A Cantonese-English Machine Translation System PolyU-MT-99

Yan Wu
Department of Computer, Harbin Institute of Technology
csywu@comp.polyu.edu.hk
Tel:(852)-2766 7328

James Liu
Department of Computing, Hong Kong Polytechnic University
csnkliu@comp.polyu.edu.hk
Tel:(852)-2766 7273

Abstract

In an information society, the application domain for Cantonese-English Machine Translation Systems is getting important day by day. In view of this, we designed and implemented a Cantonese-English Machine Translation System called PolyU-MT-99. The proposed system integrates the rule-based method, understanding-based method and example-based method together. This hybrid approach firstly applies grammar rules to analyze the syntax structure of the source Cantonese sentence, and then compares the input sentence with the example sentences in the bilingual corpus. At last, based on the calculation to compute how the input sentence is similar to those of the example sentences in the corpus, the target language sentence can be generated from the corresponding selected example sentences. In this paper, we present the ideas about the system structure and its design. Some translation examples will be given to highlight the features and the evaluation of PolyU-MT-99. We expect that the prototype system will be further enhanced to cater for real application soon.

System presentation with demo:

The system runs under the PC platform with 300Mhz CPU, 64M RAM and 1G free storage space or above. The operating system is MS Chinese Windows 95, and MS Visual FoxPro is our system development tool. Other peripherals such as monitor, mouse and printer should be compatible with the system configuration.

1 Introduction

In an age of increasing internationalization, machine translation has clear and intermediate attractions. Especially in Hong Kong, the application domain for Cantonese-English Machine Translation is gaining its importance day by day.

There are many methods for designing the machine translation system, such as rule-based method, knowledge-based method and example-based method (Sito, and Nagao, 1990; Tsujii, 1986; Brown, 1997). In recent years, with the generation of bilingual corpus, an example-based method is another alternative rather than rule-based method (Zhou, Liu, and Yu, 1997; Mclean, 1992). Because the large-scale and high-quality bilingual corpus is seldom readily available, let alone example-based method encounters a lot of problems in machine translation, such as insufficient example sentences and the redundancy of example sentences. While rule-based method can solve the problems that the example-based method often encounters, it not only reduces the scope of matching similar sentences but also increases the accuracy of matching source sentence of example-base method. On the other hand, the example-based method can solve effectively the problem of insufficient knowledge that the rule-based method often comes across during the translation process (Chen, and Chen, 1995).

In view of this, a machine translation prototype system

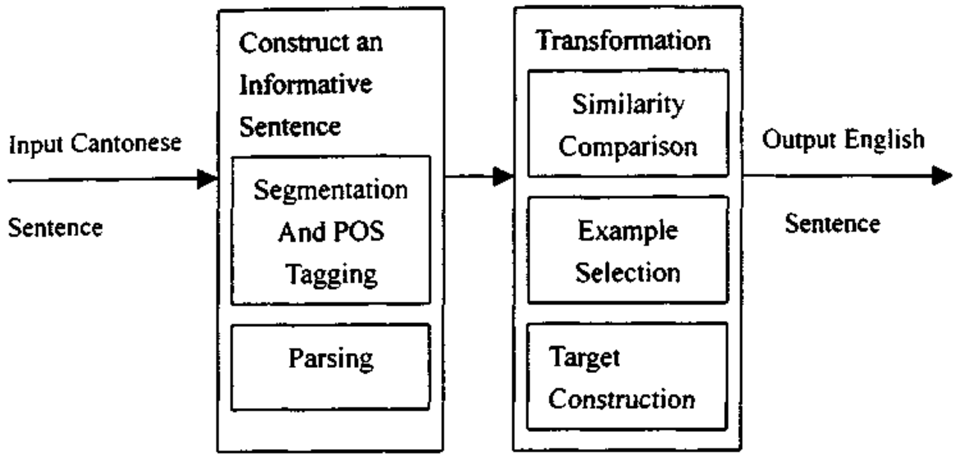

Figure 1. The architecture of PolyU-MT-99 system

PolyU-MT-99 is designed and implemented based on rules, understanding and a corpus of example sentences.

2 Design Constructs

The design of PolyU-MT-99 system is as follows:

1) Applying the rule-based method to analyze the source sentence and generating its phrase structure. The rule base of this system is established through analyzing the real corpus.

2) Applying the example-based method to convert and generate the target sentence.

3) The principle of classifying the Cantonese substantive is not only based on syntactic features of the word but also the semantic features of it; for the function word, such as "的", "被" and "因此", the principle for its classification is only based on its syntactic features.

4) The understanding model of the system includes two parts: word model and phrase model. The word model contains six parts: Cantonese word, category, frequency, corresponding English word$_1$, corresponding English word$_2$, and corresponding English word$_3$. The phrase model has the same structure with the word model. Table 1 shows the examples of these two models, where "*d*", "*c*" and "*v*" represent adverb, conjunction and verb respectively.

5) The example model includes four parts: Cantonese sentence, tagged Cantonese sentence, corresponding English sentence, and tagged corresponding English

6) The system is portable and extendable. Its dictionary, rule base and algorithm are in separate modules that can be maintained independently.

7) The translating object of this system is of Cantonese expression.

Table 1 The examples of understanding model

Attribute	Example1	Example2
Cantonese word	只是	指日可待
Category	d, c, v	v
English word1	only	Can be expected soon
English word2	however	
English word3	be only	
Frequency	0.02416	0.00046

3 Implementation

The implementation of PolyU-MT-99 system is composed of five parts: bilingual corpus, dictionaries, rule base, main program and additional function modules. The programming language of the system is in MS Visual FoxPro. The architecture of PolyU-MT-99 is shown in Figure 1.

3.1 Segmentation algorithm and POS tagging

Suppose $S=c_1c_2...c_n$ is a Cantonese character string, and there are m kinds of possible segmentation $S_1=w_{11}w_{12}\cdots w_{1n_1}$, $S_2=w_{21}w_{22}\cdots w_{2n_2}$, ……, and

$S_m = w'_{m1} w'_{m2} \cdots w'_{mn_m}$. We introduce the following mathematical model for selecting the appropriate segmentation:

$$f(w'_{i1} w'_{i2} \ldots w'_{in_i}) = \max_{i=1,m} \prod_{j=1}^{n_i} f\left(w'_{ij}\right).$$

Here $f\left(w_{ij}\right)$ is the frequency of word w'_{ij}.

We use dijkstra algorithm to implement the segmentation, the Cantonese words are tagged with POS.

For example. the input sentence “他是一個學生” can be segmented and tagged as “他/r 是/v 一/m 個/q 學生/n”; the input sentence “乒乓球拍賣完了” can be segmented and tagged as “乒乓球拍/n 賣/v 完/a 了/u”. Here ‘a’, ‘u’ 、 ‘r’、 ‘v’、 ‘m’、 ‘q’ and ‘n’ denote adjective, auxiliary, pronoun. verb, numeral, quantifier and noun respectively.

3.2 Parsing

The function of parsing is to identify the syntactic structure of a sentence. At this stage, both the input and output sentences are parsed.

This procedure works with some paring rules which can be generated from the corpus. These rules are like:

$S \Longrightarrow NP.VP$

$NP \Longrightarrow$ *adjective . noun || article . noun ||...||noun*

The sentence is scanned from its end: i.e. the last two words of the sentence will be checked first. then the next two prior to that and so on till to first word of the sentence.

After parsing. the system only needs to match out the POS. This procedure can reduce the searching time to find out a high score sentence.

For example. a tagged Cantonese sentence “他/r 足/v 一/m 個/q 學生/n” is parsed as “S=[他/r]NP[足/v[一/m 個/q 學生/n]NP]VP”. Its parsing tree is shown in Figure 2.

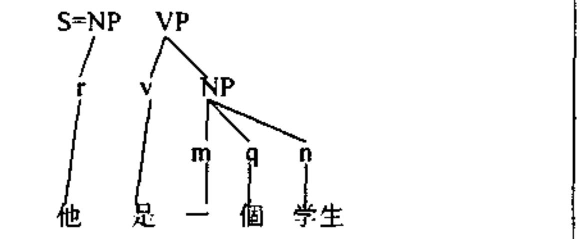

Figure 2. The parsing tree of a sentence

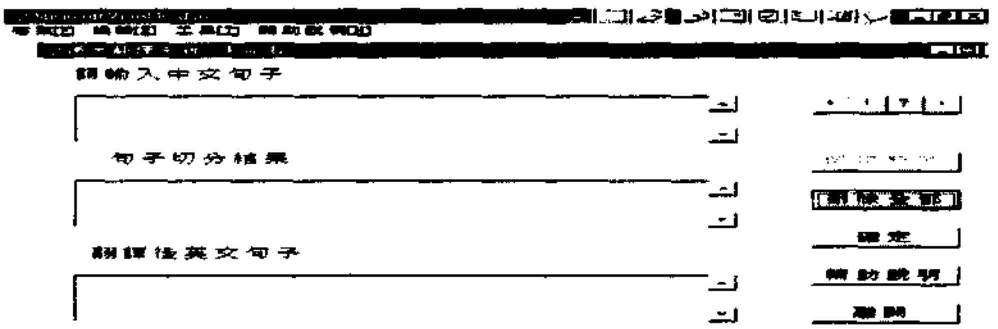

Figure 3. The interface of PolyU-MT-99 system

3.3 Similarity comparison and example selection

After the input sentence has been parsed, the system then searches for the sentence from the bilingual corpus, which is the most similar one to the input. The calculation of the similarity of two sentences operates on a phrase basis in the parsed input sentence and the parsed example sentence. The part-of-speech within the same phrase, in the phrase structure pattern of the input sentence and of each example sentence in the bilingual corpus are compared. In case of a mismatched between their part-of-speech, a penalty score is incurred and the comparison proceeds for the next part-of-speech within the same phrase. The score calculation progresses from the left-most phrase structure until the last one of the sentence. The mathematical model of this procedure is as follows:

Suppose $S_1=a_1a_2......a_n$, $S_2=b_1b_2......b_m$, $a_i(b_j), 0<i<n+1, (0<j<m+1)$ is the *ith* (*jth*) phrase in sentence S_1 (S_2). $a_i=w_{i1}w_{i2}.....w_{ik}$, $b_j=w_{j1}w_{j2}.....w_{jl}$, where w is a Cantonese word. F is the whole feature set of a certain word category, E a subset of F, and |E| stands for the number of features in E. *fea$_k$(w), sub_pos(w) and pos(w)* represent the *kth* feature, sub-category and part-of-speech of word *a* respectively. *Ss(S$_1$,S$_2$)* represents the similarity metric between S_i and S_2. *Sp(a$_i$,b$_j$)* is the similarity score between phrase a_i and b_j, while $S_w(w_i,w_j)$ the similarity score between w_i and w_j. a_1^i (b_1^j) represents the phrase string from a_1 (b_1) to $a_i(b_j)$.

$$Ss(S_1,S_2)=Ss\left(a_1^n,b_1^m\right) \qquad (1)$$

$$Ss(a_1^i,b_1^j)=\begin{cases}0, \; if\, i<1 \cup j<1 & (2)\\ Sp(a_1,b_1)+Ss(a_2^i,b_2^j), otherwise\end{cases}$$

$$Sp\left(a_i,b_j\right)=\begin{cases}-1, \; if\, category(a_i)\neq category(b_j) \\ & (3)\\ Sp(w_{i1}^{k-1},w_{j1}^{l-1})+Sw(w_{ik},w_{jl}), otherwise\end{cases}$$

$$Sw(w_{ik},w_{jl})=$$

$$\begin{cases}1, \; if\; w_{ik}=w_{jl}\\ 0.8, \; else\; if\; pos(w_{ik})=pos(w_{jl})\\ 0.5, \; elseif\; sub\_pos(w_{ik})=sub\_pos(w_{jl})\\ 0.25, \; elseif \displaystyle\bigcup_{\substack{fea_{k1}\in E\\ E\subset F\\ E\le 0.5^* F}} fea_{k1}(w_{ik})=fea_{k1}(w_{jl}) \quad (4)\\ 0.4, \; elseif \displaystyle\bigcup_{\substack{fea_{k1}\in E\\ E\subset F\\ 0.5^* F<E<F}} fea_{k1}(w_{ik})=fea_{k1}(w_{jl})\\ -1, \; Otherwise\end{cases}$$

This procedure calculates the score between input sentence and every sentence of the example base, and selects the example sentence whose score is the highest as the best-matched sentence. If an input sentence matches both a fragment and a full sentence that contains (or does not completely contain) the fragment, or matches two examples that are syntactically identical but lexically different, and so on, the highest score of the example sentence will be selected.

The example base has about 9000 Cantonese and English sentence pairs, and both sides have been annotated with part-of-speech. Moreover, many sub-dictionaries of noun, verb, adjective, pronoun, classifier, and preposition, etc, are employed. There are many specific features helpful for sentence comparison in each of these dictionaries.

For a parsed Cantonese sentence "S=[他/r]np[足/v[一/m 個/q 學生/n]np]vp", the example sentence could be "S=[她/r]np[足/v[一/m 個/q 工人/n]np]vp".

3.4 Target construction

This stage involves the use of Cantonese and English phrase structure relations of the example translation as a template to build the target English sentence. This procedure can be presented as follows:

Determining which words or phrases of the input sentence are different from those of the selected example Cantonese sentence in the same order and position. This procedure scans both the input sentence, and example sentences from the beginning to the end of the corpus. After then, the positions and words that are different will be recorded.

1) For example,
   Input sentence: S=[他/r]NP[足/v[一/m 個/q 學生/n]NP]VP
   Example sentence: S=[她/r]NP[足/v[一/m 個/q 工人/n]NP]VP
   "他" and "她" are recorded.
2) Replacing the different parts of example Cantonese sentence and its English sentence with variables, and forming the translation template.
3) Looking for different parts in the bilingual dictionary, and getting their English meanings.
4) Replacing variables of the translation template with corresponding parts of the input sentence.

5) Adjusting the metaplasm of translation template, such as verb tense and adjective comparative degree. For example,
Input sentence: 他學習英語
Example sentence: 你學習英語 (you learn English)
The system replaces the verb “learn” with “learns”.

6) Adjusting the consistency of pronoun.

For the similar example sentence “s=[她/r]np[是/v[一/m 個/q 工人/n]np]vp”, its English sentence “she is a worker” is used as the template. Then the English translation of the source sentence “牠是一個學生” is “he is a student”

4 Experimental results

PolyU-MT-99 system has been realized with MS Visual FoxPro for Windows. The system interface is shown in Figure 3. Users can easily interact with the system to perform the translation. Table 2 and table 3 displays the part of experimentation

Table 2 The experimental results I

Test target	Input sentence	Example sentence	result
Testing the sentence similarity	1 手放在口袋裡的男孩正在踢足球.	手放在口袋裡的男孩正在踢足球.	The boy with his hands in his pockets are playing football.
	2. 手放在肩上的男孩正在踢足球.	手放在口袋裡的男孩正在踢足球.	The boy with his hands in his shoulder is playing football.
	3. 腳放在桌上的男孩正在看書.	手放在口袋裡的男孩正在踢足球.	The boy with his foots in the desk is reading book
Testing the irregular verbs for past tense	1.更多業内人士讀了這個規定.	學生們借了你的茶壺 . (students borrowed your teapot)	Most of professional persons read the rule.
	2. 她去過北京.	我們去過香港(we have gone to Hong Kong)	She has gone to Beijing.

Table 3 experimental results II

Test target	Input sentence	Example sentence	result
Testing the coherence between subject and verb	1.香港公證會正式獨立.	他們正式獨立. (they are formally independent)	The notarial association of Hong Kong is formally independent
	2. 她住在香港.	工人們住在中國.(workers live in China)	She lives in Hong Kong.
	3. 物價因應市場反應而增減.	人們因應季節變化而換裝. (People change their dressing along with the season.)	The price increases or decreases along with the market reaction.

The experimental results indicate that the correction of translation of the system is 75%. Most of the translation errors are caused by the following cases:

1) The preposition and noun of sentences are misplaced, as shown from the third example English sentence in Table 2. The corrected translation of “在桌上” is “on the desk”, and not “in the desk”.

2) Errors in plural noun. For the Cantonese sentence “牠拿來兩個刀子”, its English translation is “he brings two knifes”. The corrected translation should be “he brings two knives”. Building the irregular noun table can solve this error.

3) Some Cantonese phrasal word has no corresponding English word. Insufficient Knowledge base is the usual problem in most of natural language processing.

4) Segmentation error is also one of the reasons that causes the translation errors. For example, 是/非常/常/混淆”, 她/是/非常/漂亮/的/”.

5) POS error is the other reason that causes the translation errors. Because the POS tagging is only statistic-based and it selects those categories that often occur in the corpus. For example, 書/n 在/p 桌/n 上/u”, 牠/r 上/u 山/n”. This type of

error can be solved by syntactic analysis.

5 Evaluation

We have tested 1000 Cantonese sentences that were originated from Mingpao newspaper of Hong Kong. The translation of the complicated input sentence is not perfect. The main problem occurs in the procedure for similarity comparison of example sentences. Nevertheless, PolyU-MT-99 system has the following features:

- The system has friendly human-machine interface.
- The system is simple, portable and extendable.
- The translating result of system is readable and understandable.
- The system combines effectively the advantages of rule-based method and example-based method.

Machine translation especially in the Cantonese–English domain is quite a demanding and difficulty subject. Through the research of PolyU-MT-99 system, we have discussed the example-based machine translation method, and believe that this method is feasible to solve many translating problems. It is possible and effective to acquire the bilingual knowledge from the small-scaled, representative, parallel Cantonese-English corpus. We have suggested a series of algorithms, such as Cantonese segmentation, bilingual comparing algorithm and target sentence construction algorithm. We have created some databases, which contain mass of Cantonese words and related information. For example, the Cantonese dictionary keeps the information of part-of-speech, word frequency. Bilingual corpus stores mass of Cantonese-English sentence pairs that have been segmented and tagged with POS. The bilingual dictionary stores the Cantonese words and corresponding English words. This information can be used for later development or other NLP.

Acknowledgement

The authors appreciate the support from the Central Research Grant #G-YB11 of the Hong Kong Polytechnic University.

References

Brown, R.D. (1997). "Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation". In Proceedings Of the seventh International Conference on Theoretical and Methodological Issues in Machine Translation, Santa Fe, pp. 23-25.

Chen K.H. and Chen H.H. (1995). "Machine Translation: An Integrated Approach". In Proceedings Of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 287-294.

Markman, B.A. & Gentner, D. (1996). "Commonalties and Differences in Similarity Comparisons". Memory and Cognition, 24(2), pp. 235-249.

Mclean, I. (1992). "Example-based Machine Translation Using Connectionist Matching". In Proceedings Of TMI-92, Montreal, pp. 35-43.

Sato, S. and Nagao, M. (1990). "Toward Memory-based Translation". In Proceedings of 13th International Conference on Computational Linguistics, Helsinki, pp. 247-252.

Tsujii, J. (1986). "Future Directions of Machine Translation". In Proceedings of 11th International Conference on Computational Linguistics, Bonn.

Zhou, L.N., Liu, J. and Yu, S.W. (1997). "Bilingual Parsing from the Viewpoint of Alignment". In 2nd Workshop ON Multilinguality in Software Industry, Japan, pp. 65-72.