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

Chenggang Mi[1, 2], Yating Yang[1,*], Xi Zhou[1], Lei Wang[1], Xiao Li[1] and Eziz Tursun[1, 2]
[1]Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences,Urumqi 830011, China
Email: michenggang@gmail.com
[2]University of Chinese Academy of Sciences, Beijing 100049, China
Email: {yangyt, zhouxi, wanglei, xiaoli}@ms.xjb.ac.cn, eziztursun@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 Uygur-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 is 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

*Corresponding author: Yating Yang (Email: yangyt@ms.xjb.ac.cn)

© 2014 ACADEMY PUBLISHER
doi:10.4304/jcp.9.12.2780-2786

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):

Figure 1. Uyghur-Chinese word alignment

B. Phrase Extraction

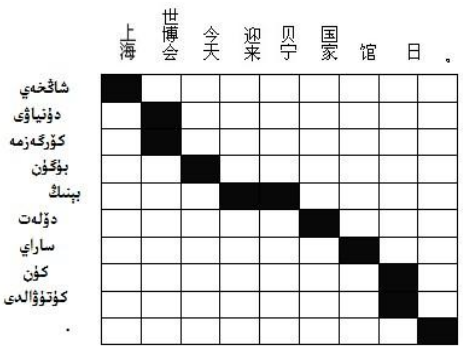

Figure 2. Uyghur-Chinese word alignment matrix

今天 迎来 贝宁 ||| بۇگۈن بېيجىڭ
今天 迎来 贝宁 国家 ||| بۇگۈن بېيجىڭ دۆلەت
今天 迎来 贝宁 国家 馆 ||| بۇگۈن بېيجىڭ دۆلەت سارای
迎来 贝宁 ||| بېيجىڭ
贝宁 ||| بېيجىڭ
迎来 贝宁 国家 ||| بېيجىڭ دۆلەت
迎来 贝宁 国家 馆 ||| بېيجىڭ دۆلەت سارای

Figure 3. Uyghur-Chinese phrase table

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:

上海 世博会 今天 ||| بۇگۈن شاڭخەي دۇنياۋى كۆرگەزمە
（PLD：1）
本届 世博会 开幕 以 ||| بۇيانقى
（PLD：3）

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.

$$F_{PLD} = 0, \text{ if } Len_s = Len_t \quad (1)$$

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.

$$F_{PLD} = \left\| Len_s - Len_t \right\| \text{ if } Len_s \neq Len_t \quad (2)$$

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: ئۇرۇمچى and ئۇرۇمچىنىڭ (“乌鲁木齐” in Chinese), however, most of Uyghur-Chinese dictionaries only include the former (the original form) one. If we use the precise matching, the last one may mismatch. 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.

© 2014 ACADEMY PUBLISHER

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):

$$F_{TWR1} = \frac{NUM(u <=> c)}{Len_s} \quad (3)$$

$NUM(u <=> c)$ is the number of translated words in a certain phrase pair, Len_s 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-formation of Uyghur, we presents a position-related translated words matching algorithm (PTWM), which extend the edit distance algorithm. According to the position information of delete operations with the edit distance, we can decide whether current word formed by added several affixes to a certain stem. That means we can also know whether the current word pair is a translated word or not. The matching algorithm can be described as follows:

$$Sim_{PMED} = \begin{cases} MED_{PMED}(uP_i, uD_j); & no\ continue\ delete\ occur \\ \min\{MED_{PMED}(uP_i, uD_j), ED_{PMED}(uP_i, uD_j) - times_{ECD}(uP_i)\}; & continue\ delete \end{cases} \quad (4)$$

$ED_{PMED}(uP_i, uD_j)$ is the edit distance of a Uyghur word uP_i in phrase table and a Uyghur word uD_j in the Uyghur-Chinese bilingual dictionary; $MED_{PMED}(uP_i, uD_j)$ is the minimum edit distance of above two words; $times_{ECD}(uP_i)$ 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):

$$Score_{TWR2} = \frac{NUM(u < \ldots > c)}{Len_s} \quad (5)$$

$NUM(u < \ldots > c)$ is the number of translated words with the fuzzy matching algorithm, and Len_s 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):

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

$Score_{TWR1}$ 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:

世博会 上 ，每位 游客 ||| دۇنياۋى كۆرگەزمە ، ھەربىر سايياھەتچى

世博会 活动 ，亲身 ||| دۇنياۋى كۆرگەزمە قاتنىشىپ ،

上海 世博会 " ||| « دۇنياۋى كۆرگەزمە شاڭخەي

上海 世博会 ，在华 ||| جۇڭگو ، دۇنياۋى كۆرگەزمە شاڭخەي

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):

$$Score_{SWR} = \frac{NUM(Symbol_s <-> Symbol_t)}{Len_s} \quad (7)$$

$NUM(Symbol_s <-> Symbol_t)$ is the number of corresponding symbol word pairs in bilingual phrases, and Len_s 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

© 2014 ACADEMY PUBLISHER

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:

$$Score_{ACC} = \frac{\sum_{i} Co\text{-}Count(u_i, c_j)}{Len_s} \quad , \quad \text{if } i < Len_s \text{ and } j < Len_t \tag{8}$$

$Co\text{-}Count(u_i, c_j)$ is the word co-occurrence frequency of Uyghur word u_i in Uyghur phrase and Chinese word c_j in Chinese phrase; Len_s and 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) are independent of each other, relatively, so we select the Naïve Bayes as our classification model. The Naïve Bayes model is one of most classic models in machine learning, which has an excellent performance in classification, and widely used in fields like text classification.

B. Construction of Classification Model

The input of Naïve Bayes model is a feature vector $x \in \chi$, the output is a class label $y \in \gamma$. We have four features in our feature vector: 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 $< ph_{uyg}, ph_{chn}, label(-1/1) >$, we compute features of the source phrase ph_{uyg} and the target phrase ph_{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.

© 2014 ACADEMY PUBLISHER

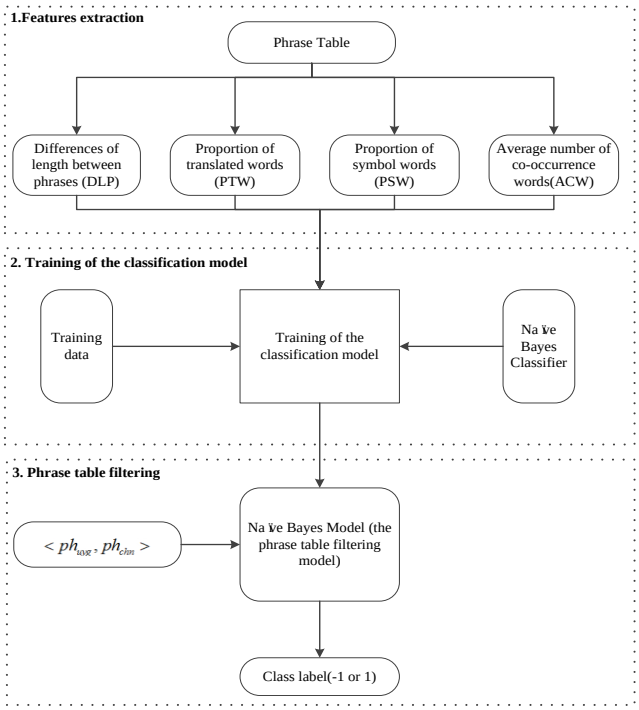

Figure 4. Framework of the phrase table filtering system

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

© 2014 ACADEMY PUBLISHER

of classification. Experimental results show that with the filtering model, the decoding efficiency and the translation performance are both improved. In our future work,

TABLE I
THE IMPACT ON SIZE OF PHRASE TABLE

	Phrase-based model				Hierarchal phrase-based model			
	n=5	n=7	n=9	n=11	n=5	n=7	n=9	n=11
BF	1,515,660	2,177,300	2,738,167	3,212,809	2,839,727	6,627,528	11,543,969	16,977,533
AF	1,143,220	1,824,981	2,140,219	2,590,156	2,019,412	5,834,990	10,594,613	14,349,078

TABLE II
THE COMPARING OF TRANSLATION PERFORMANCE BEFORE AND AFTER PHRASE TABLE FILTERING

	Phrase-based model				Hierarchal phrase-based model			
	n=5	n=7	n=9	n=11	n=5	n=7	n=9	n=11
BF	0.4701	0.4672	0.4493	0.4894	0.4287	0.4380	0.5013	0.4863
AF	0.4698	0.4677	***0.4512***	0.4980	0.4280	0.4401	***0.5105***	0.4870

TABLE III
THE IMPACT ON TRANSLATION EFFICIENCY

	Phrase-based model				Hierarchal phrase-based model			
	n=5	n=7	n=9	n=11	n=5	n=7	n=9	n=11
BF	420s	630s	712s	895s	1,445s	1,570s	1,602s	1,693s
AF	309s	445s	516s	670s	1,125s	1,330s	1,257s	1,590s

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

[1] Philipp Koehn , Franz Josef Och , Daniel Marcu, Statistical phrase-based translation[C]// Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada, 2003: 48-54.

[2] Wang Q, Zhang L, Chang C. Syntactic Function-Based Chinese Lexical Categories and Category Grammar Parsing [J]. Journal of Software, 2014, 9(5): 1270-1274.

[3] Khan M A S, Yamada S, Nishino T. How to Translate Unknown Words for English to Bangla Machine Translation Using Transliteration [J]. Journal of Computers, 2013, 8(5): 1167-1174.

[4] Peter F. Brown , Vincent J. Della Pietra , Stephen A. Della Pietra , Robert L. Mercer, The mathematics of statistical machine translation: parameter estimation[J], Computational Linguistics, 1993, 19(2): 263-311.

[5] Franz Josef Och, Hermann Ney, The Alignment Template Approach to Statistical Machine Translation [J], Computational Linguistics, 2004, 30(4): 417-449.

[6] Zhongjun He, Qun Liu, Shouxun Lin, Partial matching strategy for phrase-based statistical machine translation//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Columbus, Ohio, 2008: 161-164.

[7] Xiong W, Jin Y, Liu Z. Recognizing Chinese Number and Quantifier Prefix to Enhance Statistical Parser in Machine Translation [J]. Journal of Computers, 2014, 9(4): 867-874.

[8] Zhao Shi-qi,Zhao Lin,Liu Ting,Li Sheng. Paraphrase Collocation Extraction Based on Binary Classification [J].Journal of Software, 2010, 21(6):1267-1276. (in Chinese)

[9] Matthias Eck, Stephan Vogel, and Alex Waibel. 2007b. Translation model pruning via usage statistics for statistical machine translation[C]//In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, , Rochester, New York, 2007: 21–24.

[10] Yu Chen, Andreas Eisele, and Martin Kay. Improving statistical machine translation efficiency by triangulation[C]// Proceedings of the Sixth International Conference on Language Resources and Evaluation(LREC'08),Marrakech, Morocco, 2008: 2875-2880.

[11] Nadi Tomeh, Nicola Cancedda, and Marc Dymetman. Complexity-based phrase-table filtering for statistical machine translation[C]//In Proceedings of MT Summit XII, Ottawa, Ontario, Canada, 2009.

[12] Stephan Vogel, Hermann Ney, Christoph Tillmann. HMM-Based Word Alignment in Statistical Translation [C]//International Conference on Computational Linguistics (COLING), Copenhagen, 1996: 836-841.

[13] Franz Josef Och, Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models [J]. Computational Linguistics, 2003, 29(1):19-51.

[14] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard

[15] Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation[C]//Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, 2007: 177-180.

© 2014 ACADEMY PUBLISHER

[16] [A. Stolcke. SRILM -- An Extensible Language Modeling Toolkit[C]// Proceedings of the 7th International Conference on Spoken Language Processing, Denver, 2002: 901-904.

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 reseach 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.

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.

© 2014 ACADEMY PUBLISHER