A Phrase Table Filtering Model Based on Binary Classification for Uyghur-Chinese Machine Translation

Chenggang Mi\textsuperscript{1,2}, Yating Yang\textsuperscript{1,\ast}, Xi Zhou\textsuperscript{1}, Lei Wang\textsuperscript{1}, Xiao Li\textsuperscript{1} and Eziz Tursun\textsuperscript{1,2}

\textsuperscript{1}Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences, Urumqi 830011, China
Email: michenggang@gmail.com

\textsuperscript{2}University of Chinese Academy of Sciences, Beijing 100049, China
Email: {yangyt, zhouxi, wanglei, xiaoli}@ms.xjb.ac.cn, ezitzursun@gmail.com

Abstract—In statistical machine translation, large amount of unreasonable phrase pairs in a phrase table can affect the decoding efficiency and the overall translation performance, especially in Uyghur-Chinese machine translation. In this paper, we present a novel phrase table filtering model based on binary classification, which consider differences between Uyghur and Chinese, and draw lessons from binary classification in machine learning. In our model, four features are considered: 1) Difference in length between source and target phrase; 2) Proportion of translated words in phrase pairs; 3) Proportion of symbol words; 4) Average number of co-occurrence words in training corpus. We use this model to generate a filtered phrase table. Experimental results show that this new filtering model can improve the performance and efficiency of our current Uyghur-Chinese machine translation system.

Index Terms—Uyghur-Chinese machine translation; Phrase table filtering; Binary classification

I. INTRODUCTION

Phrase-based model [1] is one of widely used statistical machine translation models [2][3], and phrase table is the most important resource in training of a translation model. As the parallel corpus increases, the phrase table grows exponentially. There are two phases in phrase table construction: word alignment [4] and phrase extraction [5]. Phrase extraction is a process which expanding from intersection of word alignment to its union [6]. The word alignment errors will migrate to phrase extraction, so there will be many unreasonable phrase pairs; the quality of translation will also be affected [7].

Study of Uyghur-Chinese machine translation is still at its early age. Uyghur is an agglutinative language, and words are formed by joining phonetically unchangeable affix morphemes. Frequently, there exists one-to-many alignment in Uyghur-Chinese word alignment; however, the word alignment tools we use today cannot adequately deal with one-to-many and many-to-one word alignment. Otherwise, the quality of phrase table heavily depends on the accuracy of word alignment. The alignment errors will lead to many unreasonable phrase pairs during phrase extraction. Therefore, the filtering to phrase table of Uyghur-Chinese machine translation is very important.

In this paper, we consider the filtering of Uyghur-Chinese phrase table as a binary classification [8] procedure (filtering, not filtering): Using the difference in length between source and target phrase, the proportion of translated words in phrase pairs, the proportion of symbol words and the average number of co-occurrence words in training corpus as four attributes of Naïve Bayes Classifier, the phrase pair filtering or not depends on the result of classification.

II. RELATED WORK

Our research builds on several previous works: Eck’s pruning approach applies the original translation system to a large amount of text and calculates usage statistics for phrase pairs [9]. Chen described an attempt to reduce the phrase table size using additional training data in an intermediate third language; the central idea behind this approach is triangulation [10]. Tomeh presented a complexity-based filtering model [11].

Previous studies mainly focus on the improving of phrase extraction methods and filtering algorithms, but they are inadequate to handle language pairs that have significant morphological differences such as Uyghur-Chinese. The main contribution of this paper is that we investigate features of Uyghur and Chinese and highlight differences between them; we also introduce the classification theory in statistical machine learning into the filtering of Uyghur-Chinese phrase table.

III. CONSTRUCTION OF PHRASE TABLE

Phrase table is the most important resource in phrase-based machine translation model; the generating of reordering rules and decoding both depend on the phrase table heavily. Construction of a phrase table including two stages: word alignment and phrase extraction.

A. Word Alignment

Phrase-based translation model is the most common used statistical machine translation model. Word
alignment is a process that obtains word co-occurrence information from parallel corpus automatically with the help of statistical machine learning models; it is the first step when training a phrase-based model. IBM 1-5 are most famous word alignment models, which originally proposed by Brown. HMM (Hidden Markov Model) alignment model was first presented by Vogel et.al [12]. Och [13] developed the freely available GIZA++ package, which include training programs for IBM models and the HMM model.

The Uyghur-Chinese word alignment can be indicated as follows (Figure.1):

**B. Phrase Extraction**

The word alignment matrix (Figure.2) is generated firstly, and then, phrase pairs (Figure.3) are extracted from parallel corpus: 1) exhaust all possible phrases in source part; 2) find out the corresponding phrases in target part based on the alignment matrix. During the extraction, we should check whether phrase pair satisfied following two restrictions:

1) Position of words in source phrases should be continuous in source sentence
2) Bilingual phrases should be compatible with alignment matrix, which means that according to the alignment matrix, words in source phrase should align to words in target phrase or align to NULL, and vice versa.

Statistical models mainly depend on statistic information of training corpus. Due to the complexity of languages and the limitation of corpus, data sparseness may occur during training of the translation model which causes alignment errors, these errors may migrate to the phrase extraction.

**IV. FEATURE SELECTION**

In order to filter the phrase table, we take the difference in length between source and target phrase (DLP), the proportion of translated words in phrase pairs (PTW), the proportion of symbol words (PSW) and the average number of co-occurrence words (ACW) in training corpus as four features of Naïve Bayes Classifier. These four feature functions can be described as follows:

**A. Difference in Length between Source and Target Phrase (DLP)**

The differences between source phrase length and target phrase length can reflect the phrases correlation in a certain Uyghur-Chinese phrase pair. For example:

For computable of the correlation, and combine it into the filtering model as a feature function simply, we divide the algorithm of this feature function as two parts:

1) If the length of source phrase and length of target phrase are equal, we assign 0 to the feature function.

\[ F_{PLD} = 0, \text{ if } Len_s = Len_t \]  

2) If length of source phrase and length of target phrase are not equal, we assign the absolute value of the difference between the length of source phrase and the length of target phrase to the feature function.

\[ F_{PLD} = |Len_s - Len_t|, \text{ if } Len_s \neq Len_t \]  

Where \( Len_s \) is the length of source phrase, \( Len_t \) is the length of target phrase.

**B. Proportion of Translated Words in Phrase Pairs (PTW)**

In phrase-based Uyghur-Chinese translation model, there may exist two forms of Uyghur words:

For fully considering the relevance of the Uyghur phrase and the Chinese phrase, we divide it as two parts to compute the proportion of translated words in phrase pairs: the precise matching and the fuzzy matching.
Precise matching

We use the Uyghur-Chinese bilingual dictionary during the constructing of Uyghur-Chinese phrase table filtering model. In this paper, we assign a value to the proportion of translated words in phrase pairs according to number of word pairs in current phrase pair appeared in bilingual dictionary, (3):

$$F_{TWR} = \frac{NUM(u \leftrightarrow c)}{Len}$$  (3)

$NUM(u \leftrightarrow c)$ is the number of translated words in a certain phrase pair, $Len$ is the length of source phrase.

There are millions of word-pairs in Uyghur-Chinese bilingual dictionary, reading the dictionary from files during the matching process can be very time consuming. Therefore, we save the Uyghur-Chinese bilingual dictionary as a trie tree to improve the efficiency. Because of the direction of translation is from Uyghur to Chinese, we store characters of Uyghur as edges of the tree; current formed strings as intermediate nodes; complete Uyghur words as leaf nodes. Also, we add an extend node for each leaf node, which store a corresponding Chinese word. Then, we transfer the translated words matching to traversal of a trie tree.

Fuzzy matching

After we analysis of Uyghur-Chinese machine translation phrase table and the result of translation, we found that only a small number of translated words can be matched precisely. Uyghur is an agglutinative language; the word order of Uyghur is Subject-Object-Verb (S-O-V) which is different from languages like English and Chinese. Except that, Uyghur words are formed by joining phonetically unchangeable affix morphemes to the stem. Most Uyghur words have affixes. Because of complexity of Uyghur, the Uyghur-Chinese bilingual dictionary cannot include all forms of a given Uyghur word. We can only get the basic form of a Uyghur word and its translation (Chinese word). These word pair cannot be matched by the precise matching, but they are really important in the phrase table filtering.

For fully use of language resources and considering the word-alignment in bilingual dictionary, we deal with the word co-occurrence information which generated during the training of word alignment, and therefore, there will be some unreasonable translated words with fuzzy matching algorithm. Based on the intersection of Uyghur-Chinese word alignment, we extend node for each leaf node, which store a corresponding Chinese word. Then, we transfer the translated words matching to traversal of a trie tree.

$$time_{ECD}(uP_i)$$ is number of delete operations when computing the edit distance between a Uyghur word $uP_i$ and a Uyghur word $uD_j$ in Uyghur-Chinese bilingual dictionary. We store the Uyghur-Chinese dictionary as a trie tree, so we can compute the proportion of translated words with fuzzy matching algorithm as follows, (5):

$$Score_{TFR} = \frac{NUM(u \times \ldots \times c)}{Len}$$  (5)

$NUM(u \times \ldots \times c)$ is the number of translated words with the fuzzy matching algorithm, and $Len$ is the length of source phrase.

Proportion of translated words

The proportion of translated words is the sum of the precise matching score and the fuzzy matching score, (6):

$$Score_{TWR} = Score_{TWR1} + Score_{TWR2}$$  (6)

$Score_{TWR1}$ is the proportion of precise matching and $Score_{TWR2}$ is the proportion of fuzzy matching.

C. Proportion of Symbol Words (PSW)

After analysis phrase pairs in Uyghur-Chinese phrase table, we found that punctuations and special symbols in the bilingual phrases also indicate the association between Uyghur and Chinese phrase. For example:

In this paper, we consider the proportion of symbol words in phrase table as one of features in our phrase table filtering model. However, there are differences between Uyghur punctuations and Chinese punctuations, for example:

- Uyghur Punc - Chinese Punc :。 - 。 ; , (Uyghur writes from right to left)

For computing the proportion of symbol words, we keep a mapping table from Uyghur symbol words to Chinese symbol words. We can compute the proportion as follows, (7):

$$Score_{SW} = \frac{NUM(Symbol1 \times \ldots \times \rightarrow Symbol1)}{Len}$$  (7)

$NUM(Symbol1 \times \ldots \rightarrow Symbol1)$ is the number of corresponding symbol word pairs in bilingual phrases, and $Len$ is the length of source phrase.

D. Average Number of Co-occurrence Words (ACW)

We extract Uyghur-Chinese phrase pairs from results of word alignment. Based on the intersection of Uyghur-Chinese word alignment, and expand to adjacent cells of aligned words, also, we consider the union of word alignment, and therefore, there will be some unreasonable phrase pairs in Uyghur-Chinese phrase table. For fully illustrate the correspondence between Uyghur words and Chinese words, we deal with the word co-occurrence information which generated during the training of word
alignment model, and compute the average word co-occurrence frequency of every phrase pair in the phrase table, take it as one feature of the phrase table filtering model. The average word co-occurrence frequency can be computed as follows:

\[
\text{Score}_{\text{wc}} = \frac{\sum \text{Co-Count}(u_i, c_j)}{\text{Len}_s}, \quad \text{if } i < \text{Len}_s \text{ and } j < \text{Len}_t
\]

\(\text{Co-Count}(u_i, c_j)\) is the word co-occurrence frequency of Uyghur word \(u_i\) in Uyghur phrase and Chinese word \(c_j\) in Chinese phrase; \(\text{Len}_s\) and \(\text{Len}_t\) are the length of Uyghur phrase and length of Chinese phrase respectively.

V. BINARY CLASSIFICATION-BASED FILTERING MODEL

Due to the complexity of Uyghur and differences between Uyghur and Chinese, we cannot filter the Uyghur-Chinese phrase table according to any one of these four features. Here, we consider the filtering of Uyghur-Chinese phrase table as a binary classification problem. We take features which described in section IV as four attributes for our classifier, the result of classification as the decision of filtering current phrase pair or not.

A. Selection of Classifier

There are three kinds of classification models in statistical machine learning: supervised classification model, semi-supervised classification model and unsupervised classification model. Unsupervised model is mainly used in clustering; semi-supervised model is the combination of unsupervised model and supervised model. With the semi-supervised model, we can use a small number of labeled data and a large number of unlabeled data to train classification models; the supervised model depend on labeled data, which including the Naïve Bayes model, the Maximum Entropy model and the CRFs (Conditional Random Fields) model.

Based on the sampling of phrase pairs in the Uyghur-Chinese phrase table, we found that the four features (Difference in length between source and target phrase (DLP), Proportion of translated words in phrase pairs (PTW), Proportion of symbol words (PSW) and Average number of co-occurrence words (ACW)) in training corpus. The data sets of training corpus are as follows:

\[T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}\]

The feature vector \(x_i\) consists of four elements: \(<D_i, N_i, S_i, A_i>\), \(D_i\) is the difference in length between source and target phrase, \(N_i\) is the proportion of translated words in phrase pairs, \(S_i\) is the proportion of symbol words, and \(A_i\) is the average number of co-occurrence words; the output class label is \(y_i\) also known as \([-1, 1]\), -1 indicates that this phrase pair should be filter, 1 indicates the phrase pair should be kept.

C. Implementation

We implement the phrase table filtering model using Java. The entire system can be divided as three parts: 1) Features extraction; 2) Training of the classification model; 3) Phrase table filtering. (Fig. 4)

Features extraction

Features (Difference in length between source and target phrase, DLP; Proportion of translated words in phrase pairs, PTW; Proportion of symbol words, PSW; Average number of co-occurrence words, ACW) of the classification model can be computed according to algorithms described in section IV.

Training of the classification model

The classical Naïve Bayes algorithm is implemented by Java; we train the Naïve Bayes classifier (the phrase table filtering model) with training data which were annotated by hands. The format of training data is like \(<\text{ph}_{\text{uyg}}, \text{ph}_{\text{chn}}, \text{label}(-1/1)\>\), we compute features of the source phrase \(\text{ph}_{\text{uyg}}\) and the target phrase \(\text{ph}_{\text{chn}}\), firstly.

Phrase table filtering

Given a Uyghur-Chinese phrase pair, the filtering model can make a decision whether this phrase pair is reasonable. Only reasonable phrase pairs will be kept in phrase table.
Training of the classification model

Phrase table filtering

VI. EXPERIMENTS

A. Instruction of Corpus

We validate our phrase table filtering model on CWMT2011 (China Workshop on Machine Translation) evaluation sets, which consist of 50,000 (the training set), 700 (the tune set), 1,122 (the test set) Uyghur-Chinese sentence pairs. There are 640,000 Uyghur-Chinese word pairs in our bilingual dictionary. For counting symbol words in Uyghur-Chinese phrase pair, we maintain a mapping table from Uyghur symbol words to Chinese symbol words. There are about two hundreds symbol word pairs in this mapping table, which are collected from training corpus. We also have 1,000 Uyghur-Chinese phrase pairs, which were annotated with class labels (-1: filtering, 1: not filtering). These corpuses are selected manually.

B. Set up

We use Moses [14] as the baseline system. Language models used in our experiments are 3-gram; all of them are built with the SRILM toolkit [15]. We train word alignment model with GIZA++ and use ictclas4j as the Chinese segmentation tool.

First of all, we preprocess the corpus, such as Chinese segmentation, Uyghur tokenization. We train translation model with Moses and GIZA++, and extract features from Uyghur-Chinese phrase table (section IV). We use the Naïve Bayes classifier which trained on 1,000 Uyghur-Chinese phrase pairs with their labels to filter Uyghur-Chinese phrase pair in phrase table.

For fully validation of our filtering model, we also test our filtering model on the rule table which was generated by hierarchal-phrase based model.

C. Analysis of Results

Phrase extraction and phrase table filtering

We first train the machine translation model using GIZA++ and Moses, then we extract features from Uyghur-Chinese phrase table, finally we filter the phrase table with trained Naïve Bayes classifier. The sizes of phrase tables before and after filtering are performed in Table I.

With our phrase table filtering model, many unreasonable phrase pairs (rule pairs) in phrase-based translation model (hierarchal phrase-based translation model) are filtered. When we assign the maximum length of phrase (rule) 5, 7, 9, 11 respectively, the size of phrase table (rule table) all reduces significantly. There are non-terminals in rule table which is indicated by X, the proportion of symbol words in rule table improved, so the reduction of rule table size is less than phrase table.

Evaluation on the translation performance

We can get more reasonable translation candidates during the decoding with filtering of phrase table, so the quality of translation are all improved. With the help of formal syntax, the hierarchal phrase-based translation model has a strong long-distance reordering ability. (Table II) When the maximum length of rule is 9, the BLEU achieved the highest. Comparing before and after filtering, when the maximum of phrase length assign 9, the largest gain of phrase-based model BLEU achieved 0.0086; when the maximum of rule length assign 9, the largest gain of hierarchal phrase-based translation model BLEU achieved 0.0092. Because of difference between Uyghur and Chinese, when the length of phrase (rule) too small, Uyghur phrase and Chinese phrase couldn’t indicate the same meaning, so the quality of translation was decreased. (n = 5)

Evaluation on the translation efficiency

With the model described in this paper, many unreasonable phrase pairs in phrase table were filtered; the size of Uyghur-Chinese phrase table was significantly reduced. Decoding efficiency was also improved. The decoding time used before and after phrase table filtering is performed in Table III. Many unreasonable phrase pairs (rule pairs) were filtered with our model, which made translation efficiency improved effectively.

VII. CONCLUSION

In this paper, we present a phrase table filtering model for Uyghur-Chinese machine translation based on binary classification. Four features are used in our model: 1) Difference in length between source and target phrase; 2) Proportion of translated words in phrase pairs; 3) Proportion of symbol words and 4) Average number of co-occurrence words in training corpus as four attributes of the binary classifier (Naïve Bayes Classifier), the decision of whether to filter the phrase pair as the result...
of classification. Experimental results show that with the filtering model, the decoding efficiency and the translation performance are both improved. In our future work, we will further investigate the storage of Uyghur-Chinese phrase table and the Uyghur-Chinese dictionary, which have a significant impact on the efficiency of phrase table filtering model.

ACKNOWLEDGMENT

This work is supported by the “Strategic Priority Research Program” of the Chinese Academy of Sciences (XDA06030400), the West Light Foundation of the Chinese Academy of Sciences (XBBS201216), the Key Project of Knowledge Innovation Program of Chinese Academy of Sciences (KGZD-EW-501), Young Creative Sci-Tech Talents Cultivation Project of Xinjiang Uyghur Autonomous Region (2013731021).

REFERENCES


Chenggang Mi received his bachelor’s degree in Xi’an University of Posts and Telecommunications in 2010. He is a Ph.D. candidate in Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences since September 2010. His current research interests include natural language processing and machine translation.

Yating Yang received her Ph.D degree in Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences in 2012. She is an associate researcher in Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences. Her current research interests include multi-lingual information processing and machine translation.

Xi Zhou is an associate professor in Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences. His current research interests include multi-lingual information processing and application software.

Lei Wang is an associate professor in Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences. His current research interests include multi-lingual information processing and application software.

Xiao Li is a professor in Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences. His current research interests include multi-lingual information processing and machine translation.

Eziz Tursun received his master’s degree in Xinjiang University in 2008. He is a Ph.D. candidate in Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences since September 2012. His current research interests include multi-lingual information processing and machine translation.