International Journal of New Developments in Engineering and Society
ISSN 2522-3488 Vol. 3, Issue 3: 10-24, DOI: 10.25236/IJNDES.19302

Research and design of **the hybrid model translation platform based on computational linguistics**

DiYunTing[1] ,ZhangLin[2] , TianSong[*,1]

1. Computer Academy of Hubei Polytechnic University, Huangshi, Hubei, 435000,P.R.China

2. State Grid Taiyuan Power Supply Company,Taiyuan,Shanxi,030000,P.R.China

**Corresponding Author*

ABSTRACT. *Computational linguistics is a science combined with not only computer science and linguistics, but also mathematics, cognitive science ,and so on. With the application of computer technology into language process ,it emerged and developed gradually into an independent system. Study on linguistics influence on the development of computational linguistics and translation platform construction from the perspective of computational linguistics, promotes deep understanding ,the further research explore and effective practice on translation. There seems a certain plausible division of labor between linguistics analysis methods and language process in computational linguistics, which are combined to form a hybrid system. The approach to design the hybrid system is explored from the application of translation platform construction from the perspective of computational linguistics. The paper aims to provide some new perspectives to the translation and the teaching of foreign languages, firstly, analyzes the basic theory of the hybrid model translation platform based on computational linguistics, and then designs the translation platform in detail from the system design idea, overall structure design, and function module design three aspects.*

Keywords: *Computational Linguistics; Hybrid Model System; Translation Platform*

1. INTRODUCTION

With the coming of information age, to master and skilled use of computer technology has become the era and society the most basic requirements for each

worker. The use of language for human is the main carrier of information and knowledge. In the Internet age, the study of human language of computer generated and understanding of the language information processing become one of the modern hot subject. And the development of society and science and technology ask for general linguistics, computer technology, mathematics and talent of cognitive science. Computational linguistics is a combination of computer science, mathematics and linguistics and other disciplines of knowledge, not only in-depth study and linguistic phenomenon, but also provides scientific theoretical guidance for the computer application technology. Computational linguistics and linguistic analysis method combining to form hybrid system, produce positive effects on translation platform construction.

Computational linguistics and natural Language information processing research at the core of the problem is the automatic understanding of Language and automatic generation of it. The former from the surface of sentence the words string identifying the syntactic structure of the sentence, judge the semantic relationships between components, finally find out to express the meaning of the sentence; The latter from to express the meaning of the word choice, according to the semantic relationship between words structure between each component of the semantic structure and syntax structure, eventually create a sentence of grammar and logic. The current computational linguistics mainly engaged in natural language processing, their goal is to make using natural language communication between man and computer. Specifically, it is establishing various computer application software system of natural language processing, such as machine translation, natural language understanding, automatic speech recognition and synthesis, text automatic recognition, computer aided teaching, information retrieval, automatic text classification, automatic summarization, and in the text information extraction and intelligent search on the Internet, as well as a variety of electronic dictionary and terminology database. But these studies are more or less affected by linguistics and the guidance of. In the study of linguistics, appeared a lot of expression of the concept of similar terms, such as the traditional theory of linguistics, linguistics, this article unified the term by linguistics. [1]

Previous studies mostly discuss computational linguistics influence on linguistics, Feng Zhiwei with new theory of the level of human understanding of natural

language symbols, computational linguistics is a challenge to traditional linguistics are discussed. [2] Bumairemu argue that computational linguistics branch of linguistics brings different impact and challenges at the same time to promote the development of linguistics. [3] Computational linguistics to traditional syntax, morphology, semantics, logic grammar, vocabulary, and so on, has important influence. At the same time, the development of computational statistics also bring new perspectives to linguistics, for example, Ji Tieliang combines linguistics and statistical methods such as establishing Chinese verbs subclasses frame type set. [4] Yao Minfeng described a Chinese-English machine translation system based on phrases in combination, to build a platform of Chinese-English machine translation has a positive effect. [5] Computational linguistics and linguistic effect between should be mutual. These studies focused on the influence of computational linguistics on linguistics, this article from the perspective of computational linguistics to explore translation platform design and construction of hybrid model system studies.

2. THE HYBRID MODEL SYSTEM

As the computational linguistics research in theory and application aspects of language processing continuous evolution and development, many facets of the between fuzzy linguistics and computational linguistics gradually formed a relationship. In the process of the development of computational linguistics, linguistics plays an important role. Computational linguistics combines computer science and linguistics and formed a clear division of responsibilities between the two hybrid model systems, the hybrid system on construction of translation platform has a strong practical guidance.

2.1 Summary of computational linguistics

Computational Linguistics originated in the 1950s, also known as the Natural Language Processing, computational linguistics is to use computer technology to study and an emerging discipline of dealing with the Natural language, is related to linguistics, psychology, psycholinguistics, brain science, artificial intelligence, computer science, philosophy, logic, mathematics, information theory, literature,

aesthetics, and many other areas of a cross discipline. [6]

Computational linguistics has experienced the start-up phase (50s American linguist Chomsky's generating method and the function of the sixties and seventies grammar makes the modern linguistics into the formalization, standardization and systematization of contemporary linguistics, is the negation of American structuralism grammar and to the return of the traditional grammar.), the low tide stage (ALPAC report of the national academy of sciences in 1964 as the mark), golden age (70s), and entered a stage of vigorous development of the 80s. Its research has expanded from the original machine translation to the natural language understanding, information retrieval, speech recognition and synthesis, computer assisted instruction, etc.

Machine translation in China started earlier, but there is a period of stagnation until the late 1980 s is the first translation software “YiXing”. Since the 90 s, the translation software is available, but the overall quality is not high, mainly because of the computer automatically analyze and understand the problem of Chinese has not been very good solution, linguistics studies also is unable to provide more law of Chinese natural language for use by the computer expert, the difficult problems in computer automatic segmentation - not login and ambiguous segmentation problem hasn't been solved yet. Hu Mingyang in computational linguistics lecture series, especially when it comes to 1989 years of the joint chief conference, Tsinghua university computer experts put forward to solve the modern Chinese grammar question at the meeting is computer processing natural language need to first solve the problem. Since the 1990s, the main research direction is the man's language knowledge systematically articulated in the formal way.

2.2 The influence of linguistics in computer linguistics

Computational linguistics is not about computer language discipline, not break about mathematical linguistics or a new branch of linguistics, applied linguistics, it is neither a binary machine language, the study is not to write a computer program used in programming language, but in the human to know the world and the creation of civilization in the process of the formation of the natural language. In the 1980 s, Lauri Karttunen found application of the theory of computational linguistics and

computational linguistics coexistence and mutual promotion, at the same time, the branch of computational linguistics theory to understand and use play an important role in human language structure. However, relationship between linguistics and computational linguistics a lot has changed over time. These changes are reflected by five paradigm of computational linguistics, in each paradigm, linguistic theories play a role, have different influence on computational linguistics research.

The first paradigm is directly enable processing language program. Accepted the relevant linguistic theory education operators, directly to enable such as FORTRAN, COBOL, etc. A computer program or assembler language processing, etc. This stage of linguistic knowledge and no systematic difference between processing method. The second model is the development of professional language processing algorithm and methods, such as parsing algorithms, finite analysis and expanding phrase structure grammar. Under this paradigm found the distinction between linguistic knowledge and processing program, but the improvement of research methods can't do without the guidance of linguistic theories, require a certain degree by using the theory of linguistic knowledge. The third paradigm is the emergence of linguistic form system. The 1980 s saw the emergence of a series of new system of grammatical forms, such as HPSG (Head-Driven Phrase Structure Grammar), LFG (Lexical-Functional Grammar) theory system has influence on the computational linguistics, such as the collection system of grammar patterns form and the semantic system, the form model and the linguistic theories are closely linked, so many model system is placed in the linguistics course professor.

When these linguistic formalism model cannot satisfy the practical application, the fourth paradigm used in soon and become the dominant method of natural language processing, namely professional method of natural language processing. So the researchers to focus on the improvement of processing technology, to reduce the importance of language and linguistics. The appearance of the fifth kind of paradigm is in computational linguistics to statistical methods in some application fields, natural language processing began to reconsider the method and the knowledge source of linguistics. Statistical methods in natural language processing experts try to return to linguistics in lexicology, or try to establish a statistical model based on the structure of phrases. The combination of statistical and linguistics method contributed to the generation of computational linguistics is the fifth kind of

paradigm, namely statistics and statistical machine learning methods and innovative combination of linguistic methods.

With the development of computer technology and deep research in the theory of language, the first three paradigm gradually withdrew from the study of the center position, the latter two paradigm will be important method of computational linguistics statistics combined with linguistics, as a new progress of natural language processing paradigm. Linguistics and the reasonable application of statistical methods in computational linguistics can promote in-depth development of language research. Therefore, for the division of labor and the combination of the two discusses the formation of the hybrid system is especially important.

2.3 The division of labor and the combination of Computational Linguistics and Linguistics

The development of the Statistics constantly change the relationship between the computational linguistics and linguistic. Statistics student was used in computational linguistics, and the combination of linguistic theories, its role in the study of hybrid system. In some areas of language processing, the design of the hybrid system approach has shown the prospect of results. The first design hybrid system contains both linguistics also include the composition of computer technology, make the two languages analysis method together to complete the lexical phrase sentence processing tasks. In mixing machine translation study, the task of the hybrid system is systematically for the input of language to explore the combination of statistical and language rule is the most ideal results. By experienced linguist input language for a detailed semantic analysis, found the best statistical system by the corresponding output language vocabulary phrases or sentences chains, and decide which kind of output is the most appropriate translation. System using a given linguistic grammar transformation rules in advance of the lexical phrase sentence chain set selection combination, so as to get the corresponding output language statement. The use of computational linguistics technology combined with language rules system to explore lexical phrase sentence translation method just try for hybrid processing system. Another design method of hybrid system is based on the theory on the study of the whole discourse. This discourse hybrid system is the supplement to the first hybrid system, it is not only the phrase structure, more will match the

phrase structure rose to the height of the discourse, is a higher level of exploration. In this way, the development of computational linguistics and linguistic research on hybrid system, mixed machine translation and translation platform construction play an important role.

2.4 Statistical collocation model

Collocation is usually composed of two or more words in people's expression habit. According to different application, match usually have different emphases. For example: from the perspective of linguists, collocation are fixed phrases and words can be combined freely between a kind of language phenomenon; While statistics linguists think, Collocation usually refers to not accidentally often appear together phrases [7]. In this article, Collocation by the customary use of two words together, they can be adjacent, also can't adjacent. Match these include proper nouns, idioms, the combination of conjunctions and other word, for example: verb and noun, adjective and noun, adverb and verb and adverb + adjective, and so on.

2.4.1 The monolingual statistical word alignment model [8]

First of all, we copy the each sentence in monolingual corpus, and get one with the same sentence to sentence, and then use the monolingual statistical word alignment algorithm search potential on the collocation of word in the sentence. According to the monolingual statistical word alignment algorithm, given a single sentence $S = w_1^l$. Using monolingual word alignment model (In short MWA3) algorithm to get the optimal alignment result. It shown below:

$$A^* = \arg\max_A \left\{ P_{MWA3}\left(A \mid S\right) \right\} \tag{1}$$

$$P_{MWA3}\left(A \mid S\right) \propto \prod_{i=1}^{l} n\left(\phi_i \mid w_i\right) \cdot \prod_{j=1}^{l} t\left(w_j \mid w_{a_j}\right) \cdot d\left(j \mid a_j, l\right) \tag{2}$$

Among them, *A* is the word alignment sequence, any one word cannot be aligned and in its own right, therefore, alignment collection is expressed

as $A=\{(i,a_i)\mid i\in[1,l]\&a_i\neq i\}$; ϕ_i is the number of words corresponding to w_i. The model of formula (2) mainly use three kinds of probability model: based on the collocation of word probability model $t\left(w_j \mid w_{a_j}\right)$, based on the location of collocation probability model $d\left(j \mid a_j, l\right)$ and word derivative model $n\left(\phi_i \mid w_i\right)$.

Figure (1) shows the sentence "team leader plays a key role in the project." the single words of alignment results, among them, the alignment of the word for the potential of the collocation of word, such as a "key role", "play a key role" and so on.

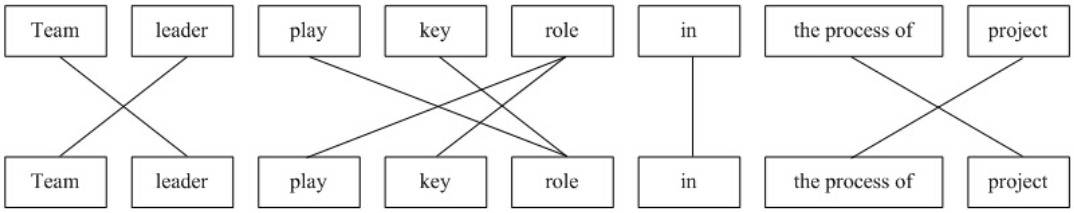

Fig. (1). MWA Example

2.4.2 The Statistical Collocation Model

After the monolingual statistical word alignment processing of corpora, the calculation on the frequency of alignment word. In the experiment, we filter out the word for frequency is less than two. Aligned based on word frequency, we calculate the probability of each word on the alignment of:

$$p\left(w_i \mid w_j\right)=\frac{freq(w_i,w_j)}{\sum_{w'} freq(w',w_j)} \tag{3}$$

$$p\left(w_j \mid w_i\right)=\frac{freq(w_i,w_j)}{\sum_{w'} freq(w_i,w')} \tag{4}$$

Among them, the collocation of words is equivalent to the two words. As a result, the collocation of word for probability using the average of the probability of two above it shown in formula (5).

$$r\left(w_i, w_j\right) = \frac{p\left(w_i \mid w_j\right) + p\left(w_j \mid w_i\right)}{2} \tag{5}$$

Formula (5) describes the probability of the collocation of two words, it can see that if the match between the two words probability is higher, so the collocation of the two words have stronger relationship. On the other hand, the collocation relationship is weak.

This method is effective to describe the internal relations between words, statistical collocation model was successful in other NLP applications, such as raising bilingual word alignment, function of the translation system of the tuning sequence, and so on.

3 THE DESIGN OF THE HUBRID MODEL TRANSLATION PLATFORM

In China's foreign trade, culture, and the rapid development of science and technology exchange under the background of rising demand for translation industry, the progress of the language information processing technology bring to translation career of great change and impact. Changes in the environment for language services companies find a new business model, adopt a new strategy and new management model, and improve the production efficiency. A lot of language services companies monthly million word level translation projects have become common, requirements in a short period of time according to the predetermined quality standards to complete a lot of translation. Put forward new requirements to the language service workers. However, the traditional mode of small workshops “proof”, careful, manual translation process is clearly no longer adapt to today's mass, teamwork translation business process. Modern language information processing and other industries need to master the principle and application of machine aided translation talents and related natural language processing technology development, thus explores the translation platform construction in computational linguistics perspective is very important, especially in hybrid system under the

research of the machine translation system.

3.1 The Design Idea

The design of the hybrid model translation platform based on computational linguistics mainly based on the following design thought.

Firstly, organic fusion of various translation strategies. Existing practical rules of machine translation systems are mostly based on the analysis of the transformation, but due to the complicated natural language, and use the conceptions, makes the translation system is not only need to build a large system of rules describe the complicated language phenomenon, also need to keep adding personality rules in order to enhance adaptability of the system. As the number of rules increased to a certain extent will inevitably cause the phenomenon such as redundancy and conflict, so that the high quality MT knowledge base construction and maintenance of the difficulty is high, the quantity is big.

And based on the statistical analysis of corpus and based on the analogy of the translation memory in the establishment of large-scale corpus, and statistical model of analogical reasoning method instance library construction, the representation of a language patterns and require a lot of work in such aspects as analogy inference implementation and implementation complexity. But if, as a rule analysis method is an effective supplement, organically combined with the advantages of several translation methods, USES the many kinds of translation strategies for parallel processing and optimal choice, not only can improve the quality and speed of automatic translation, but also provide various levels of linguistic knowledge accumulation, the way and efficiency of knowledge acquisition will help to improve the translation.

Secondly, the efficient translation editing environment. In general, the machine translation system for translation error is several kinds of common situation. As a result, the system should be able to all kinds of analysis results based on the translation process, and knowledge, to provide intelligent editing tools, efficiently choose the correct translation string, adjustment means location in the translation. At the same time of the modifier on can be limited enumeration easily realize add, delete and modify, thus quickly obtain high quality translation.

Thirdly, using the object-oriented knowledge database management technology. In a multi-strategy machine translation system, knowledge rules and statistical knowledge, the knowledge dictionary, pattern, and based on the surface of the string instance of knowledge, etc. In the storage of prior knowledge, after processing, editing and translation of translation knowledge storage each processing link, such as the knowledge and other knowledge is based on knowledge related to complex, quality and efficiency of management system of knowledge base for processing performance of the system is very important. Therefore, it is necessary to adopt the object-oriented knowledge representation and organization structure of knowledge organization and processing.

3.2 The Overall Structure Design

According to the above design idea, the overall design scheme of the hybrid model translation platform based on computational linguistics is shown figure (2).

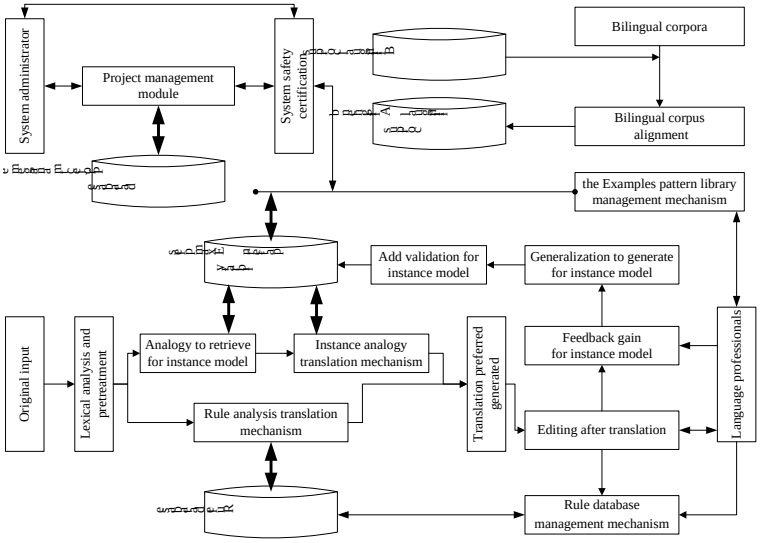

Fig. (2). The Overall Design Scheme of the Translation Platform

Among them, the lexical analysis and pretreatment of the source text input lexical analysis and retrieval model of pretreatment, the strip generated source

substructure components constitute the characteristics of the model. Instance model analogy model retrieval based on the current statement of retrieval characteristics, the characteristics of instance model repository candidate model retrieval. Instance analogy translation mechanism to retrieve a candidate model and heuristic analogical matching source sentence, and match the pattern solution was obtained from the instance in the pattern library mode, construct a translation. Instance model repository to store a large number of various natural language sentence structure patterns and corresponding solution of the model. Rule analysis translation mechanism analysis of sentences based on the rules of conversion, and get the corresponding translation. Inventory to put all the language rules. The preferred generated according to the analogical matching analysis of similarity and rule the credibility of the corresponding translation as the ultimate goal of translation. Aligned bilingual corpus and bilingual corpus, respectively, to hold the bilingual corpora and aligned bilingual corpus. Aligned bilingual corpus annotation of bilingual text corpus sentence level and paragraph level of alignment and feature labeling. Project management database and project management module to achieve storage and management of project. Instance model example of library management mechanism in storage and various management functions.

3.3 The Design of Function Module

In the process of the construction of system platform, the function module design is very important. Only good design the specific function of each function module in detail can better and faster to complete the construction of system platform. According to the overall design of the system, the hybrid model system translation platform can be divided into six functional modules, the specific function of each function module design is shown in figure (3).

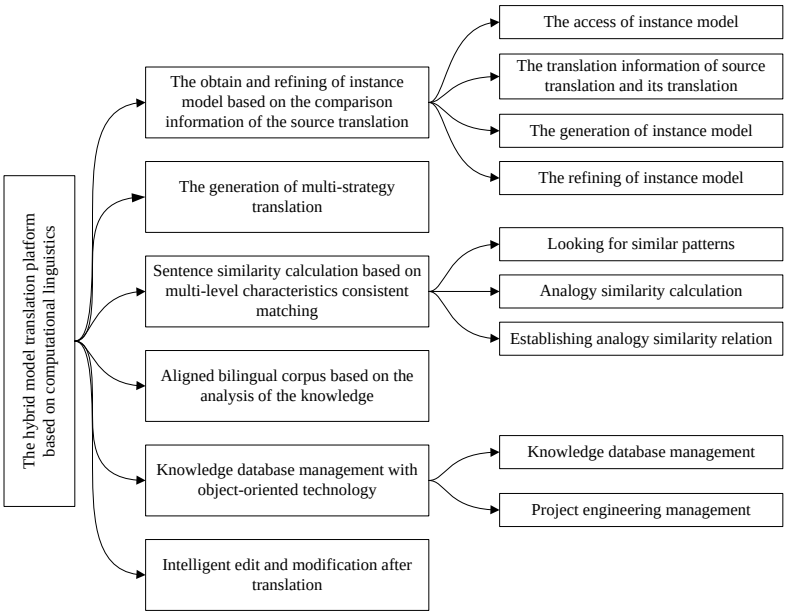

Fig. (3). The Specific Function of each Function Module Design

4 CONCLUTION

The development of computational linguistics and linguistics and its related theory research provides technical support for translation platform construction and theory of guarantee. The core power of platform construction is the design of the mixing machine translation system. Language rules for translation platform construction of hybrid system design provides the premise condition. Corpus resources construction and the improvement of language information processing technology is the important resources in the hybrid system research and technology guarantee. The mutual promotion between the computational linguistics and linguistics to translation platform construction plays an important role. At present although computational linguistics in some linguistic research field has made good progress, but with the wide popularity of the Internet, language information processing needs bigger and bigger, there is an urgent need to use the means of automated processing language information, still need the further research of language workers. Therefore, future research should pay attention to the theory of

linguistics and the influence of computational linguistics, further exploring computational linguistics in language study, important application in the field of language information processing, and so on. This paper firstly analyzed the basic theory of the hybrid model translation platform based on computational linguistics, and then designed the translation platform in detail from the system design idea, overall structure design, and function module design three aspects.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGMENTS

This work was supported by the No.1 project :Teaching and research project of Hubei Polytechnic University(project No.: 2017C39), Reform and research of “embedded operating system” course under “15personnel training plan”. No.2 project: Innovation and entrepreneurship project of college students in Hubei (project No.: 201710920019), Design of fun running planning system for “cool running”.No.3 project:2019 Guidance projects of department of education of Hubei Province: Research on Development and Transplantation of Embedded Operating System Kernel.

REFERENCES

[1] Zhang Xiaoyan, and Song Tiehua(2013). Computational Linguistics and Translation Platform Construction from the Perspective of Computational Linguistics,” Journal of Shanxi Agricultural University (Social Science Edition) (In Chinese), vol. 12, no. 4, pp. 359-362.

[2] Feng Zhiwei(1992), “The challenges of the calculation of theoretical linguistics,” Applied Linguistics (In Chinese), no. 1, pp. 84-87.

[3] Bumairemu Abula(2004). Introduction to computational linguistics and its influence on the theory of linguistics,” Journal of Hetian Teachers College (In Chinese), no. 1, pp. 79-80.

[4] Ji Tieliang, Sun Weiwei, Sui Zhifang(2007). The Acquisition of Chinese Verb’s

Subcategorization Frame Types Based on Linguistic Theory and Statistical Algorithm. Journal of Chinese Information Processing (In Chinese), vol. 21, no. 5, pp. 118-125.

[5] Yao Minfeng(2010). A Chinese-English Machine Translation System Based on the Combination of Target Language Phrases. Journal of Guangdong University of Foreign Studies (In Chinese), vol. 21, no. 2, pp. 75-77.

[6] Huang Jianshuo(1991). The Review of Computational Linguistics Research. International Academic Developments (In Chinese), no. 4, pp. 24-31.

[7] McKeown KR, Radev DR. Collocations(2000). A Handbook of Natural Language Processing. NewYork: Marcel Dekker, pp. 507-523.

[8] Liu ZY, Wang HF, Wu H, Li S(2009). Collocation extraction using monolingual word alignment method. In: Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing, pp. 487-495.

[9] Cao Lingmei(2015). 4T Teaching Mode with Internet Translation Platform: A Case Study of Zhejiang College of Shanghai University of Finance and Economics. Journal of Southwest Jiaotong University (In Chinese), vol. 16, no.1, pp. 69-74.

[10] Zhang Jian, Li Sujian, Liu Qun(2002). Statistical N-gram Method Used in Machine Translation System. Computer Engineering and Applications (In Chinese), no. 8, pp. 73-75.

[11] Huang Heyan and Chen Zhaoxiong(2004). Overall Design of an Intelligent Computer-Aided Translation Platform Based on Hybrid Strategy. Journal of computer research and development (In Chinese), vol. 41, no. 7, pp. 1266-1272.

[12] Liu Zhanyi, Li Sheng, et al(2012).Improving Example-Based Machine Translation with Statistical Collocation Model. Journal of Software (In Chinese), vol. 23, no. 6, pp. 1472-1485.

[13] Somers H(1999). Review article: Example-Based machine translation. Machine Translation, vol. 14, no. 2, pp.113-157.

[14] Liu ZY, Wang HF, Wu H(2006). Example-Based machine translation based on tree-string correspondence and statistical generation. Machine Translation, vol. 20, no. 1, pp. 25-41.

[15] Zhao SQ, Zhao L, Liu T, Li S(2010). Paraphrase collocation extraction based on binary classification. Journal of Software, vol. 21, no. 6, pp. 1267-1276.