ISSN : 0974 - 7435 Volume 10 Issue 17

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(17), 2014 [9618-9623]

Research on machine translation based on key technologies of bilingual corpus

Di Lu
[1]Huaibei Normal University, AnhuiHuaibei 235000, (CHINA)
E-mail: lu2069@163.com

ABSTRACT

With the development of the technology of statistical natural language processing, the role of parallel corpus in statistical machine translation and cross-language retrieval cannot be ignored. In this paper, we examines the translation equivalent pairs could be extracted from parallel corpus. An iterative algorithm based on degree of word association is proposed to identify the multiword units for Chinese and English. Then a hypothesis testing approach is used to extract the Chinese English Translation Equivalent Pairs. We present a tree-tree model by mapping between the syntactic tree and the ITG tree, the model limits the reordering of the phrases in the global scope. While in the local scope, the tree-tree model takes the TTG-based local reordering model as one feature, in which the reordering probability of two blocks is decomposed into the product of the reordering probabilities of the child blocks respectively. So the model is able to estimate the reordering of two blocks with arbitrary lengths.

KEYWORDS

Bilingual corpus; Machine translation; ITG tree.

INTRODUCTION

Machine translation is the use of computer simulation of human translation intelligence, the one language (source language) is converted into another language (target language) automated process. In order to achieve this process usually requires cooperation by the linguists and computer scientists, we need to use the special software. This software is generally referred to as machine translation system. machine translation is very obvious advantage that it's old words of memory and function to absorb new words, especially technical terms and phrases fixed sentence memory, making the interpreter can immediately get in when needed, greatly improving the translation speed and efficiency. In some occasion, we only need a rough translation, such as access to information from the machine translation can help us understand the general content of the mouth slogans in a shorter amount of time[1].

Machine translation can promote social progress, a developed translation software can promote economic development, social progress, eliminating the world among people language barriers, communication throughout the world, promoting the exchange, accelerated the process of globalization. Machine translation can improve the efficiency with the continuous improvement of research levels, and translation has also made unprecedented achievements. In our translation work, we can make use of a variety of translation systems or software to complete the translation task, which can improve the speed and accuracy of translation, to improve the efficiency of the translation work. Machine translation problems caused due polysemy mistranslation, which can't accurately identify the part of speech. Depending on context and make appropriate translation machine translation is also an important issue. Semantic ambiguity machine translation often can't flexibly convert passive to active, and therefore translated will appear stiff. In addition, with linguistics, the scientific quality of training and further development of mathematics, machine translation of a computer will be getting better. Machine Translation biggest shortcoming lies in its translation quality finish off, causing some speech-based transformation, polysemy, cultural differences, such as the translation ambiguity, improper word order, structure and poor readability.

Corpus-based machine translation is extracted from the corpus of knowledge related to the translation, so it is crucial to the quality of the corpus, the need for effective corpus finishing and processing. According to language contained in the corpus, it can be divided into: single-and multi-language corpus. With the improvement of computer performance improvement and corpus construction, the past ten years, corpus-based machine translation has gradually become the mainstream[2]. Figure 1 shows the applications of translation in English translation

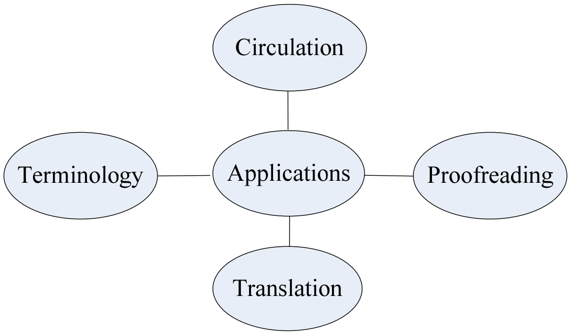

Figure 1 : Applications of translation in English translation

BASIC MODEL OF MACHINE TRANSLATION BASED ON CORPUS

Bilingual parallel corpus of machine translation process using broadly similar pattern, corpus-based bilingual machine translation methods generally use similar translation process, which differs mainly in the way of knowledge and its application translation extracted from the corpus[3]. Figure 2 shows the basic model of machine translation based on corpus.

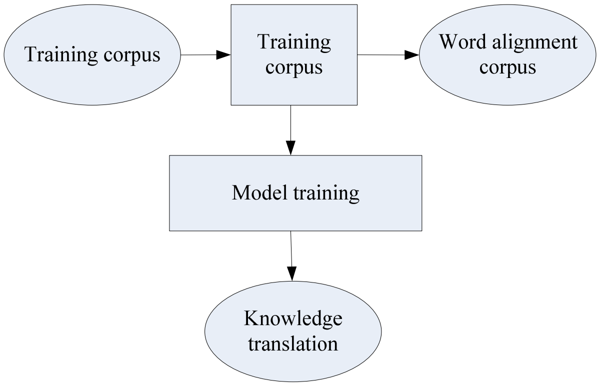

Figure 2 : Basic model of machine translation based on corpus

Given training corpus, on which a variety of training to acquire the appropriate knowledge translation. Typically, the training corpus is sentence aligned parallel corpus. To get a better model would be trained before the model training corpus word alignment process. Depending on the model, the model used in the training process is not the same, but the content is generated within the knowledge contained in the translation varies. The use of translation and other linguistic knowledge, the source text is decoded to generate the corresponding target text. In machine translation mode, linguistic knowledge can be applied separately to the training phase and decoding phase, which contains different types of knowledge, such as POS, syntactic knowledge. While the basic model based machine translation corpus same method, but the translation of knowledge they use in the decoding process is different, and the difference in solving language problems brought to the methods used are different. Therefore, the following will be reviewed typical corpus-based translation model, compare the similarities and differences between them current.

Because purely statistical model can't explain the complex structure of the text, even after the introduction of the phrase, it can only deal with the local context of the word, so the statistical machine translation since the proposed model, in order to improve the quality of machine translation, many researchers improve various improved models to combine linguistic knowledge, in order to explain the complex structural relationships translation of the word. Which uses the largest linguistic knowledge is syntactic knowledge, according to the syntactic forms of knowledge can be divided into two categories: linguistic syntax traditional syntactic knowledge is generally the source or target sentence parsing sentences, then using the syntax tree constraint sentence words. Formal syntax,this syntax is an intermediate grammar, which can simultaneously output in two different language texts. They are mainly on account of two languages with great differences can be difficult to obtain the corresponding syntax, so we have the use of a compromise solution. In order to reduce manual search bilingual corpus difficulty of the work, it is necessary to study an efficient bilingual corpus building programs, and can easily be applied to all areas of research work, as an alternative to the conventional way of artificial acquisition bilingual corpus. In order to solve practical problems provide accurate solutions has a very important practical significance[4].

AUTOMATIC ALIGNMENT OF NOUN PHRASE

Recognition of English noun phrases basis, according to the bilingual word association information to identify the appropriate Chinese noun phrases, resulting in the Chinese-English phrase corresponding method. This method will be dealt with separately in English phrases into high-frequency and low-frequency phrases. For high-frequency phrases in English, according to the bilingual corpus associated information, using an iterative algorithm revaluation statistics, generate and English noun phrases corresponding Chinese words corresponding vocabularies. Forlow-frequency phrases, using the dictionary definition and the use of Dice coefficient similarity of English words and Chinese words, find the corresponding Chinese word string from the corresponding Chinese sentence. After identifying the English noun phrase, the English noun phrase into high-frequency and low-frequency phrases phrase categories, respectively, using different strategies to identify the corresponding Chinese phrases and English phrases. For high-frequency phrases, we can get a more accurate word from the corpus corresponding information: For the low-frequency phrases, because there is no accurate statistics are available, we have some full use of bilingual dictionaries to obtain the corresponding information[5]. Figure 3 shows the flowchart of bilingual corpus.

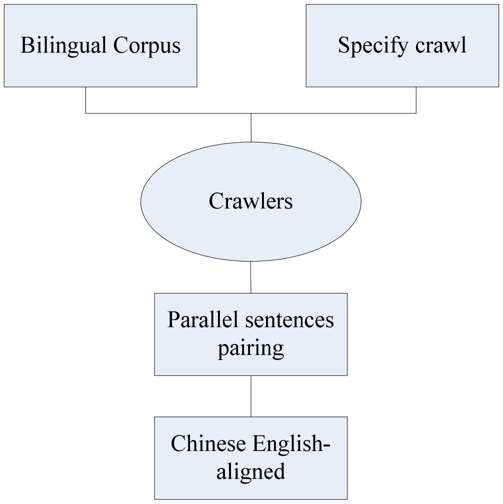

Figure 3 : Flowchart of bilingual corpus

For high-frequency phrases, the first use of statistical methods to obtain each English noun phrases associated with Chinese words, and their degree associations arein the bilingual corpus; then associated information corresponding to determine an English noun phrase in a sentence in accordance with the bilingual aligned Chinese noun phrase. Associate degree in which English phrases and words between Chinese learning algorithm uses an iterative algorithm revaluation. Low collar noun phrase, we take full advantage of bilingual dictionary words corresponding information corresponding phrase. Since the English phrase turned into a Chinese word, the sequence of words may vary, so we first must obtain the appropriate Chinese word order rules in the first term and the tail word, then using low frequency corresponding notional Chinese economic policy from the corresponding sentence find the corresponding terms in order to obtain the corresponding knot soft. Bilingual have a lot structure of long sentences, and translation of the text is long or short translations, proofreading when not easy to accurately match the original translation, proofreading the translation is not conducive to post. The Collaborative Translation Platform provides original and the translation window up and down and left and right control proofreading mode, double-click, you can automatically align the positioning so that the original translation, proofreading is very convenient, can greatly improve the speed of proofreading. In the manmachine cooperation process,there is the circulationbetween the user and the knowledge base of knowledge. Knowledge management platform enables users to tacit knowledge explicit. The translation process through interactive translation system and the translator between the two translation capabilities and interoperability gradually increase, which is a dynamic peer process. In short, not only it can ensure the translation of collaborative translation uniform and accurate results, also greatly can improve the speed of our translation and proofreading

High-frequency identification and the corresponding noun phrases steps are: identification of English noun phrases in the corpus. We can obtain relevant information in English and Chinese words in noun phrases from English bilingual corpus. According to English noun phrases and more relevant information,we determine the alignment of the corresponding Chinese sentence noun phrase. We sequential scan English Corpus, for each English phrase, if it belongs to the high-frequency phrases from the statistical results in the above table to obtain the corresponding Chinese word, they use the word in a sentence with the corresponding Chinese noun phrase may appear string matching, the level corresponds to standard heart to match the length of the longest continuous scale. If the length is the same but different contents match the word string, and calculate the probability of each word corresponds to a string and the largest for the corresponding logo appears. Corpus corresponding low frequency phrases which sequential scan English Corpus, for each low frequency of English phrases, words function calls end to end, and last word will be conducted Lemmatization then constructed based on the low-frequency electronic dictionary notional, resulting in a corresponding Chinese and last word of each sequence[6].

TREE-REE MAP STATISTICAL MACHINE TRANSLATION

Tree-tree of statistical machine translation model integrates formal syntactic knowledge (ITG model) and linguistic syntactic knowledge, considering the reordering of local and global constraints. In the tree - the tree model, the use of a source sentence parsing tree as the basic framework, to generate the corresponding ITG tree, the same basic structure of the two, ITG tree structure will be adjusted according to the specific needs of the build process, in general, these adjustments are fine-tuning. In addition, the model in the basic guarantee source sentence syntactic tree structure at the same time, according toreordering model selection ITG synchronous tree synthesis direction. Therefore, the tree a tree from the global model can be controlled on a pair of phrases, while modeling the local direction of the phrase belongs to the global and local reordering of combining models[7].

Tree-tree model incorporates linguistic syntactic knowledge, and through the use of ITG model, which can be explained by differences in language syntax text. Since the model is able to ITG model to analyze the tree as the basic framework for fine-tuning, therefore, enhances the parsing tree for fault tolerance. In addition, the tree without a tree model extraction rules, to avoid considering the problem of extracting effective rules to improve the ease of use of the model. Translation model for the training, the training is usually obtained by the use of the word in the associated aligned corpus, the phrase extraction using heuristic rule, the frequency of the model calculation of the translation of the latter in accordance with the phrase appears. Because of our training corpus is to meet ITG constraints word alignment corpus, it is possible to build a more simple way translation model. Existing Parallel Corpus still unable to meet the requirements of practical applications, bilingual corpus becomes the bottleneck of the development of statistical machine translation systems and the cross-language information retrieval. The domestic and foreign researchers are paying attention to the further research of bilingual corpus.

Since the tree - the tree translation model combines the ITG model, the decoding process can be viewed as a sequence ITG rules apply, with the constitutive models feature tree structure can guarantee ITG parsing tree and the tree are similar. We give a bundle with the search CYK types of decoders can search for the best tree ITG in the space of all trees. CYK algorithm is a bottom-up analysis of the dynamic programming algorithm, the objects it deals with a Chomsky grammar paradigm (CNF) is. In grammar plus probability, we called probabilistic C YK algorithms. To improve efficiency, the standard CYK algorithm, for each phrase holds a maximum probability translation option, so we can use algorithmic functions to compare the probability of size options. Since the noisy channel model, especially the maximum entropy model, for the machine translation have been proposed, one of the central tasks is to integrate more useful knowledge, especially linguistic knowledge, to improve the translation quality further. This paper focuses on the machine translation between the Chinese-English texts. And we make an in-depth and systematical research on how to incorporate the syntactic knowledge into the bilingual corpus-based machine translation and implement a complete system in the end[8].

MACHINE TRANSLATION SYSTEM

In statistical machine translation model mix based machine translation can only use in the decoding process instance analogy fragments generated translation, while the translation fragments generated using statistical models to evaluate. Examples of the use of the appropriate translation template is generated for the target word ordering constraints "These templates are obtained in the decoding process, that is placed in the decoding process of training, the training phase without extracting the template, and therefore a simple realization of the system; Moreover, examples of the itself contains the syntactic structure of the target sentence, so you can find similar examples in the case, to obtain a smooth translation[9].

Decoding is similar instances at a given set of circumstances, to construct the various translations clips, the process of obtaining the target sentences in different combinations, the process is divided into two steps: l) Matching: According called translation template, referred to as a template. They are in the middle of the translation piece, which will use these various models of fragments were evaluated; 2) a combination of Merging: a combination of two instances of a template to form a bigger match, circulating the above process until the completion of the translation, we select the best translation. English literature, science and technology have a lot of jargon, requires a lot of specialized vocabulary translator reserves, which for non-professional translators areas is a huge challenge in terms of collaborative translation platform provides us with a lot of jargon resources; only need to click the mouse you can get the term, you can use these resources to quickly translate this term savings terminology recognition and retrieval time; and translation of the entire document, the term repeated reminders for consistency, so that a unified terminology, accurately grasp the translation result. In addition to the terms of scientific literature, there are a large number of repeated occurrences of words and sentences. Collaborative Translation Platform provides bilingual interactive translation mode, you can always see all the terminology and reference information for the current sentence translated to complete the confirmation by the mouse or keyboard shortcuts, modify work; collaborative platform provides translation memory function can be repeated intelligent matching of translation can save on repeated re-sentence translation, greatly improving the speed and quality of translation. Figure 4 shows the decoding structure based on instances.

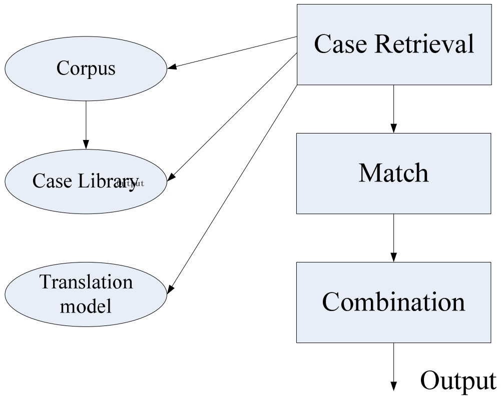

Figure 4 : Decoding structure based on instances

System is divided into three levels: Training layer data layer and the application layer. Training is training and building the interface layer is the developer of the corpus finishing and translation model parameters. Parameters Training: First extraction block, then build a translation model and reordering model, and finally establish a double index structure corpus. Model training: the right value of each translation model for training, mainly on the MER method of implementation, here draws on the realization of Moses MERT module. Other modules: Other tools, such as language training module comprises a model, we directly use the tool as a result of its language model. Data layer contains all kinds of corpus, the model parameters, such as the translation model parameters which correspond to the result of the training modules. The application layer is the interface with the end users. Receiving user input, i.e., to be translated Chinese sentence, translation output, if necessary, one or more of the output result. We also made comparison between different statistical association measurement and proposed to use categorical hypothesis to improve the performance of extraction[10]. Figure 5 shows themachine translation system.

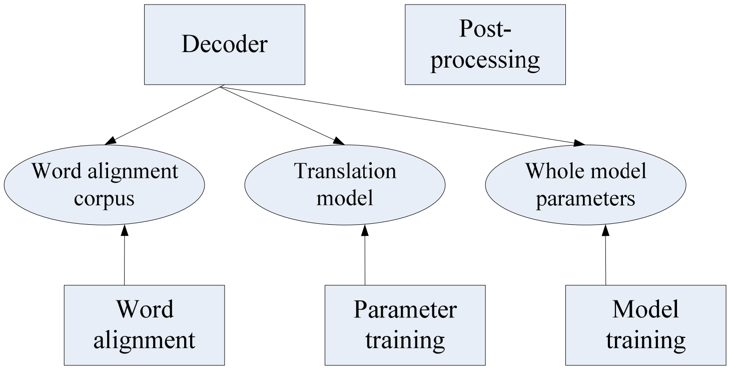

Figure 5 : Machine translation system

CONCLUSION

More and more researchers have recognized the potential value of the parallel corpus in the research on Machine Translation. In this paper, we present a related machine translation model, which expands the tree-tree model, combining the example knowledge to ensure the translation's fluency and consistency. In the same time, we present an example-based decoder, which makes use of both of the knowledge within the translation examples and the statistical knowledge, to improve the quality of translation. With the rapid development of the computer technologies, and the improvement of the corpus construction, the machine translation based on the statistical knowledge becomes possible, and the quality of translation has the chance to get closer to the expectation of human beings. Bilingual corpus has important research value in machine translation, cross-language information retrieval, bilingual dictionary automatically generated research. The model can provide a very important practical significance for the development of related research.

ACKNOWLEDGMENTS

The authors wish to thank the National Natural Science Foundation of China for contract 61100194, the foundation of education department of Zhejiang province of China for contract Y201120520 and the research foundation of Hangzhou Dianzi University for contract KYS105612008 and YB1205, under which the present work was possible.

REFERENCES

[1] Colin Cherry, Dekang Lin; *A comparison of syntactically motivated word alignment spaces, EACL-2006: 1l th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, April,* **3-7**, 145-152, **(2006)**.

[2] Kemal Oflazer, Durgar EIKahlout; *Exploring Different Representation Unit in EnglishtoTurkish Statistical Machine Translation[A], In Proceedings of the Second ACL workshop on StatisticalMachine Translation[C], pages, Prague, Czech*, 25-32 **(2007)**.

[3] Masaaki Nagata, Kuniko Saito; *A clustered global phrase reordering model for statistical machine translation, Proceedings of the 21st International Conference or Computational Linguistics and 44th Annual Meeting of the ACL, pages* **713-720**, 2006 **(2006)**.

[4] Kemal Oflazer; *Statistical machine translation into morphologically complex language[A], Alexander F.Gelbukh: computational linguistics and intelligent text processing, 9th international conferece[C], CICLing 2008, Proceedings.Lecture Notes in Computer Science (4919), Springer, Haifa,Israel.*, 376-387 **(2008)**.

[5] Ren Feiliang, Zhang Li, Hu Minghan, Yao Tianshun; *EBMT based onfinite automata state transfer generation, TMI-2007: Proceedings of the 11th International Conference or Theoreticaland Methodological Issues in Machine Translation, Skovde (Sweden), September*, **7-9** ,65-74 **(2007)**.

[6] Shuyi Zheng, Pavel Dmitriev, C. Lee Giles; *Graph based crawler seed selection [C], Proceedings of the 18th international conference on World wide web, Madrid, Spain,* **1089-1090, (2009)**.

[7] F.Huang, Y.Zhang, S.Vogel; *Mining key phrase translations from web corpora[C], In: Proceedings of the Conference on Hum an Language Technology and Empirical Methods in Natural Language Processing*, 483-490 **(2009)**.

[8] R.Zhang, C.Wu; *A hybrid approach to large-scale job shop scheduling[J], Applied Intelligence,* **32(3)**, 47-59 **(2010)**.

[9] M.R.Sierra, R.Varela; *Pruning by dominance in best-first search for the job shop Scheduling problem with total flow time [J], Journal of Intelligent Manufacturing,* **21(1)**, 111-11 **(2010)**.

[10] Nicola Ferro, Carol Peters; *CLEF 2009 Ad hoc track overview: TEL& persian tasks[M], Text Retrieval Experiments Lecture Notes in Computer Science,***6241**, 13-35 **(2010)**.