ISSN : 0974 - 7435 Volume 10 Issue 9

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(9), 2014 [4167-4172]

Bilingual corpus based machine translation research on key technologies

Zhen Li[1]*, Changhan Zhao[2]

[1]Ningbo Polytechnic, Ningbo, 315800, (CHINA)

[2]Wuhan Institution of Technology, Wuhan Hubei, 430070, (CHINA)

E-mail: lizhenlzhen@163.com

ABSTRACT

With the development of computer hardware and software, as well as corpus and bilingual corpus based machine translation has become the mainstream of current machine translation. This article provides an overview of the history of machine translation, especially machine translation model based on statistics and case histories and, finally, explore the introduction of formal syntax for statistical machine translation, to reach for a bilingual corpus-based machine translation technology to provide a theoretical basis for and support purposes.

KEYWORDS

Bilingual corpus; Based on statistical machine translation; Example- based machine translation.

INTRODUCTION

Based on statistical machine translation and example-based machine translation systems are two kinds of corpus-based machine translation system, the main difference between the two methods based on statistical machine translation does not require language knowledge base, rely on bilingual corpus through the estimation of the model parameters of the machine translation, mainstream methods are based on words, phrases and syntactic structure, occupies the dominance of machine translation at the present stage. Example-based machine translation using parallel corpus resources, based on the corpus of the source language and target language exchange and alternative translation results are obtained, to improve translation speed[1-3]. This study aims to further the study of corpus-based machine translation, explore the problem of machine translation method based on statistics and case histories, thereby improving the quality of machine translation.

MACHINE TRANSLATION PROCESS

To solve a unique civilization and written exchanges with foreign countries as a result of problems raising machine translation solutions. Machine translation refers to the use of computers from one natural language to another, automatic translation of natural language technology, with the improvement of the development of computer hardware and software, as well as corpus and machine translation show a spiral of history, history of machine translation[1] as shown in TABLE 1:

TABLE 1: Translation course

Time	Methods and defects
Early stages (50-60 's of the last century)	Method: direct translation. By using bilingual dictionaries or translation and simple mapping rules as knowledge, directly between the source language and target language translation. Defect: due to underestimating the difficulty of natural language processing, direct translation quality falls far short of investors ' expectations.
Middle stage (70-80 's of the last century)	Methods: the rule-based approach. Large amounts of linguistic analysis of source language and target language of the rules, including transformation based machine translation, based on the intermediate language translation and machine translation based on knowledge. Defect: compared to the direct translation, translation quality became the backbone of the business world, but it takes a lot of manpower and material resources and, in addition, because of language differences between some transformation rules could not be established.
Recent stage (Since the 90 's of the last century)	Methods: a corpus-based machine translation has become mainstream, contains two types of example-based machine translation and statistical machine translation.

BILINGUAL CORPUS BASED MACHINE TRANSLATION

The basic model of corpus-based machine translation

Computer to improve performance and perfection of the corpus, corpus-based machine translation in the past more 10 years has become the mainstream. Its basic model is shown in Figure 1:

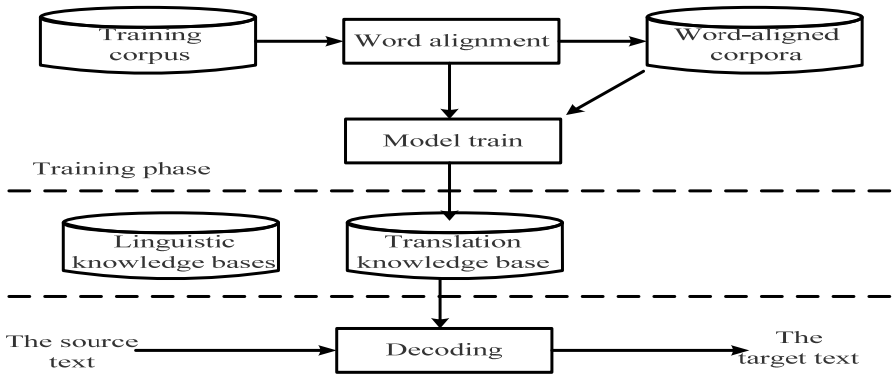

Figure 1 : The basic model of corpus-based machine translation

From the foregoing, machine translation generally consists of two stages: the stage of training or learning stage and a decode stage or translation stages. Align training is mainly used to solve word problems and related model design and training issues, reached the stage for subsequent decoding the provision of knowledge translation. Decode stage through the use of linguistic knowledge translation Knowledge Base training as well as other, to decode the original text, generate the corresponding target text.

Corpus-based machine translation method is the same basic pattern though, because the decoding process in addition to the translation of knowledge is different from addressing the problem of language differences will vary, currently a corpus-based machine translation system based on statistical machine translation system and can be divided into two types of example-based machine translation system[3]. Based on statistical machine translation

It is by relying on a large number of bilingual corpus and machine estimate the model parameters and to translate[4]. Braun and others propose statistical machine translation model, basic principles for translation as a process of information transmission, the noisy channel model is shown in Figure 2:

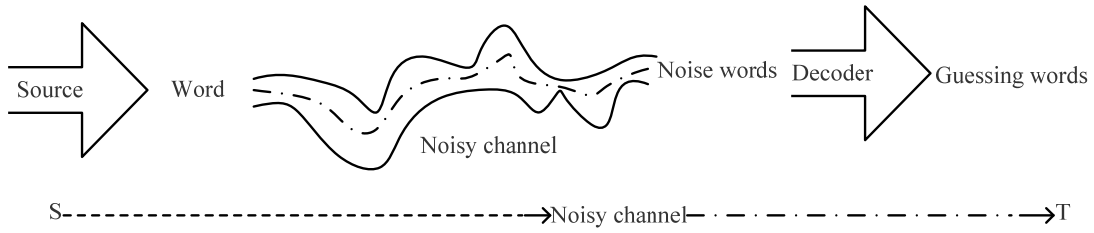

Figure 2 : Noisy channel model

Observe the noisy channel model, s language (target language) due to noise after a channel is distorted, so while the channel output is language t(source language), based on statistical machine translation by finding out where the most likely sentence as the source language translations of t, with a probability value is calculated as:

$$P(T|S) = \frac{P(T)P(S|T)}{P(S)} \quad \textbf{(1)}$$

Due to the denominator of the right side of the equation $P(S)$ and T, so, and $P(T|S)$ the maximum value corresponds to looking for a T; two products $P(T)P(S|T)$ are the largest of the molecule on the right side of the equation, namely:

$$T = \arg\max P(T)P(S|T) \quad \textbf{(2)}$$

$P(T)$: Language model of the target language;

$P(S|T)$: Cases for a given T-S model

Pure statistical machine translation can't satisfy translation requirements, introduced by IBM Corporation the Candide system is based on statistical machine translation is the most famous, however, the system only until 1995. Machine translation would then be made based on maximum entropy model and maximum entropy in statistical models for machine translation:

$$\Pr(E|C) = \frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(C,E)\right]}{\sum_{E'}\exp\left[\sum_{m=1}^{M}\lambda_m h_m(C,E')\right]} \quad \textbf{(3)}$$

$h_m(C,E)$: Features; λ_m : Corresponding feature weights. The decoding process is as follows:

$$\begin{aligned} E^* &= \arg\max_E \Pr(E|C) \\ &= \arg\max_E \left\{\exp\left[\sum_{m=1}^{M} \lambda_m h_m(C,E)\right]\right\} \\ &= \arg\max_E \left\{\sum_{m=1}^{M} \lambda_m h_m(C,E)\right\} \end{aligned} \quad \textbf{(4)}$$

Relative to the noisy channel model and maximum entropy model to provide a framework for an open, flexible and effective integration of all kinds of useful knowledge, so the translation quality is relatively high. Typical statistical machine translation system based on maximum entropy models are: decoder Pharaoh phrase-based statistical machine translation, Chinese colleges and universities, research institutions, construction of the Silk Road, and foreign researchers to develop the Moses[1, 5].

Example-based machine translation

Basic idea is to translate the instance of example-based machine translation based on the generated based on similar principles: Yes, usually consisting of three issues: 1) search match in the translation memory segments, through source examples finding the corresponding target language sentences; 2) establish efficient instance retrieval mechanisms, example-based machine translation is to take full advantage of the potential instances of debris at the phrase level, that is, aligned at the phrase level, establish a set of similarity criteria in order to determine whether two sentences or phrases, fragments are similar; 3) regroup the final translations translation fragment[3,6].

At present, example-based machine translation system[3] are as follows:

(1) Japan Kyoto University, the MBT1 and MBT2 systems of SATO, Makoto Nagao: MBT1 system selected for the example-based translation system, MBT2 is a complete example-based machine translation system; the translation process is divided into three-step decomposition, transformation, and synthesis.

(2) United States-Carnegie Mellon University PANGLOSS system of multi-engine machine translation system: the main engine is a knowledge-based translation system, example-based machine translation system as one of its engines, provides candidates for multi-engine machine system as a whole.

(3) The ETOC and EBMT system of Japan research laboratory ATR spoken language translation communication: ETOC system can retrieve the instance with the given sentences similar to the source language, machine translation systems for instance can use the case to clear up ambiguities, however, these two example-based machine translation systems is still incomplete.

In addition, China's Tsinghua University Department of computer science has established case-based Japanese-Chinese machine translation system.

THE STATISTICAL MACHINE TRANSLATION INTO THE FORMAL SYNTAX

While machine translation achieved fruitful research results, however, there are still many issues to be resolved, especially with long text, apart from polysemy, ambiguity structure and semantic ambiguity problems in natural language processing, but at all stages of machine translation there are special difficulties, for example, translation between different languages reordering issues remain to be resolved, such as[6]. Taking into account the differences between the languages and the difficulty of Chinese syntactic analysis, thus research in statistical machine translation into formal syntax in order to improve the interpretation of bilingual translation ability and fault tolerance, put forward the following contents[1]:

Word alignment based on syntactic knowledge model

Word alignment is one of the basic components of corpus-based machine translation, in view of the present word alignment is facing some sentence syntax tree is difficult to get, hard to ensure the quality problems, such as Wu, puts forward the reverse transcription of grammar, this paper improved the reverse transcription grammar, so that it can explain a many-to-many alignment, and on the basis of reverse transcription grammar does not destroy, which implied the structure of the constraint is converted into an explicit position judgment, thus will reverse transcription in grammar contains the syntactic constraints effectively integrated into the word alignment model, at the same time, design a more efficient algorithm, reduce the complexity of the line search algorithm.

To further enhance the order of word alignment constraints, design syntax tree and ITG similarity metrics between trees, constraint syntax tree into word alignment based on ITG's model, through the integration of these two types of syntactic knowledge, and effectively improve the quality of the word alignment.

Tree-tree model for statistical machine translation

Sequencing problem for machine translation in the target word proposed tree-tree-based model for statistical machine translation. In global range, the model, through in original sentences of syntax tree and ITG tree for mapping, to constraint target word of order changes, in local range within, model in the contains has based on ITG of local heavy set sequence model features, by putting the two pieces of the direction of the prediction to transform both the direction of the adjacent blocks in pairs, which can predict between any two pieces of translation direction. The integration of local model and global model, effectively explain the complicated relationship between the source and the target sentence, improve the quality of machine translation.

Similar case retrieval methods based on bilingual information

Translation of example-based machine translation using the principle of analogy, in similar conditions for instance, can produce a smooth translation. So how to retrieve similar cases in the massive case, it is important for the quality of the EBMT. We present a novel method of similar case retrieval, based on the translation of word alignment information in the instance, designed a series of similarity, is used to calculate the similarity between sentences and examples of your input sources, so as to improve the quality of search, meanwhile, devised a two-tier training corpus indexing structure, improving the efficiency of retrieval.

Case-based model for statistical machine translation

As the source sentence tree - tree model is the perspective of syntax tree, there is no guarantee that the target sentence syntactic rationality and fluency. Therefore, put forward a kind of statistical machine translation and examples of the hybrid model of machine translation, he is an extension of the tree - a tree model, which combined with the instance of knowledge, using the instance contains the target sentence structure information, as well as the existing source of syntax tree information, further constraints of target word selection and order change, improve the quality of translation, at the same time, presents the corresponding decoder based on instances, it combined with statistical information, knowledge, and strength and improve the quality of the decoding efficiency.

CONCLUSION

Machine translation is a classic problem, but in the highly developed computer technology, under the background of basic forming corpus construction, based on the corpus of machine translation is a burgeoning field. Machine translation system based on statistics and the instance is based on the corpus of the two systems, the maximum entropy model is put forward, machine translation for the integration of different knowledge platform provides an effective framework. For now, the main task to study how to machine translation in the system combined with more and more useful knowledge, based

on the illustration of the corpus-based machine translation after two proposed a model of word alignment based on syntactic knowledge, tree - statistical model, similar case retrieval method based on bilingual information and statistical machine translation model based on the instance, hope to improve the quality of machine translation.

REFERENCES

[1] W.H.Chao; Bilingual corpus based machine translation research on key technologies, Harbin, National University of Defense technology, **(2008)**.

[2] L.Chen; Study on english and chinese parallel corpus based machine translation knowledge acquisition, Beijing, Beijing Jiaotong University, **(2012)**.

[3] Z.W.Feng; A corpus-based machine translation system, Terminology Standardization and information technology, **1**, 28-35 **(2010)**.

[4] J.J.Ma; Rule-based and statistical machine translation method of comparative analysis of the ambiguity problem, Journal of Dalian University of technology (social sciences Edition), **31(3)**, 114-119 **(2010)**.

[5] H.S.Liang; Research on the key issues of the training on statistical machine translation models based on one language, Weihai, Harbin Institute of technology, **(2013)**.

[6] L.Yin; Corpus-based machine translation in the translation of knowledge acquisition, Beijing, Beijing Jiaotong University, **(2014)**.