_M(W) + \Pr(SEG_S(W))PE_S(W)}{\Pr(SEG_M(W)) + \Pr(SEG_S(W))} \tag{1}$$

$\Pr(\bullet)$ is the LM probability for the sub-word sequence. In this work, the unigram LM is used and $\Pr(\bullet)$ takes the form as,

$$\Pr(\mu_1, \mu_2 \ldots \mu_m) = \prod_{i=1}^{m} \Pr(\mu_i) = \prod_{i=1}^{m} \frac{f(\mu_i)}{N} \tag{2}$$

for all word types T in the corpus **do**
  select $SEG_M(T)$ or $SEG_S(T)$ as splitting result for T randomly
end for
calculate $Loss(R)$
temp ← $Loss(R)$
for i = 1 2 ... **do**
  for all word types T in the corpus **do**
    $\overline{SEG} = \arg\min_{SEG \in \{SEG_M(T), SEG_S(T)\}} Loss(R)$
    select $\overline{SEG}$ as splitting result for T
  end for
  calculate $Loss(R)$
  if temp - $Loss(R) < \delta$ **then**
    break
  else
    temp ← $Loss(R)$
  end if
end for

Fig. 2 Pseudo-code for the optimization procedure.

where $f(\mu_i)$ is the occurrence count of μ_i in the corpus and N is the total number of the sub-words. The objective function is written as,

$$Loss(R) = \sum_{W \in R} E(W) \tag{3}$$

where R is the whole reference transcripts. The use of unigram is mainly aiming to reduce the computational complexity. When the segmentation of the corpus changes in the optimization procedure, the calculation of the unigram probability is simple and straightforward. It can be seen reduction in $Loss(R)$ means the sub-word sequence with fewer erroneous phones has its unigram LM probability increased. Although adjustment of unigram probability cannot guarantee the optimality for a trigram LM based recognition system, it has relation with higher order n-gram intrinsically.

3.2 Optimization Algorithm

Greedy algorithm is utilized to minimize the objective function. The splitting method for each word type is initially selected randomly. Then for each distinct word type, the morphological parsing and the data-driven splitting are proposed and the one yielding less $Loss(R)$ is selected. After all word types have been processed once, the calculation of $Loss(R)$ is performed. The procedure continues until no significant change of $Loss(R)$ is obtained. Let δ be the threshold for the change of $Loss(R)$, the optimization procedure is illustrated in Fig. 2.

In the procedure of minimizing the objective function, the tree structure is used to organize variables in $Loss(R)$. Figure 3 shows such trees for the word types "kitablirim" and "kitabimning". The child nodes of the word type node represent the sub-word types occurring in its morphological parsing and data-driven splitting results. Each sub-word type has a tree for the word tokens in the reference transcript. If a sub-word type is present in the $SEG_M(W)$ or $SEG_S(W)$ of a word token W, it will be the parent node of this word token. The sub-word type's occurrence count can

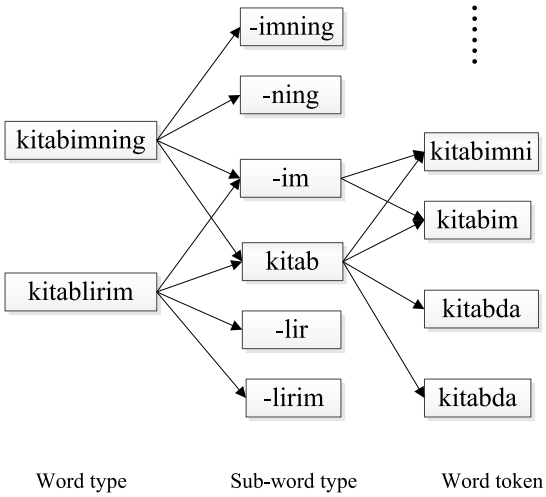

Fig. 3 Examples of the tree structure for two Uyghur word types.

for all sub-words s in current splitting result of word type T **do**
  decrease count(s) by count(T)
for all sub-words s in $SEG_M(T)$ **do**
  increase count(s) by count(T)
$Q = \varnothing$ //Q is a set for word tokens//
Accumulation = 0
for all sub-word nodes s rooted at word type T **do**
  for all word token nodes W rooted at sub-word s **do**
    increase Accumulation by $E(W)$ **if** $W \notin Q$
    $Q = Q \cup \{W\}$
return Accumulation

Fig. 4 Pseudo-code for evaluating morphological parsing.

affect $E(W)$ of its child nodes. The word type node is associated with its occurrence count in the text corpus. The occurrence count of a sub-word node equals the sum of the counts of its parents. The child node of the sub-word type is associated with $PE_M(W)$ and $PE_S(W)$, which are need for the calculation of $E(W)$.

In the optimization procedure, it is important to know the difference in $Loss(R)$ and the absolute value is not crucial. When selecting the sub-word sequence for a word type, occurrence counts of its child nodes, the sub-word type nodes, will change. $E(W)$ of word tokens rooted at these sub-word types need be accumulated and compared. Figure 4 gives the accumulation process when morpheme sequence is proposed as the splitting result. The same process is carried out to evaluate the statistical sub-word sequence. Then the splitting result yielding less accumulation value is selected for this word type and the counts of its sub-words are updated.

4. Speech Database and Text Corpora

4.1 Speech Data

We build a speech database for the Uyghur CTS transcription task. Telephone conversations between close friends or family members are recorded. More than two thousands native speakers are involved and the ratio of female to male is balanced. These conversation recordings are segmented and transcribed manually. 200 hours of this database are used to train models. Separate from the training data, disjoint develop (1 hour) and test (1 hour) sets are used for parameter optimization and final performance evaluation respectively.

4.2 Text Corpus

Reference transcriptions of the speech data mentioned above are considered as the in-domain text corpus for the CTS recognition task. There are 2.13M word tokens in this corpus. A general text corpus is built by collecting sentences from novels, essays, newspapers and webs. There are 27.3M word tokens in the general corpora.

5. Recognition System

5.1 Acoustic Model

The speech data is represented by a stream of 39-dimensional feature vectors, which are produced through heteroscedastic linear discriminant analysis (HLDA) of 52 perceptual linear prediction coefficients (13 base coefficients plus first, second and third derivatives). In Uyghur, the word spelling is almost phonetic, with each letter corresponding to a phone. We split the word into the letter sequence to represent its pronunciation. Speaker independent decision-tree state clustered cross-word triphone models are built in the HMM-GMM framework. There are 6964 HMM states in the acoustic model, and each state has a GMM with 32 mixture components. First maximum likelihood estimation is used in acoustic model training. Then minimum phone error training is proposed to refine the parameters.

5.2 Language Model

We use the SRILM toolkit [9] to build trigram language models. A CTS domain LM is trained on the reference transcripts of the acoustic training data. The general corpus is used to train another LM. The LM used in the recognition system is built by interpolating the CTS domain LM with the general LM linearly. The interpolation weights are determined by minimizing the perplexity on the develop set.

6. Recognition Experiments

6.1 Baseline Systems Results

We evaluate the performance of the word-based and sub-word-based systems on the CTS transcription task. Vocabularies are built by selecting the 65K most frequent units from the text corpus represented in the word or sub-word level. We perform ASR experiments for various recognition units with the same acoustic model. Table 1 shows the letter error rates (LERs) of these systems on the develop set and

Table 1 Results for different language modeling units.

Recognition Units	LERs (%)	
	Develop	Test
Words	39.5	45.7
Morphemes	38.7	44.8
Statistical Sub-words	38.8	44.7

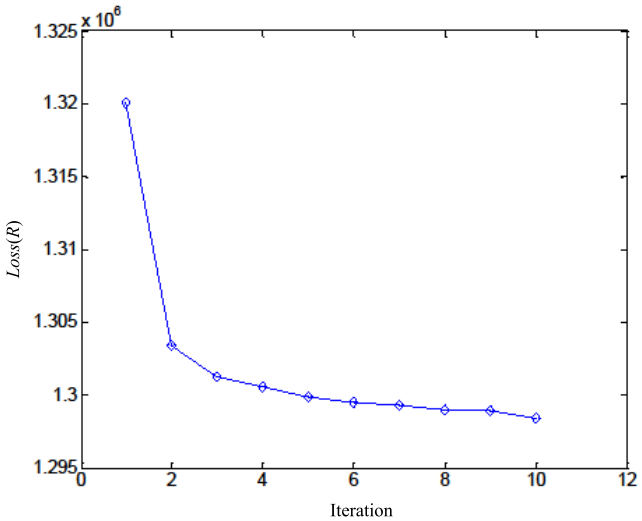

Fig. 5 Values of *Loss*(*R*) after iteration 1-10.

the test set.

Compared to the word-based system, the morpheme-based system and the statistical sub-word system achieve 0.9% and 1.0% LER reduction on the test set respectively. There is no statistically significant difference between these two sub-word systems, although only 21K sub-word types both occur in the morpheme vocabulary and the statistical sub-word vocabulary. These experiment results inspired us to study the complementarity of morphemes and statistical sub-words in speech recognition.

6.2 Minimization of Objective Function

Acoustic training data is decoded with the morpheme-based and the statistical sub-word-based systems mentioned above. Recognized sentences are aligned to the reference transcripts with the phone level edit distance minimized. In the reference transcripts, there are 631,256 word tokens whose misrecognized phone count differs in the two systems. Only these words can affect the value of the objective function.

The process of selecting the optimal splitting method for each word type is carried out as presented in Sect. 3.2. To reduce the amount of calculations, general text corpus is not involved in the splitting selection procedure. Values of the objective function after iterations 1-10 are shown in Fig. 5. It can be seen there is no significant change in *Loss*(*R*) after iteration 6.

Table 2 Results of different combined vocabularies.

Splitting Method for Out-domain Words	LERs (%)	
	Develop	Test
Morphological Parsing	38.0	43.9
Data-driven Splitting	38.3	44.0

6.3 Hybrid Systems Results

The text corpus needs to be segmented into sub-words before building a hybrid vocabulary. Each word token in the reference transcripts is split according to the selected splitting method after iteration 6. The word tokens which cannot be found in the reference transcripts are split all through morphological parsing or data-driven splitting. As a result, we get two different sub-lexical level representations of the text corpus. After collecting the 65K most frequent units from each segmented corpus, we build two different hybrid vocabularies. Two recognition systems are built based on these two hybrid vocabularies seperately. LERs of these two hybrid systems on the develop set and the test set are given in Table 2. Compared to the statistical sub-word system mentioned in 6.1, the best hybrid system achieves 0.8% LER reduction on the test set. This improvement is statistically significant at the level of $p < 0.001$ in the NIST MAPSSWE significance test.

7. Conclusion

In this letter, we propose a discriminative approach to exploit the complementarity between morphemes and statistical sub-words in the automatic transcription task of Uyghur conversational telephone speech. An objective function including unigram LM probability of the sub-word sequence and misrecognized phone count of the word is defined. After minimizing the objective function with a greedy algorithm, the optimal splitting method is found for the in-domain word. The text corpus is converted to the sub-lexical level and the most frequent morphemes and statistical sub-words are selected to build a hybrid vocabulary. Experiment results suggest recognizers based on this hybrid vocabulary outperform the morpheme system and the statistical sub-word system. In future work, we will try higher order n-gram in the objective function.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124, 11074275, 11161140319, 91120001), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), the National 863 Program (No. 2012AA012503) and the CAS Priority Deployment Project (No. KGZD-EW-103-2).

References

[1] A. Puurula and M. Kurimo, "Vocabulary decomposition for Estonian

open vocabulary speech recognition," Proc. 45th Annu. Meeting Assoc. Comput. Linguist., pp.89–95, 2007.

[2] P. Mihajlik, Z. Tüske, B. Tarján, B. Németh, and T. Fegyó, "Improved recognition of spontaneous Hungarian speech—Morphological and acoustic modeling techniques for a less resourced task," IEEE Trans. Audio Speech Language Process., vol.18, no.6, pp.1588–1600, 2010.

[3] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to Finnish," Computer Speech and Language, vol.20, no.4, pp.515–541, 2006.

[4] E. Arısoy, D. Can, S. Parlak, H. Sak, and M. Saraçlar, "Turkish broadcast news transcription and retrieval," IEEE Trans. Audio Speech Language Process., vol.17, no.5, pp.874–883, 2009.

[5] E. Arısoy, H. Dutağacı, and L.M. Arslan, "A unified language model for large vocabulary continuous speech recognition of Turkish," Signal Process., vol.86, no.10, pp.2844–2862, 2006.

[6] M. Ablimit, T. Kawahara, and A. Hamdulla, "Discriminative approach to lexical entry selection for automatic speech recognition of agglutinative language," Proc. ICASSP 2012, pp.5009–5012, 2012.

[7] K.R. Beesley and L. Karttunen, Finite State Morphology, CSLI Publications, Stanford, CA, USA, 2003.

[8] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0," Helsinki University of Technology, Publications in Computer and Information Science, Technical Report A81. 2005.

[9] A. Stolcke, "SRILM—An extensible language modeling toolkit," Proc. ICSLP 2002, vol.2, pp.901–904.