Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (22): 5230-5234, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.5230.5234

Improved of Phrase Extraction Algorithm in Tibetan and Chinese Statistical Machine Translation

Cao Hui and Dong Xiaofang
Chinese National Institute of Information Technology, Northwest University for Nationalities, Lanzhou, Gansu, China, 730030

Abstract: The extraction of the bilingual phrase is one of the key steps in the phrase-based translation model of Statistical machine translation. Extracting bilingual phrase accurately and sufficiently is the focus of the study. By improving the phrase extraction algorithm get the final phrase translation probability table. There is the situation that a Tibetan word aligned to many Chinese in the word alignment matrix. Using the Och algorithm extracts phrase pairs. When it does not meet Och' condition, adding Tibetan dictionaries information. Comparing results by two methods which is the same size between different linguistic corpus and different sentence pairs of Tibetan-Chinese parallel corpora, the improved will be better in the experiment.

Key words: Statistical machine translation, phrase extraction, translation model, tibetan-Chinese bilingual phrase pairs

INTRODUCTION

Phrase extraction technology is development with the development of the phrase based Statistical Machine Translation (abbr. SMT). Whether the phrase extraction result is good or bad will directly affect the final translation result. Many foreign companies, universities' institutions carry out the SMT's research and achieved very good results based on Och algorithm, such as Microsoft, Systran, Google, IBM, Language Weaver (LW), etc.

Phrase extraction can effectively improve the performance of the phrase based SMT system and become a hot spot. Scholars have proposed many different phrase extraction algorithm: He *et al*.. (2007) improved phrase extraction method based on loose scale, this method relax the constraint of Och phrase extraction; David (2005) proposed the hierarchical phrase Deng *et al*. (2008) mentioned phrase extraction as problem statement for information retrieval process; Vogel (2005) proposed the model of word alignment based on Viterbi not phrase alignment methods; Zhao and Vogel (2005) proposed a phrase extraction algorithm does not need word alignment information, to avoid mistakes due to word alignment error.

PHRASE TRANSLATION MODEL AND TRAINING

SMT model based on phrases, adopt the noisy channel model to decode the target language. The noisy channel model is the procession that obtained the source language by distorting the target language. In order to transform the direction of the source language and target language, you need to use the Bayes rule and draw into the language model of PLM.

Phrase translation model includes the processes as below: The pretreatment of the training corpus, bilingual alignment, word alignment, acquire vocabulary probability tables and phrase extraction and the phrase evaluation. The flow diagram shows in Fig. 1.

Pretreatment: Pretreatment is to get rid of the noise for bilingual parallel sentences, including: Remove empty

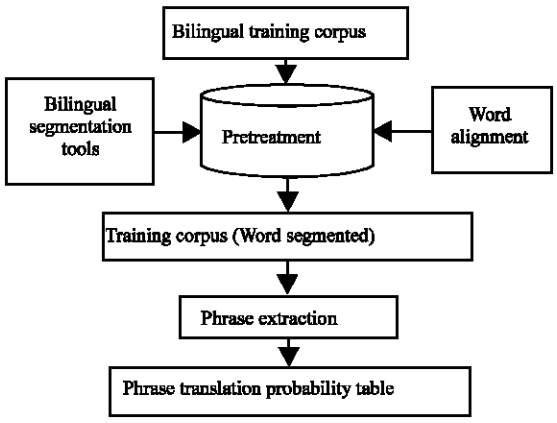

Fig. 1: Training process of Phrase translation model

Corresponding Author: Cao Hui, Chinese National Institute of Information Technology, Northwest University for Nationalities, Lanzhou, Gansu, China, 730030 Tel: 13893694060 Fax: 0931-2977686

lines, redundant spaces and word segmentation for bilingual sentence, etc.

Bilingual sentence alignment: Bilingual GIZA++ is a two-way bilingual alignment tool for bilingual sentences having been pretreated, such as English-Chinese parallel sentences, if translating Chinese into English, the alignment is not only the need for Chinese to English, also need English to Chinese of the alignment. By this tool, we can get word alignment files including bi-direction, Chinese vocabulary table and English vocabulary table.

Word alignment: GIZA++ using IBM's Model can only send a map between multiple source languages to single target language words but can't realize the many-to-many mapping. To achieve a many-to-many mapping, need refining of bilingual sentence pairs. Refining method include insertsect, union, etc. After word alignment, get word alignment matrix.

Obtain vocabulary translation probability table: We calculate the probabilities by co-occurrence frequency based on word alignment matrix, the calculation as (1) and (2) are shown below:

$$p(f \mid e) = \frac{N(f, e)}{\sum_{f'} N(f', e)} \quad (1)$$

$$p(e \mid f) = \frac{N(e, f)}{\sum_{e'} N(e', f)} \quad (2)$$

Among them, N (f, e) and N (e, f) are respectively the number of phrase (f, e) and (e, f) appeared in the corpus.

Then, get the word translation probability by using maximum likelihood estimation. It shows in 3:

$$lex(\bar{f} \mid \bar{e}, a) = \prod_{j=1}^{n} \frac{1}{|\{i \mid (j,i) \in a\}|} \sum_{\forall (i,j) \in a} p(f_j \mid e_i) \quad (3)$$

Usually, p (f | e) and p (e | f) (the phrase translation probability) and lex (f | e, a) and lex (e | f, a) (lexicalization translation probability) are used bi-directional. Using the minimum error method adjust characteristic parameters of the four probabilities together with other characteristics.

Phrase extraction: The phrase refers to a continuous string (n-gram), not necessarily means phrase defined in linguistics. When making the phrase extraction must also ensure that the phrase is "continuous" and bilingual phrase must compatibility with the aligned matrix. Through the phrase extraction algorithm, obtain the pair of phrase don't bring a probability.

Phrases score: Through the phrase score, we can get the pair of phrase with a probability. Extracting phrase score is consistency evaluating to the phrase which use dictionary probability score to evaluate the phrases. And also to evaluate the phrases that have no grammar relations but have transition effect on sentences coherent.

PHRASE EXTRACTION ALGORITHM IMPROVEMENT AND IMPLEMENTATION

Phrase extraction: Phrase Extraction is a process that finds and extracts the source language Phrase corresponding to the target language phrases based on the bilingual sentence alignment. The number and correctness of extracted bilingual phrase will directly affect the correctness of the late model of translation and translation system performance as a whole. It is a key step.

A bilingual sentence is:

f = f1... fm, e = e1... en.

If the source language word fj is corresponding to target word ei, then it is called (j, i) as aligned points, inside 1?j?m,1?i?n. For all connection word sets of (f, e) is called a alignment. The alignment can be expressed into a matrix as A which is m * n order. A (I, j) = 1 when (i, j) is a connection, otherwise A (i, j) = 0.

Bilingual phrase defined in Och is such as 4:

$$BP(f_1^J, e_1^I, A) = \{(f_j^{j+m}, e_i^{i+n}) : \forall (i', j') \in A : j \le j' \le j+m \leftrightarrow i \le i' \le i+n \wedge (i', j') \in A : j \le j' \le j+m \wedge i \le i' \le i+n\} \quad (4)$$

Bilingual translation refers to the bilingual sentence pairs with the corresponding relation of continuous word sequence of binary group. The numbers of the source language words are m and the target language words are n.

There are Two limit conditions of bilingual phrase:

- The original position of the words from the phrase must be continuous in the sentence
- Bilingual phrase must be compatible with aligned matrix. That is to say in aligned matrix, the words in source language phrase either aligned to empty, or the corresponding target language words must be in phrases set and vice versa

For Tibetan-Chinese bilingual sentence, Tibetan language sentence structure is given to SOV (Subject+ Object+ Verb) and Chinese sentence structure is given to SVO(Subject + Verb + Object), such as Fig. 2. shows:

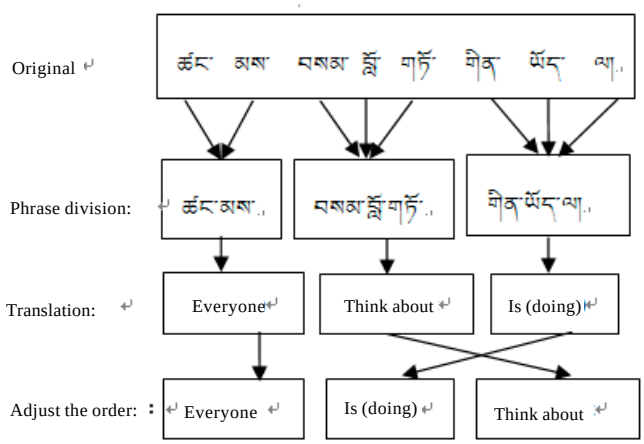

Fig. 2: Example of Tibetan-Chinese phrase translation

ཚང་མས་བསམ་སྟོག་གི་ཨིད་ལ། (Everyone is thinking about.)

Among them, "ལ" are function words, "ཨ" Presents continuous tense, "I" said the end of the sentence.

Och phrase extraction algorithm: First, need to get word alignment sentences; Then, word alignment matrix is obtained. For each source phrase to search the target in the matching phrase and find the minimum and maximum positions the target phrases; Then, in turn, determine the target phrase alignment between the two places to the location of the source phrase is not in the scope of the source phrase. If within the scope of the extraction, otherwise don't extraction. For the target alignment empty phrase boundary extraction under the condition of the phrase is also added to the phrase alignment in the collection.

Och algorithm is the most frequently used in SMT phrase extraction algorithm, this method is to directly using the open source word alignment tool GIZA++ to extract bilingual word alignment. The method is simple and has high accuracy. But since it has two limit conditions and extraction was conducted on the basis of word alignment phrases it also have shortcomings, such as leading to some information lost and low recall rate, higher requirements on the accuracy of word alignment, the difference between particular languages having no take into account, only extracting continuous phrases and no discontinuous phrases.

IMPROVED PHRASE EXTRACTION ALGORITHM

Pretreatment to word alignment: The word obtained through GIZA++ tools training can't be completely aligned because the size of Tibetan-Chinese bilingual corpora hasn't so much and Tibetan sentence is without definite punctuation. This will cause further mistake to decode. So the improvement needs to pretreatment Tibetan-Chinese bilingual alignment corpora before the phrase extraction.

First, obtain the bidirectional word alignment using GIZA++ to Tibetan-Chinese bilingual corpora: From the source language into the target language and also the target language to the source language, Get aligned matrix with points of matrix showing the alignment relations between the Tibetan and Chinese. Word alignment is not necessarily a one-to-one, some words may have multiple alignment points and some may have no one.

In the process of bidirectional word alignment, aligned ambiguity may arise which is the case that positive direction word-pairs aligned and the reverse word-pairs don't. we use a dictionary to optimize for these ambiguity words in this kind of situation. In addition, Tibetan language and Chinese language have different grammar order. There will be a long distance adjustment sequence phenomenon in translation.

Adopt traverse method in checking ambiguity words of Tibetan-Chinese sentence-pairs. Dictionary optimization steps are as follows. If word-pairs having alignment information are not in the dictionary it can reverse check the Chinese words whether align to other Tibetan words or not. If can find the match, then cancel the alignment information; If not, then according to the maximum and minimum range of the Tibetan words to decide Keeping it or not which is two words before and after the word in Tibetan. If in this range, the alignment persists. This process is called the length of the jump method.

Improvement: Och algorithm is an algorithm which has strict qualification, high accuracy and low recall rate with large scale training corpus. But because Tibetan-Chinese SMT research is still in its infancy, Tibetan-Chinese bilingual parallel corpus collected scale and coverage are limited, many linguistic characteristics and knowledge cannot learning through statistical which makes extraction of bilingual phrase on the number and accuracy are not enough, translation results are not perfect and accurate. The basic idea of the improved phrase extraction algorithm is.

First of all, traverse the Tibetan sentences, drawn out the word pairs which are one to more or more to one, from alignment matrix.

Secondly, using Och phrases extraction algorithm extracts word pairs in inner loop. For local continuous phrase in series of discontinuous phrases, Using local continuous phrase extraction method extracts word pairs. Otherwise, add dictionary information.

Finally, sequence the target language sentences s in order considering word order structure.

The description of the improved phrase extraction algorithm is as below Fig. 3.

Tibetan dictionaries include AnDuo Oral Dictionary, Tibetan-Chinese Gussie Dictionary and Lhasa Oral Dictionary. Characteristics of the improved algorithm:

- Considering the continuous phrases exist in the discrete phrase which is extracted by the method of Local continuous phrase extraction
- The improved can extract more phrases than Och algorithm
- The dictionary information is needed when the phrases can't be extracted by mentioned algorithms
- Dictionary query will cost a lot of time which makes the improved algorithm increases the recalling rate of the phrases at the cost of increasing Time complexity. Och's time complexity is $O(n^2)$ and the improved algorithm's is $O(n^3)$

EXPERIMENT AND CONCLUSION

The experimental data:

- The corpus Experiment used are parallel corpora, whose content is the government work report, articles of law, a total of 101629 pairs. The Tibetan corpus' code is Unicode
- Before experiments, the parallel corpus is dealt with segmentation and redundancy (Chen *et al*.., 2003)
- Selecting different scales of sentence pairs
- There are four Chinese reference answers to each test corpus, using Srilm toolkit to train language model

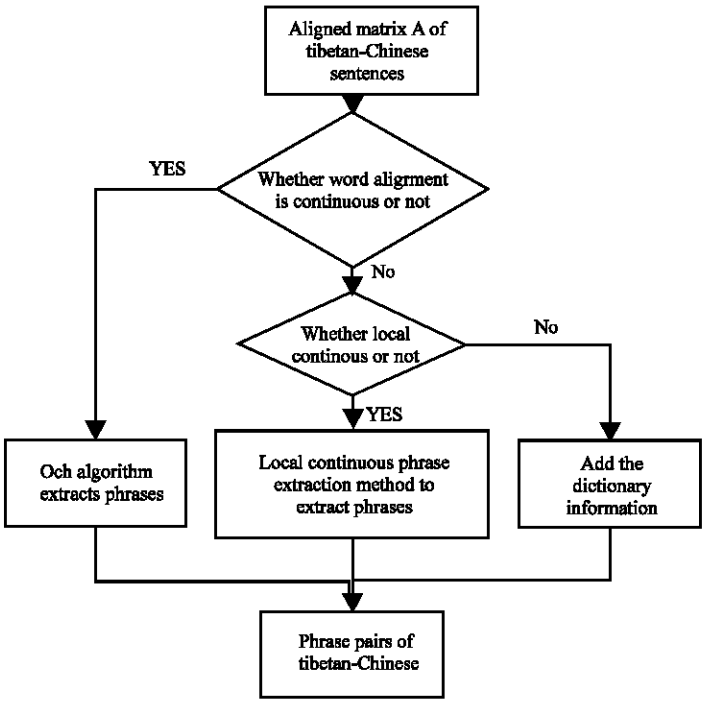

Fig. 3: Description of improved phrase extraction algorithm

Table 1: The comparing results between E-C and T-C

ITEM Languages	Average lengthy of sentences	phrases extracted Och	Improved
E-C	E: 13.9, C: 22.8	59244	68475
T-C	T: 34.7, C: 28.3	68305	83584

Table 2: The comparison in different sentence pairs

Pairs phrases	5000	10000	30000	50000	80000	100000
Och	115679	261528	873053	150454	381546	1216485
Improved	185364	352753	1421451	235641	416544	1918513

- Chinese word segmentation tool adopt the software that is developed by Pattern recognition, State Key Laboratory of Natural Language Processing Research Group
- Tibetan word segmentation software developed by Qi (2006) who is the associate professor in Northwest Universities for Minorities

EXPERIMENT AND RESULT ANALYSIS

Training of translation model: At first, Tibetan-Chinese bilingual parallel corpus is preprocessed, including word segmentation, redundancy, etc. Using word alignment tools GIZA++ get Tibetan-Chinese bilingual alignment corpora and then phrase translation probability table is obtained. Adopting open source tools Moses train the corpus and set up the translation model based on phrases.

Tibetan language encoding method has a lot of kinds, the codes of the corpus should be converted and unified to utf-8 format (Li *et al*.., 2009).

The result of phrase extraction experiment: Make the comparison to the extracted results which are respectively obtained through the improved algorithm and Och algorithm:

- Table 1 shows the comparing results in the same size (3000 sentence pairs) between the English-Chinese corpus(E-C) and Tibetan-Chinese corpus(T-C)
- Table 2 shows the comparing results in different sentence pairs of Tibetan-Chinese parallel corpora

The average lengthy of T-C parallel corpora is longer than that of E-C.

When the number of sentence pairs is equal, the improved can extract more bilingual phrases than Och, no matter E-C or T-C.

For Tibetan parallel corpus of different scales, the improved also can extract more Tibetan-Chinese bilingual phrases.

Based on the phrase of Tibetan-Chinese SMT research need to do more in-depth studies, mainly includes: extended Tibetan-Chinese bilingual parallel corpus scale, expand the covers the areas of parallel corpora, join the lexical and syntactic information, etc.

ACKNOWLEDGMENTS

This study thanks to the support of the 2011 general project of the State Language Commission in the 12th five year plan (YB125-2).

REFERENCES

Chen, Y.Z., B.L. Li and S.W. Yu, 2003. The design and implementation of a Tibetan word segmentation system. J. Chin. Inform., 17: 15-20.

David, C., 2005. A hierarchical phrase-based model for statistical machine translation. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, June 25-30, 2005, USA., pp: 263-270.

Deng, Y.G., J. Xu and Y.Q. Gao, 2008. Phrase table training for precision and recall: What makes a good phrase and a good phrase pair? Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, USA., pp: 81-88.

He, Y., Y. Zhou, C. Zong and X. Wang, 2007. Method of phrase translation extraction based on loose scale. J. Chin. Inform., 21: 91-95.

Li, Y., X.Z. He, J.Y. Ai and H. Yu, 2009. Tibetan encoding and its transformation. Comput. Appl., 29: 2017-2018.

Qi, K.Y., 2006. Information processing in Tibetan word segmentation research. J. Northwest Univ. Nationalities, 4: 92-97.

Vogel, S., 2005. PESA: Phrase pair extraction as sentence splitting. Proceedings of the Machine Translation Summit X, September 14, 2005, Phuket, Thailand, pp: 251-258.

Zhao, B. and S. Vogel, 2005. A generalized alignment-free phrase extraction. Proceedings of the ACL Workshop on Building and Using Parallel Texts, June 29-30, 2005, USA., pp: 141-144.