Advances in Economics, Business and Management Research, volume 49
2nd International Conference on Education, Management Science and Economics (ICEMSE 2017)

The Application of Paraphrasing Technology of Machine Translation in the Construction of Corpus

Wang Jing
Department of translation, College of foreign languages
Zhejiang International Studies University
Hangzhou 310023, China
zxdningbo@126.com

***Abstract*—At present, with the updating of knowledge and information, bilingual libraries are unlikely to contain all linguistic phenomena. In the course of Machine Translation, it is impossible to process unknown texts that are not present in the corpus. With paraphrasing technology, the unknown text fragments can be translated into idioms with relatively close meaning. The goal of this paper is to combine multiple resources to improve the accuracy and comprehensive performance of paraphrasing technology. This can solve the problem of data sparseness and make the expression more diversified. When bilingual corpus is enriched, the expressions are diversified; the accuracy of machine translation is improved.**

Keywords—Machine Translation; paraphrasing technology; linguistic; bilingual corpus; multiple resources

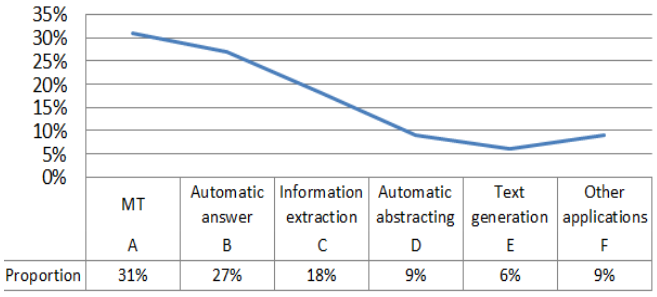
I. INTRODUCTION

The Machine Translation system is the transformation of a natural language into another form of natural language using computer programs. The field of research belongs to computational linguistics, and the main research direction in this field is based on mathematical statistics. The core idea is to establish a statistical model of translation through statistics and analysis of a large number of bilingual parallel corpora. By using the statistical model to decode and translate, the most popular statistical model is the log linear model.

The aim of Machine Translation is to make a mathematical statistics of large-scale bilingual corpus and extract the rules of text translation. But these rules often only deal with literal vertical translation, and do not have the ability of intelligent translation. But with the updating of knowledge, no bilingual corpus can exist, including all language phenomena. This will not be the translation of unknown text is not stored in the corpus, but paraphrases technique can solve the unknown text conversion relative synonymous idioms in the corpus of text fragments, paraphrasing technology is synonymous with the text monolingual expression form conversion. The function of Machine Translation system is to realize the conversion of the form of cross language text fragment expression, so the paraphrases technology is closely related to Machine Translation. Paraphrases technology is developed with the development of various specific technologies of Natural Language Processing. Between 2006 and 2009, Chinese scholars, such as Liu Ting and Zhao Shiqi, began to pay attention to the application of paraphrases technology. In 80s, the famous linguist Halliday and DeBeaugrande defines repeat: "approximate equivalence" concept, Brarzilay to repeat the same information technology as alternative forms of expression. Glickman and others argue that rehearsal techniques exactly reflect the core form characteristics of linguistic diversity. Rehearsal technique is the equivalent expression of the same meaning. To sum up, retelling is the exchange of language fragments with the same semantics and different expressions in a language, which fully reflects the diversity and flexibility of language. Therefore, paraphrases technology can also provide a powerful support for the study of natural language computing.

II. RESEARCH CONTENTS

The main contents of this study based on the statistical model of paraphrase generation technology and application technology in the field of paraphrase generation Natural Language Processing including Machine Translation, automatic answering, information extraction, automatic text summarization, text generation and other related direction. At present, rehearsal generation technology has the largest proportion in the field of Machine Translation applications. According to statistics, the proportion of applications in various fields is shown in Fig. 1.



	MT	Automatic answer	Information extraction	Automatic abstracting	Text generation	Other applications
	A	B	C	D	E	F
Proportion	31%	27%	18%	9%	6%	9%

Fig. 1. The proportion of rehearsal technology in Natural Language Processing's various applications

The core idea of text rehearsal generation includes.

(1)The system needs to establish a unified framework for text data statistical model, which can generate repetitive text fragments for different application tasks.

(2)The system needs to integrate all kinds of available resources so that it can improve the accuracy and matching performance of paraphrase generation. The paraphrase generation technology not only solves the sparse problem of

text data, but also makes the generated repeats rich in expressions.

(3)Wealth of text resources is obtained by using a pivot based method and a repeated generation method of multiple Machine Translation engines. The method first uses the multi pivot method to obtain candidate text fragments for the source language sentence, and then uses the selection and decoding techniques to generate the corresponding rehearsal fragments for the source language sentences. This method can also be used to switch from one application domain to another application field according to the application target, and generate abundant rehearsal fragments. The multi pivot method can obtain a large number of high-quality candidate retelling texts simply and efficiently, which can further improve the performance of rehearsal generation[1,2].

III. KEY PROBLEM

This research needs to solve three aspects of the problem:

The first is to establish methods and standards for evaluating the text to be obtained. At present, the quality evaluation of the automatically generated retelling text still relies on manual judgment, which is an important problem that needs to be solved;

The second problem that needs to be solved is how to solve the quality rehearsal resource acquisition so as to improve the quality of rehearsal. High quality Machine Translation is dependent on rich repetition resources, but it is usually long time to acquire rich quality retelling resources. This problem often becomes a bottleneck in the development of Machine Translation.

The third problem is to solve the problem of text reproduction based on statistics with multi task. The application of rehearsal technique can include sentence compression, sentence simplification and auxiliary sentence similarity calculation.

The experiments in this study can be performed using a Machine Translation decoder, the most famous one in the statistical Machine Translation research is the “MOSE” decoder. In the initial stage of the system work, the source sentence should be pretreated first. This process requires the tagging tools (SVMTool) and the syntactic analyzer (MSTParser) respectively. To use the English Gigaword system in language modeling process (English Gigword LDC of the United States to establish the corpus, it includes more than 1200 megabytes of words), which is the source of a variety of related news media, the library is expected to conduct training on the text using N-Gram language model. In order to be able to use the learning management system, system configuration of HTK (Hidden Markov Model Toolkit, hidden Markov model in the field of artificial intelligence has many successful applications, such as speech recognition, speech recognition system is the current international mainstream based on Hidden Markov model.). A dictionary corpus training set corresponding to the speech recognition system is also needed. The system uses the results of machine translation and manual reference to compare the experimental data, providing several artificial reference sentences for each test sentence. These reference meanings are relatively close. By comparing and calculating the similarity between the automatically generated text and the artificial version, and then, the text with similar similarity is selected according to the result of the calculation [3,4].

In order to improve the usability of the system, it is necessary to design a manual evaluation criterion for statistical rehearsal generation, which mainly includes evaluation of correctness, fluency and applicability". However, it is not enough to measure the performance of the paraphrase generation system, so it is necessary to design an automatic aided test system, which can reflect the quality of paraphrase generation from different perspectives. One parameter is the proportion of the test sentence to be repeated, which reflects the generality of the system. Another parameter is the number of units to be replaced by the rehearsal, which reflects the ability to repeat the system. If the source text is replaced by a larger number of units, the paraphrase phrase is more abundant. In general, if the system generates too many repeat substitutions, it may introduce errors or affect fluency, so it is necessary to weigh the accuracy and fluency of the repetition.

The system needs to design different paraphrase phrase extraction methods for each corpus, and can extract different scale and different types of paraphrase phrase list according to different application goals. These generated paraphrase phrase tables should be used for subsequent statistical paraphrase generation. In order to improve the flexibility and low coverage rate of paraphrase phrases, the system uses a large-scale (English-Chinese) bilingual parallel corpus, and uses the pivot method to extract rehearsal templates. The system uses this method to extract more than 10 thousand pairs, with an accuracy rate of around 70%.

Paraphrase technology can improve the performance of Machine Translation systems in many ways. Based on the rehearsal generation technique, the input sentences of MT (Machine, Translation) systems can be rewritten, and the sentences generated by repetition can be translated more easily. Especially for irregular grammar, spoken sentence translation is more effective. This method can effectively reduce the difficulty of translation system processing. Rehearsal technique can also alleviate the data sparse problem of statistical Machine Translation system to a certain extent. For example, when the phrase S1 needs to be translated does not exist in the MTS (Machine, Translation, System) training corpus, it will be impossible for S1 to be translated. However, if the translation F2 corresponding to the paraphrase S2 of S1 exists in the training corpus, the F2 output can be used as the translation result of S1, so that the data sparseness problem can be effectively solved. Paraphrase technique can also be applied to the automatic evaluation of MT. The basic principle of MT automatic evaluation method, such as BLEU standard, is to compare the similarity between the generated translation T and the reference translation C. The greater the similarity is, the higher the score of the translation T is. In this comparison process, if the T and C can be identified in some literal expression of different forms, but the same meaning of the repeat fragments, which shows that the similarity between the two is very high. Paraphrase technique can also be applied to

the modification and adjustment of work parameters in MT systems[5,6,7].

IV. RETRIEVAL OF REHEARSAL RESOURCES BASED ON MULTI PIVOT

Before discussing how to obtain candidate paraphrase generation based on multi pivot method, a single pivot paraphrase generation method is described here. The system can first use a Machine Translation engine TE1 (Baidu translate, Youdao translation or Google translator, etc.). The input sentence source statement S translation as the pivot language PL (any kind of language is different from the corresponding source statement language), then use Machine Translation search engine TE2 (any kind of translation engine is different from that of TE1) will sentence re translated back to the source language text fragments, and repeat language segment S', S' expression with the source language. A single pivot rehearsal generation system can be described as a three tuple <MT1, PL, MT2> system. TE1 represents a translation engine that converts source language into some axis language PL, and TE2 represents a translation engine that is different from TE1[8].

A multi axis paraphrase generation system is a complex system consisting of several single axis repeat generation systems. Each single pivot system in the system can use different pivot languages and Machine Translation engines. As shown in Fig 2, the multiple pivot method is used to obtain the candidate paraphrase process, which is assumed to be composed of N translation engines, and assumes that each engine can handle M languages. The system can input any source language fragment S through any of the N inputs, and then obtain the representation of the target language fragment T from any of the N output engines[9,10].

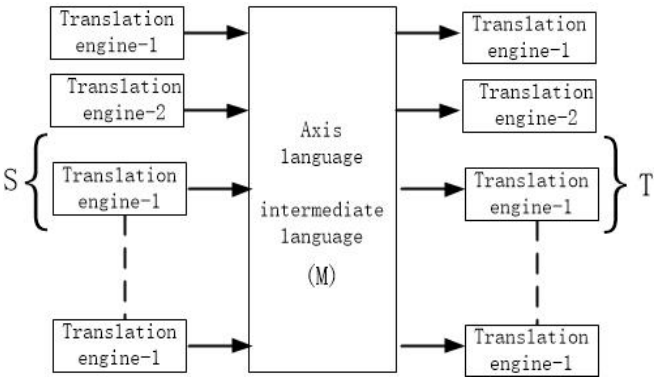


Fig. 2. The paraphrase generation principle based on the multi pivot method

The multiple pivot based paraphrase generation method has the following two advantages. First of all, this method can effectively utilize the huge bilingual resources and translation rules of Machine Translation engine. It can easily obtain rich and high-quality paraphrase resources. Second, the method is very simple and takes full advantage of the online translation engine on the current network. The system only needs to send the input sentences and the sentences of the generated pivot language to the corresponding translation engine and get the corresponding results, without any additional corpus resources or processing tools at all.

Since the hypothetical system includes M pivot languages and N Machine Translation engines, the system is a multi-pivot paraphrase generation system consisting of single pivot systems. Suppose this N Machine Translation engine can achieve the bidirectional conversion between source language and pivot language, theoretically obtained N=n*m*n different repeat expression, so the system can increase the expression of paraphrase generation, have great help to solve the problem of the scarcity of text resources.

Nothing can be perfect. Therefore, these technologies and methods also have an unavoidable defect. The deficiency of the system is that it can only generate (select) the result already existed in the candidate paraphrase set. If every single axis system produces some flaws in the results, even if it is obvious error, these methods cannot intelligently generate a correct paraphrase text. In order to solve this problem, we must use the decoding based paraphrase generation technique and the selection based paraphrase generation method to optimize the generated repeating sentence fragments. The use of a decoding technology optimization paraphrase generation based on the results of the method still need to be decoded by the decoder used in Machine Translation, and the different translation system is generated by multi candidate paraphrases pivot system set to train a translation model[11].

V. TECHNOLOGY ROADMAP

The statistical rehearsal generation method consists of three main steps. That is, (1) pretreatment of source language sentences; (2) rehearsal goal planning; (3) target language rehearsal generation. The first step is the sentence preprocessing stage, which is mainly responsible for the tagging of the input sentence S, and the syntactic analysis of the dependency of each component. In the subsequent rehearsal, goal planning and rehearsal generation, the input statement is matched, followed by the paraphrase template and the rehearsal collocation. This preprocessing phase requires the use of sentences, parts of speech, and dependencies. The second step is the planning of rehearsal generation technology. According to repeat target uses to determine the choice of input sentence in S can be repeat units from all kinds of resources in the repeat (called "source unit"), a plurality of candidate repeat unit corresponding to each source unit (called "target unit"). Finally, in the rehearsal generation stage, an optimal target cell is selected for each source unit based on a statistical model. The target cell is generated from the text of the source statement to reproduce the text[12,13].

Statistical rehearsal generation requires a large number of rehearsal resources with different particle sizes, including rehearsal phrases, rehearsal templates and rehearsal collocations. Retelling the planning process involves several steps. A) system will use various resources to repeat are stored into the list for a repeat, each repeat record in the table of the corresponding record repeat unit (source element and target element) and repeat similarity scoring parameters; B) a system that extracts all the resources in an input sentence that can be used in a rehearsal table; C) system according to the source unit

obtain the candidate information table in all repeat extracted as target unit; D) system according to the application of task specific, who can meet the requirements of the source element and target element data will be retained. These steps are the rehearsal of the planning process.

Method of generating retell the text of the statement is similar to Machine Translation, repeat the sentence generation process is actually a process of decoding, when the input of a sentence S, the system will be decomposed into I unit sequence, then these units are repeated into another sequence consisting of I units.

The so-called language model is a computational model which is widely used in Machine Translation and information retrieval research, and the system can use this model to measure whether the paraphrase sentence is coherent and reasonable.

The parameter estimation of the system refers to the model parameters which can make the training samples get the minimum error rate. That is, the weights of the template, the weight of the language model and the applicability, the weights in the model and so on.

VI. Summary and Outlook

With the development of Machine Translation, rehearsal generation technology is becoming more and more important in natural language. Since the quality of Machine Translation is largely dependent on the quality of corpus, rehearsal technique can well solve the problem of sparse corpus data. At the same time, rehearsal generation technology can also play an important role in automatic question answering, automatic abstracting, information extraction, and the generation of natural language. Therefore, the technology of resource acquisition and rehearsal generation for repetition has gradually become a new research focus in the field of Natural Language Processing. Many universities and research units are being carried out on corpus construction, if the paraphrase generation technology makes full use of, it is bound to preprocessing of the corpus resources play a great help, can save a lot of time and manpower. The transformation of achievements in this direction will have considerable economic benefits. Another important application of paraphrase is abstraction extraction, which can be applied to extract interesting information from sea data, which has great commercial value. The application of repetition in the direction of automatic answering questions also has enormous commercial potential. Due to the automatic answer question, the coverage content range is relatively fixed. So the use of special training in text resources, can provide accurate answer text fragments, can be very good to meet customers various sales or service industry demand, save manpower cost, but also has broad application prospects.

References

[1] Li Maoxi, Zong Chengqing. Machine Translation [J]. Chinese information system integration technology review journal, 2010, 24 (4): 74-81.

[2] Du J, Way A. An incremental three-pass system combination framework by combining multiple hypothesis alignment methods[J].International Journal on Asian Language Processing, 2012, 20(1): 1-16.

[3] A. Fujita and S. Sato. A Probabilistic Model for Measuring Grammaticality and Similarity of Automatically Generated Paraphrases of Predicate Phrases. Proceedings of COLING. 2008: 225-232.

[4] Li Weigang. Research on Chinese rehearsal examples and rehearsal template extraction. Ph. D. Thesis, Harbin Institute of Technology. 2008: 7-25

[5] O. Uzuner, B. Katz, and T. Nahnsen. Using Syntactic Information to Identify Plagiarism. Proceedings of the 2nd Workshop on Building Educational Applications Using NLP. 2005:p37-44.

[6] Rosti A V I, Ayan N F, Xiang B, et al. Combining Outputs from Multiple Machine Translation Systems[C]. HLT-NAACL. 2007: 228-235.

[7] Watanabe T, Sumita E. Machine translation system combination by confusion forest[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 1249-1257.

[8] Zukerman I, Raskutti B. Lexical query paraphrasing for document retrieval[C]. Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2012: 1-7.

[9] Bond F, Nichols E, Appling D S, et al. Improving statistical machine translation by paraphrasing the training data[C]. IWSLT. 2008: 150-157.

[10] Nichols E, Bond F, Appling D S, et al. Paraphrasing training data for statistical machine translation[J]. Information and Media Technologies, 2010,5(2): 950-971.

[11] Callison-Burch C, Koehn P, Monz C, et al. Findings of the 2011 workshop on statistical machine translation[C]. Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2011: 22-64.

[12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing systems. 2012: 1097-1105.

[13] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neuralnetworks for large-vocabulary speech recognition[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2012, 20(1): 30-42.