**УДК 004.9**
**МРНТИ 81.93.29**

# A PHONE-BASED TRANSLATION APPLICATION

**A. RAZAQUE, S.T. AMANZHOLOVA, G.A. TOLGANBAYEVA,**
**N. RAPILBEK, B. KEMERBAY**

*International IT University*

***Abstract:*** *we design a phone-based translation application to translate Chinese text into English. As, this application should recognize the most signs. However, there would be a challenge of maintaining the higher accuracy because one word possessesseveral meanings. In addition, longer text translation and connecting the application with phone's camera would be problematic andrequires proper attention. To handle these issues, Rule-based machine translation (RMT) method is implemented. Finally, the proposed RMT is compared with Google translation and Kingsoft PowerWord from accuracy perspective. The results demonstrate the higher accuracy of RMT.*

***Keywords:*** *text translation, phone-based translation, rule-based machine translation, Google translation, Kingsoft PowerWord*

## ТЕЛЕФОН АРҚЫЛЫ АУДАРУҒА АРНАЛҒАН ҚОЛДАНБА

***Аңдатпа:*** *Біз қытай тіліндегі мәтінді ағылшын тіліне аудару үшін телефонға негізделген аударуға арналған қосымша жасаймыз. Осылайша, бұл бағдарлама көптеген белгілерді тануы керек. Алайда, жоғары дәлдікті сақтау мәселесі пайда болады, өйткені бір сөз бірнеше мағынаға ие болуы мүмкін. Бұдан басқа, мәтінді ұзағырақ аудару және қосымшаның телефон камерасына қосылуы қиындық тудырады. Сондықтан да осы мәселеге айрықша назар аударуды қажет етеді. Осы проблемаларды шешу үшін ереже негізіндегі машиналық аудармасы (RMT) әдісі енгізілді. Ақырында, ұсынылған RMT Google және Kingsoft PowerWord-пен дәлдікке аударумен салыстырылады. Нәтижелері RMT-ның жоғары дәлдігін көрсетеді.*

***Түйінді сөздер:*** *мәтінді аудару, телефондық аударма, машиналық аударма, Google аудармасы, Kingsoft PowerWord аудармасы*

## ПРИЛОЖЕНИЕ ДЛЯ ПЕРЕВОДА НА ТЕЛЕФОНЕ

***Аннотация:*** *Авторы разрабатывают приложение для перевода на телефоне, чтобы перевести китайский текст на английский. Это приложение должно распознавать большинство символов. Здесь возникает проблема поддержания более высокой точности, поскольку одно слово может иметь несколько значений. Поэтому, более длинный перевод текста и подключение приложения к камере телефона оказываются проблематичными и требуют должного внимания. Для решения этих вопросов был реализован метод машинного перевода на основе правил (RMT). Наконец, предлагаемый RMT сравнивается с переводом Google и Kingsoft Power Word с точки зрения точности. Результаты демонстрируют более высокую точность RMT.*

***Ключевые слова:*** *перевод текста, машинный перевод, телефонный перевод, перевод Google, перевод Kingsoft Power Word*

## Introduction

Translation has been playing an important role in the future. It can be convenient for people when having problems in communication. The translation App is of high interest for those individuals who are involved in learning foreign language. Therefore, maintaining the proper accuracy of translationis of paramount significance to understand the language. As, less accuracy-providing app could create the problem. We know one word has many meanings. Thus, there is challenge of translating the long text accurately. These words in the text should choose appropriate meanings. Thus, handling the issue of accurate translation, there is need of robust translation App. The difficulty of app development does not only depend on the platforms, but it also depends on the nature of translating language [1-2]. In this paper, we introduce rule-based machine translation for translating the text. This translation can be grouped into three translationmeasures: direct translation, interlingua and transfer approaches. We choose the direct translation method because it provides simple and easy standard when translating from English to Chinese language. Furthermore, we also use other measures like Interlingua approach to compare these measures to conclude which is the most efficient. Besides, we should pay attention to how to product on translation of signs and the connection between phone's camera and application. We apply Rule-based machine translation algorithm to translate the text accurately.

The remainder of the paper is organized as. Section 2 signifies the problem identification. Section 3 presents the salient features of the existing work. Section 4 describes the parsing process and machine rule-based translation algorithm. Section 5 presents result and evaluation and finally entire paper is concluded in section 6.

This paper contributes as

- RMT accurately translates English to Chinese words.
- It provides better bilingual word chunk recognition as compared to Google translation and Kingsoft PowerWord.

## Problem Identifying

People have many complaints on these machine translation applications. They have advantages on translating every single word, even it is uncommon. However, when it comes to translate a whole sentence, most applications fail to express the appropriate meaning, and only Google Translate can do it well. Some people find that some applications is too slow to translate an article which is more than 200 words. They may take 10-20 seconds. Google Translate also does best during the applications. There are some more problems like translations cannot be copied, some translations cannot be revised even there are some obvious mistakes during them and so on. Let me take Google Translation[3] as an example. Google Translate is a free tool that can help you instantly translate sentences, files, and even the entire website. The computers used a program called statistical machine translation. It means that the computer is based on a variety of patterns found in a large number of texts. If you want to teach someone a new language, you may first teach him vocabulary and grammar rules to explain how to construct a sentence. Computers also learn a foreign language by the same way - by referring to words and by a series of rules. When you try to include all the special cases and exceptions in a computer program, the quality of translation starts to decline. Google Translate takes a different approach. Instead of teaching computers all the rules of the language, we let the computer discover rules themselves. Computers discover the rules by analyzing tens of millions of files that have been artificially translated. The results are from books, institutions such as the United Nations and websites around the world. Our computers scan these texts, looking for patterns that are statistically significant -- that is, there is no accidental pattern between translation results and the original text. Once the computer finds these patterns, it should be able to use these patterns to translate other similar texts in the future.

## Related Work

First, we need to parse the source statements both in Chinese-to-English and English-to-

Chinese translation. It almost includes automatic hyphenation, part-of-speech tagging, word sense disambiguation, parsing and semantic analysis [4].

1. automatic hyphenation

Automatic hyphenation means that the words that are not clearly delimited are automatically cut into strings. It includes dot symbols, figures, mathematical symbols, tags, names, locations, organization and so on. These unregistered words need to be identified by machine. I take a sentence as an example, “They are reading.”The machine uses the segmentation module to cut it into:They / are / reading. So that it means that the statement is made up with three words. There are two problems which are hard to deal with in the segmentation module. One is that vocabulary in the dictionary required in the segmentation module must be comprehensive. Another is that we need to provide a proper measure to segmentation ambiguity. It may be a long-way work to do because it requires large amounts of segmentation ambiguity rules and participation of many linguistics experts. In this module word segmentation algorithm is divided into lexical participle and lexiceless participle. Lexical participle is main word segmentation measure. It is divided into measures based on rules and statistics.

2. Part-of-speech tagging

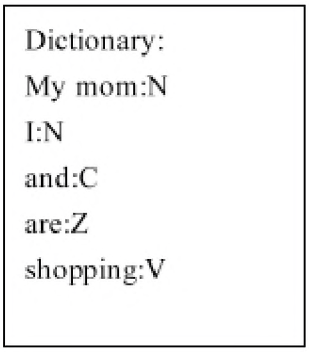
In Chinese, a word may take different part of speech in different situations. Part-of -speech tagging means that the machine determine every word’s grammar category in the sentence and ensure its part of speech to tag it. Take an example,” He is editing files.” After the machine does automatic hyphenation and part-of-speech tagging, it should show us “He/n is/z editing/v files/n”. Here, n means noun, z means adverbial and v means verb. The algorithm takes advantage of measures based on rules, and its principle is to do disambiguation to the words which have many part of speeches by using the rules that have been designed already and keep the last and right part of speech. It mainly includes:

- A separate annotation rule database is established for part of speech ambiguity.
- When tagging, if some word has many part of speeches, the machine should search for the rule database.
- Identify and eliminate the ambiguity with the same pattern. If not, the machine should save it.
- The program and rule database are two separate parts: Parsing and application’s machine dictionary.

## Parsing Process Rule-Based Machine Translation Algorithm

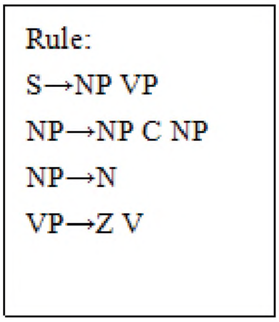
It is the progress of making word strings to syntactic structure. This syntactic structure should be a tree. We need to choose a proper syntax theory to do parsing, and here we choose context-free grammar. Now look at the sentence “My mom and I are shopping. “The rule table 1 and dictionary table 2 and syntax tree depicted in Figure 1.

**Table 1 – Dictionary table**



Dictionary:
My mom:N
I:N
and:C
are:Z
shopping:V

**Table 2 – Rule table**



Rule:
S→NP VP
NP→NP C NP
NP→N
VP→Z V

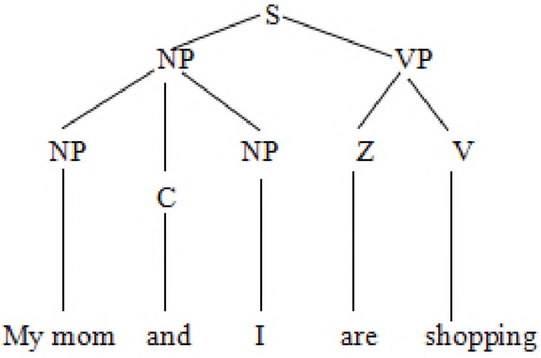


*Figure 1 - Syntax tree*

Some other translation system does semantic analysis to the source language and most of them use case grammar. Machine dictionary uses sememe method to store meanings.

Second, our application's machine dictionary uses database as storage. Every record stores one word, and part of word storage rules are described in Table 2.

**Table 3 – Part of word storage rules**

#D:noun, singular form
#F:noun, plural form
G1:nominative case
G2:objective case
V1:room form of the verb
V2:past tense of the verb
V3:verb, present participle
V4:verb, past participle
So, some examples are as shown below:
我:#N,G1-I,G2-me/
今天:T-today/
买:V1-buy,V2-bought,V3-buying
了:have done/
书:book/

The quality of translation application depends on the increase of the words in the machine dictionary, so we have to set a project to enrich our machine dictionary. Because the whole grammatical analysis period requires machine dictionary, the storage and rules of the dictionary are highly important.Third, after we prepare the machine dictionary and finish syntax analysis, we can set transformational rules. Transformational rules are serious rules which are set to transform vocabulary sequence after syntax analysis to proper target language. Our application have four main transformational rules

- Verb-predicate-translation choice rule;
- Noun-subject-translation translation rule;
- Translation-position choice rule;
- Auxiliary-word-deletion-or-not rule.

Concrete steps:

a. Find main verb of the word sequence which requires analysis and find its case frame in the verb dictionary.

b. Fill the content accordingly.

Judge the modality by the sign in the sentence.

A. Chunk boundary definition

a. Bi-gram chunk boundary definition

There is mutual information, t-value, $\chi^2$ statistics relevance evaluation function which measure the degree of closeness between words in statistical methods. Because mutual information behaves better than other functions, we can use choose mutual information as the method to get the candidate word chunk [5]. However, there exists some data sparse problems when advantage of point mutual information is taken to get multiword units. Under the same circumstances, mutual information of two-tuple of low-frequency phrases may be bigger than that of two-tuple of high-frequency phrases. This is the reason, there is need of improvement in the relevance functions of mutual information: Collocation [6-7].

$$Collocation\,(w_1, w_2) = \frac{VMI(w_1, w_2)}{H(w_1)+H(w_2)} \quad (1)$$

In this function, H(w) means the entropy of a word, and VMI ($w_1$, $w_2$) means average mutual information. Table 3 shows the result of relevance function.

**Table 4 – Results based on relevance function**

| Relevance function | Average accuracy | Achieve MWU(number) |
|---|---|---|
| mutual information | 81.0 | 24476 |
| $\chi^2$ statistics | 76.0 | 24711 |
| Logarithmic possibility | 53.0 | 40602 |

c. Multi-gram chunk boundary definition

Collocation is designed to calculate Bi-gram. So it is efficient for the chunk recognition. But the chunk may be more than two words, so algorithms need to be recursively called, then it should mark the multi-gram which includes more than two words[8].

To recognize the multi-gram chunk, a window sliding mechanism is used for the sentence which needs analysis.The observation window is set that consists of'K' size of the words. Mutual Information Mean (MIM) and Mutual Information Variance (VMI) are used [9].The MIM and VMI are given by equations (2) and (3).

$$EMI = \frac{1}{C_N^2} \sum_{\substack{0\le i,\\ j\le N}} Collocation(w_i, w_j) \qquad (2)$$

$$VMI = \frac{1}{C_N^2-1}\sqrt{\sum_{\substack{0\le i,\\ j\le N}}(Collocation(w_i, w_j) - EMI)^2} \qquad (3)$$

The meaning of VMI is that the smaller it is, the more stable the combination of the various words in the window is.

B. Inheritance and delivery of grammar attributes

Phrase chunk can be divided into noun phrase chunk, verb phrase chunk and adjective phrase chunk in grammar attributes. Chunk grammar attributes are very important to further natural language process and machine translation. However, the former simple statistical method can't achieve and ensure the chunk grammar attributes and rationality [10].

The multi-word units which constitute a chunk are not any combination of words. From a linguistic point of view, a chunk should have a proper inner grammar structure. So we can take advantage of some specific syntactic pattern rules to filter and delete candidate chunks when we use statistical relevance method. Scott uses part of speech information to achieve candidate multi-word units.Thus, Combination of chunk grammar attributes should improve the accuracy of candidate word chunks. On one hand, we uses phrase rules to filter candidate word chunks which come from statistics methods in order to remove some junk chunks. On the other hand, every word in the chunk can get inheritance and delivery in a proper way, so that it can provide better service for machine translation. In our proposed approach, the rule analysis method is applied to make syntax analysis to phrase chunks and constraint them in grammar so that every word in the chunk should get inheritance and delivery in a proper way, then the new chunk should obey some specific grammar rules.

One example is shown as follows:

We first defines basic noun phrase to: BaseNP.

BaseNP→BaseNP+BaseNP | BaseNP+noun | qualitative+BaseNP | qualitative+noun

Qualitative → adjective | distinguishing | words | adverb | verb | noun | locality category|Englishstring|numeral+quantifier

Thus, the noun phrases are divided into BaseNP and ¬BaseNP. And some typical examples are shown in the Table 4.

**Table 5 – Examples of BaseNP and ¬BaseNP**

| BaseNP | ¬BaseNP |
|---|---|
| laid-off workers, product structure | Complex climate phenomenon |
| study method, space travel | Well-developed economy |
| Enterprise production management | Research and development |

C. Recognition of Bi-gram chunk

Many researchers contributed a lot in this area.

As Dagan and Church [11] designed Termight system.Frank Smadja [12] designed Champollion system. McEnery [13] designed ASMT method. There are two problems trying to achieve multi-word unit translation equivalence pair:

First one is the achievement of monolingual candidate multi-word unit. There are two common ways. One is that we can use grammar rules and language analysis technology. The other is to use statistic methods to make n-unit strings as candidate multi-word units.Second one is how to build the corresponding relationship between bilingual multi-word units. One is to take advantage of word alignment technology, and the other is to calculate the relevance of two languages.Our method is to implement bilingual chunk alignment and build bilingual chunk recognition model based on word alignment technology to overcome the defects of monolingual model. Basic idea is to implement feedback verification and evaluation to improve the accuracy of recognition and choices of bilingual chunks. The concept Fuzzy Matching Degree (FMD0 is lead out that is calculated by the function proposed in[14].

$$FMD = \sum_k \arg\max\left(\frac{2*|WordTsr_{k,i} \cap WordTg_j|}{|WordTsr_{k,i}|+|WordTg_j|}\right) \quad (4)$$

In this function, WordTsr means translation words of the original chunk words in the bilingual dictionary, WordTg means words in the aimed chunk. When FMD is bigger than a certain threshold, we can make a conclusion that the bilingual chunk is aligned.

## Result and Evaluation

To evaluate the effectiveness of proposed RMT algorithm, experiment is conducted on the small scale using bilingual chunk recognition on bilingual corpus. From this bilingual corpora, 71814 bilingual chunks are received. Considering all aspects of limitations, 200 are randomly chosen to perform artificial judgement.First,we check the accuracy of the word chunk recognition. In these 200 chunks, 168 Chinese chunks are identified correctly by using RMT, whereas, Google translation and Kingsoft powerword have 152 and 157 respectively. The candidate word chunk recognition is depicted in Figure 2 can accurately determine as

$$A = CW_{cr} \times \frac{100}{Ch} \quad (5)$$

The candidate word chunk accuracy found to be 168 * 100/200=84%. Then, we check the accuracy of bilingual chunk recognition. We confirm that 145-word chunk are correct, which means when the word chunk recognition is correct, the bilingual chunk recognition accuracy is 145 * 100/168=86.3% depicted in Figure 3. Whereas, Google translation and Kingsoft powerword have 128 and 142 respectively. As, in all these 200 chunks, the accuracy of getting bilingual chunk recognition found to be 72.5%.

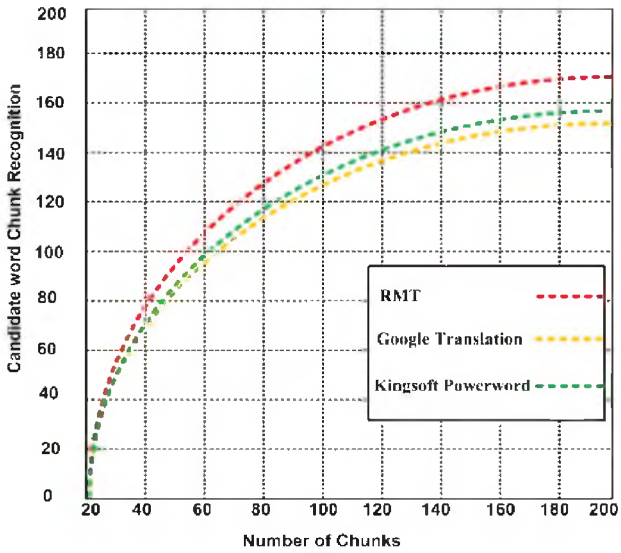


*Figure 2-Candidate word chunk recognition*

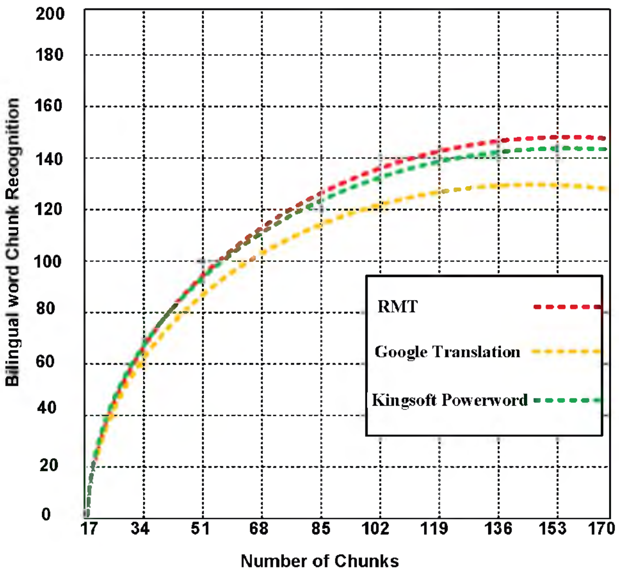


*Figure 3-Bilingual word chunk recognition*

## Conclusion

The translation application based on chunk boundary method has been introduced in this paper. The main idea of this paper is to use rule analysis method for making syntax analysis to make a phrase of chunks and constraints into grammar. The proposed method ensures a proper sequence of words into a sentence according to the meaning of words. The proposed method tries to improve the accuracy of machine translation. Based on the testing, we obtained the results that show the effectiveness of proposed method from accuracy point of view. In the future, there is need to conduct more tests to obtain the results from reliability and efficiency perspective.

## REFERENCES

1. Rieger, C. and Majchrzak, T.A., 2019. Towards the Definitive Evaluation Framework for Cross-Platform App Development Approaches. Journal of Systems and Software.
2. Kolk, Richard, and Abdul Razaque. "Scalable and energy efficient computer vision for text translation." In 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), pp. 1-6. IEEE, 2016.
3. Zhang, Xiangyu, Sihan Tao, Zhitao Gong, Bo Wu, Ruixin Wang, and Bogdan M. Wilamowski. "An improved English to Chinese translation of technical text." In 2015 IEEE 19th International Conference on Intelligent Engineering Systems (INES), pp. 79-83. IEEE, 2015.
4. Yang, Y. Y., B. Z. Li, J. M. Wang, C. S. Yuan, R. Lin, G. M. Lu, and J. L. Wang. "Chinese-English-Yi Public Opinion Information Database Construction and Implementation." In 2016 4th International Conference on Advanced Materials and Information Technology Processing (AMITP 2016). Atlantis Press, 2016.
5. Søgaard, Anders, and Yoav Goldberg. "Deep multi-task learning with low level tasks supervised at lower layers." In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, pp. 231-235. 2016.
6. Nuo, Minghua, Huidan Liu, Congjun Long, and Jian Wu. "Tibetan unknown word identification from news corpora for supporting lexicon-based Tibetan word segmentation." In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), vol. 2, pp. 451-457. 2015.
7. Kumar, Ishan, RenuDhir, and Sanjeev Kumar Sharma. "Various Parsers Available for Indian and Foreign Languages: A Survey." International Journal of Computer Applications & Information Technology 9, no. 1 (2016): 168.
8. Kovář, Vojtěch, VítBaisa, and MilošJakubíček. "Sketch Engine for bilingual lexicography." International Journal of Lexicography 29, no. 3 (2016): 339-352.
9. Liu, Yang, Jiajun Zhang, ChengqingZong, Yating Yang, and Xi Zhou. "A Bilingual Discourse Corpus and Its Applications." (2016).
10. Schneider, Nathan, and Noah A. Smith. "A corpus and model integrating multiword expressions and supersenses." In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1537-1547. 2015.
11. Olohan, Maeve. Scientific and technical translation. Routledge, 2015.
12. Garcia, Marcos, Marcos García-Salido, and Margarita Alonso-Ramos. "Using bilingual word-embeddings for multilingual collocation extraction." In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pp. 21-30. 2017.
13. Camacho-Collados, José, Mohammad Taher Pilehvar, and Roberto Navigli. "A unified multilingual semantic representation of concepts." In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), vol. 1, pp. 741-751. 2015.
14. Van der Lek-Ciudin, Iulianna, Ayla Rigouts Terryn, Geert Heyman, Els Lefever, and Frieda Steurs. "Translator's methods of acquiring domain-specific terminology. Information retrieval in terminology using lexical Knowledge Patterns." In Proceedings of the 21st European Symposium on Languages for Special Purposes (LSP). University of Bergen; Bergen, 2018.