Computational Linguistics and Chinese Language Processing
Vol. 11, No. 2, June 2006, pp. 137-158

A Structural-Based Approach to Cantonese-English Machine Translation

Yan Wu*, Xiukun Li* and Caesar Lun+

Abstract

In this paper, we present an integrated method to machine translation from Cantonese to English text. Our method combines example-based and rule-based methods that rely solely on example translations kept in a small Example Base (EB). One of the bottlenecks in example-based Machine Translation (MT) is a lack of knowledge or redundant knowledge in its bilingual knowledge base. In our method, a flexible comparison algorithm, based mainly on the content words in the source sentence, is applied to overcome this problem. It selects sample sentences from a small Example Base. The Example Base only keeps Cantonese sentences with different phrase structures. For the same phrase structure sentences, the EB only keeps the most simple sentence. Target English sentences are constructed with rules and bilingual dictionaries. In addition, we provide a segmentation algorithm for MT. A feature of segmentation algorithm is that it not only considers the source language itself but also its corresponding target language. Experimental results show that this segmentation algorithm can effectively decrease the complexity of the translation process.

Keywords: Example-Based Machine Translation (EBMT), Rule-Based Machine Translation (RBMT), Example Base (EB).

1. Introduction

Although Machine Translation has been an important research topic for many years, the development of a useful Machine Translation system has been very slow. Researchers have found that developing a practical MT system is a very challenging task. Nevertheless, in our age of increasing internationalization, machine translation has a clear and intermediate

* Department of Computer, Harbin Institute of Technology, Harbin 150001, China
Phone Number: (00852) 95810688 Fax Number: (00852) 2626 1771
E-mail: wy98hk@yahoo.com
+ Department of CTL, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
E-mail: ctslun@cityu.edu.hk

[Received July 19, 2005; Revised February 21, 2006; Accepted February 25, 2006]

attraction.

There are many methods for designing machine translation systems [Carl 1999; Carpuat 2005; Kit 2002b; Mclean 1992; Mosleh and Tang 1999; Somers 2000; Knight and Marcu 2005; Tsujii 1986; Brown 1997; Zhou *et al.* 1998; Zens 2004], such as the rule-based method, knowledge-based method, and example-based method. In recent years, with the development of bilingual corpora, the example-based method has become a better choice than the rule-based method, although statistical MT systems are now able to translate across a wide variety of language pairs [Knight and Marcu 2005]. This is because the rule-based MT system has some disadvantages, such as a lack of robustness and poor rule coverage [Zhou and Liu 1997]. On the other hand, the large-scale, high-quality bilingual corpora are seldom readily available, so the example-based method has encountered a lot of problems in machine translation, such as a lack of sufficient example sentences and redundant example sentences. The good performance of an EBMT system depends on there being a sentence in the example base which is similar to the one that is to be translated. In contrast, an SMT system may be able to produce perfect translations even when the sentence given as input does not resemble any sentence in the training corpus. However, such a system may be unable to generate translations that use idioms and phrases that reflect long-distance dependencies and contexts, which are usually not captured by current translation models [Marcu 2001]. On the other hand, the example-based method can effectively solve the problem of insufficient knowledge that the rule-based method often encounters during the translation process [Chen and Chen 1995]. In view of this fact, a machine translation prototype system, called LangCompMT05, has been implemented. It integrates rule features, text understanding, and a corpus of example sentences.

In this paper, a brief review of the MT method is given first. This is followed by an introduction to the framework for LangCompMT05. In section 3, a detailed description of this system, whose implementation involves combining example-based and rule-based methods, is presented. Experimental results are discussed in section 4. The last section gives conclusions and discusses future work.

2. Design Constructs

Figure 1 shows the architecture of the LangCompMT05 system.

The implementation mechanism of the LangCompMT05 system is as follows:

1) The source Cantonese sentence is segmented with a new segmentation algorithm, whose implementation is based on the word frequency, and the criterion for segmentation considers not only the source sentence itself but also its corresponding translation. The source sentence "*她有些神經過敏*" (She is a little bit hypersensitive), for example, can

be segmented as "*她/有些/神經/過敏*" in general. Because "*神經過敏*" can be translated into the English word "hypersensitive", for MT, the sentence is segmented as "*她/有些/神經過敏*".

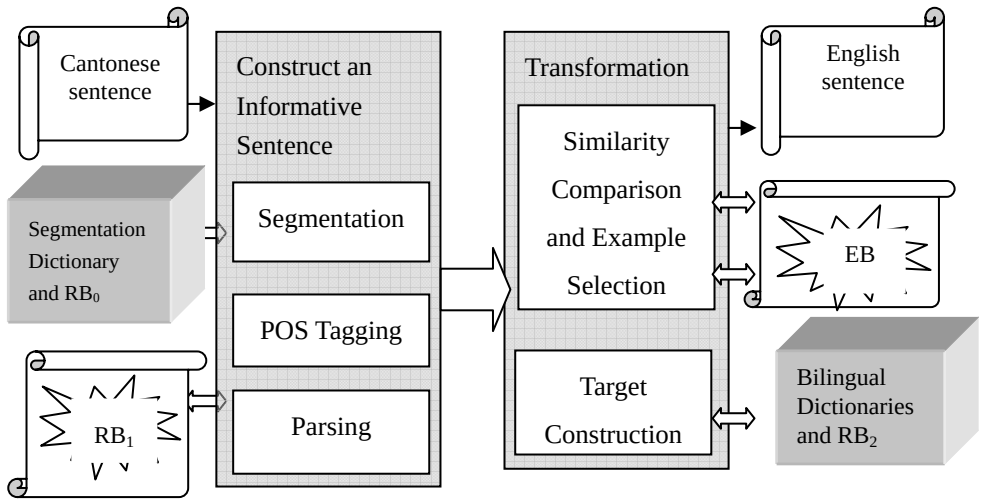

Figure 1. The architecture of the LangCompMT05 system

2) The rule-based method is applied to analyze the source sentence, and its phrase structure is generated. The Rule Base (RB) of this system is established through analysis of the real corpus. The phrases are classified as noun phrases (NPs) or verb phrases (VPs). Some of the rules for phrases are as follows:

NP= : [a] [n] | [m] (q) (n),

VP= : [d] (v) .

Here, "a", "n", "m", "q", "d", and "v" denote adjective, noun, numeral, quantifier, adverb, and verb, respectively.

3) A new knowledge representation, called SST, is applied to store the sentence structure. The target sentence can be generated with this tree.

4) The example-based method and rule-based method are combined and used to select, convert, and generate the target sentence.

5) The principle for classifying a Cantonese content word, such as "單車 (bike)" or "*返工 (go to work)* ", is dependent not only on the syntactic features of the word but also its semantic features; for a function word, such as "*的*", "*被*", or "*因此 (so)*", the principle for classification is only based on its syntactic features.

6) The understanding model of the system includes two parts: a word model and a phrase model. Both of them consist of six parts: a Cantonese word, a category, a frequency, and three corresponding English words: word1, word2, and word3. The phrase model has the same structure as the word model. Table 1 show examples of these two models, where "*d*", "*c*", and "*v*" represent adverb, conjunction and verb, respectively.

Table 1. Examples of understanding models.

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

7) The example model consists of four parts: a Cantonese sentence, a tagged Cantonese sentence, a corresponding English sentence, and a tagged corresponding English sentence.

8) The system is portable and extendable. Its dictionaries, rule bases, and algorithms are in separate modules (see Figure 1) that can be maintained independently.

9) The system can translate written Cantonese into English.

3. Implementation

The implementation of the LangCompMT05 system is composed of the following parts: an example base, dictionaries, rule bases, the main program and five additional function modules (see Figure 1). It integrates rule features, text understanding, and a corpus of example sentences. For the preprocessing stages, it uses a rule-based method to deal with the source sentence. Then, the EBMT method is used to select the translation template. In the target sentence construction stage, which involves the translation of sentence components, the system is mostly based on a rule-based method.

3.1 Segmentation Algorithm

Word segmentation is the basic tack in many word-based applications, such as machine translation, speech processing, and information retrieval. Chinese word segmentation, being an interesting and challenging problem, has drawn much attention from many researchers [Hu 2004; Kit 2002a; Dunning 1993; Hou 1995; Liu 1994; Nie 1995]. We will present the segmentation algorithm in detail in another paper.

3.2 POS Tagging

Parts of speech can help us analyze the syntax structure of a sentence, and they are fundamental to the understanding and transformation of MT. A knowledge base and rules are used to tag each Cantonese sentence.

The knowledge base consists of records that contain words and their parts-of-speech. After segmentation, all of the words in the source sentence are tagged. For ambiguous words that have more than one part-of-speech, the rules in RB_0 are used to perform disambiguation.

Suppose $T = \{n, np, m, q, r, v, a, p, w, d, u, f, c, t, b, g\}$ is the tag set of the system, and *A* is the set of all Cantonese words. The formal presentation of the disambiguation rules is as follows:

$$\begin{aligned} &\alpha \aleph \beta \to \alpha \ell \beta, \\ &\alpha, \beta \in \{A \cup T\}^*, \\ &\aleph \subseteq T, \\ &\ell \in T. \end{aligned} \tag{1}$$

Here, χ is the subset of POS set *T*, ℓ is the element of *T*, and α and β are null, a Cantonese word or an element of *T*. $\to$ denotes that if an ambiguous word that has the POS χ is preceded by POS α and succeeded by POS β, then it can be tagged as ℓ. For example, the POS rule ($m\{u,n\} \to mn$) means that if a word has the property of an auxiliary word (*u*) or a noun (*n*) and is preceded by a quantifier, then it is a noun.

The following is an example of this process:

兩/m 地/(u,n)相距/n 三/m 哩(u,q) $\xrightarrow{m\{u,n\}\to mn,\ \ m\{u,q\}\to mq}$ 兩/m 地/n 相距/n 三/m 哩/q (The distance between the two locations is 3 miles)

他/r 騎/v 單車/n 追/v 上來/(u,v) $\xrightarrow{v\{u,v\}\to vu}$ 他/r 騎/v 單車/n 追/v 上來/u (He catches up by bike)

她/r 終於/d 上來/(u,v)了/u $\xrightarrow{d\{u,v\}u\to dv}$ 她/終於/d 上來/v 了/u (Finally, she comes up)

3.3 Parsing

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

This procedure works with some paring rules that have been generated from the corpus. These rules in RB_1 include the following:

S → *NP.VP*,

NP → *adjective . noun || article . noun ||...||noun*.

The sentence is scanned backwards from the end; i.e. the last two words of the sentence are checked first, then the next two prior words, and so on till the first word of the sentence is scanned.

After parsing, the system only needs to match out the POS. This procedure can reduce the searching time needed to identify the most similar example sentence in the EB.

For example, a tagged Cantonese sentence *他/r 是/v 一個/q 學生/n (He is a student)* is parsed as *S=[他/r]NP[是/v[一個/q 學生/n]NP]VP*. Its parsing tree is shown in Figure 2.

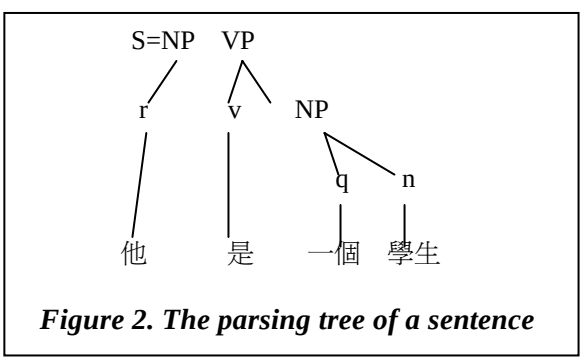

Figure 2. The parsing tree of a sentence

After parsing, the sentence is converted into SST as shown in Figure 3.

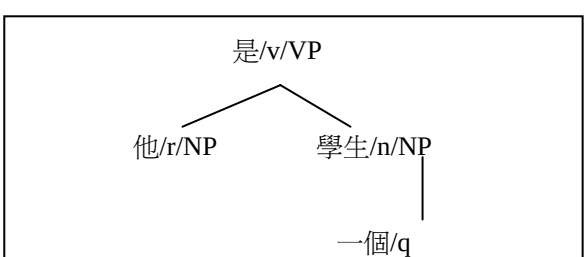

Figure 3. An example of SST

Definition 3. SST is a Binary Tree; it is used to store the natural language sentence. Let $s=w_1w_2...w_n$ be a sentence:

1) w_i is a root if and only if w_i is the center word of the predicate in the sentence.

2) $w_1...w_{i-1}$ forms the left sub-tree of the root, while $w_{i+1}...w_n$ forms the right sub-tree of the root.

3) The left sub-tree and the right sub-tree are formed as follows:

   a) If $w_1...w_{i-1}$ or $w_{i+1}...w_n$ is a sub-sentence, then go to 1).

   b) If $w_1...w_{i-1}$ or $w_{i+1}...w_n$ is a phrase, then the root of the sub-tree is the center word (or content word), while the following word is the modifier of the center word.

This type of knowledge representation can easily reflect the structure of a sentence, and can be implemented for the translation process.

3.4 Similarity Comparison and Example Selection

In general, an example-based MT system should address the following problems:

1) building the map relation of bilingual alignment, based on characters, words, phrases, sub-sentences or sentences;

2) similarity calculation and example selection;

3) constructing a target.

Among these problems, problem 2 is the most important one in example-based MT. Many researchers have focused on the above problems [Li 2005; Chen 2002; Church 1994; Fung 1993; Carl 1999; FuRusE 1992; Mosleh 1999; Carl 1999] and tried to solve it in different ways.

For problem 2, our research addresses three important questions as follows:

1) *Determining the matching level*:

The matching level includes the sentence level and sub-sentence level. For the former, it is easy to determine the boundary of a sentence. Because the sentence can contain a certain number of messages, the possibility of having an exact match is very low, so the system lacks flexibility and robustness. In contrast, matching at the sub-sentence level has the advantage of exact matching and the disadvantage of boundary ambiguity.

In addition, there are no exact chunking or cover algorithms. Our matching algorithm is sentence-based.

2) *The algorithm for calculating the similarity*:

There is no exact definition for the similarity between sentences. Many researchers have addressed this issue and presented similarity algorithms based on words. Some of the algorithms [e.g., Sergei 1993] firstly calculate the word similarity according to the word font, word meaning, and semantic distance of words, and then calculate the sentence

similarity based on word similarity. Other algorithms [Brown 1997; Carl 1999; Markman *et al.* 1996; Mclean 1992; Mosleh *et al.* 1999; Zhang *et al.* 1995] are based on syntax rules, characters and hybrid methods.

Our similarity algorithm is based on the phrases in the sentence; it has the following features:

a) The example base consists of a variety of sentences whose phrase structures are different.

b) The phrases of a sentence are the fundamental calculating cells for aligning the content words of the input sentence and example sentence, i.e., calculate the similarity between the same positional phrase in the input and example sentence. For example:

更多業內人士 /NP *讀了/VP* *這個規定 /NP* *(More professional people have read the regulation.)*

學生們/NP *借了/VP* *你的茶壺 /NP (Students borrowed your teapot.)*

For the same positional phrases, the similarity calculation is based on the content words. This is based on the principle that in a natural language sentence, the content words form the framework of the sentence and depict the central meaning of the sentence.

c) The system does not need lexical, syntax, and semantic analysis to perform similarity comparison.

d) The system can deal with a variety of Cantonese inputs, such as sentences, sub-sentences, and phrases.

3) *The efficiency of this algorithm*:

Normally, there will be a lot of example sentences in the example base. The algorithm proposed here has to calculate the similarity between the input sentence and every sentence in the example base. So the efficiency of the algorithm is very important.

The example base contains the different structures of Cantonese sentences. For sentence with the same structure, we select the shortest one as an example sentence. So the example base will keep the smallest number of sentences yet maintain the largest number of sentence structure types. In addition, the similarity algorithm is not recursive, and it saves computing time.

3.4.1 The Example Base

Each translation example in the example base consists of four components: a Cantonese sentence, a tagged Cantonese sentence, an English sentence, and a tagged English sentence. A Cantonese-English translation example is given as follows:

> *他騎單車返工。; 他/r 騎/v 單車/n 返工/v。/w; he goes to work by bike. he/He goes to work/V by/P bike/N ./W;*

In the example base, the four components of an example sentence have no relationship with each other and don't need to align Cantonese to English sentences. All the Cantonese sentences in the example base are segmented and tagged. Cantonese segmentation is based on English translation, i.e. if the English translation is a phrase; then the corresponding Cantonese part is segmented as a word, such as "*返工*". This part of the English sentence serves as a translation template, the tagged Cantonese sentence and tagged English sentence are to construct a target (see section 3-5).

3.4.2 Similarity Comparison

Similarity comparison is used to choose the most similar Cantonese example sentence in the example base with the input sentence, and then its corresponding English translation sentence will serve as the translation template to translate the input Cantonese sentence. The similarity of two sentences is calculated on the basis of a phrase in the parsed input sentence and the parsed example sentence. The parts-of-speech within the same phrase, in the phrase structure pattern of the input sentence, and in each example sentence in the bilingual corpus are compared. In case of a mismatch between the parts-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 to the last one of the sentence.

In fact, the similarity comparison mechanism is mainly based on the content words in the sentence. The example base can only store Cantonese framework sentences. For sentences that have the same phrase structure, the shortest is stored in the example base so as to avoid information redundancy in the example base. The mathematical model of this procedure is as follows [Wu and Liu 1999; Zhou and Liu 1997]:

Suppose A=$w_1w_2...w_n$=$p_{A1}p_{A2}.....p_{Ak}$, B=$w_1w_2...w_m$=$p_{B1}p_{B2}.....p_{Bl}$, where $w_{Ai}(w_{Bj})$, $p_{Ai}(p_{Bj})$ is the i^{th} (j^{th}) Cantonese word and phrase, respectively, in sentence *A (B)*. F is the whole feature set of a certain word category, *E* is 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 k^{th} feature, sub-category, and part-of-speech of word *w*, respectively. $Ss(S_1,S_2)$ represents the metric between S_1 and S_2;

$$Ss(S_1, S_2) = \sum_{i=1}^{\max(k,l)} Sp\left(p_{Ai}, p_{Bi}\right), \tag{2}$$

$$Sp(p_{Ai}, p_{Bi}) = \begin{cases} -len(p_{Ai}),\ if\ \ len(p_{Bi}) = 0 \\ -len(p_{Bi}),\ if\ \ len(p_{Ai}) = 0 \\ Sw(p_{Ai}^c, p_{Bi}^c) + Sw(p_{Ai}^f, p_{Bi}^f) \end{cases}, \tag{3}$$

$$Sw(p_{Ai}^f, p_{Bi}^f) = \begin{cases} 1.5,\ if\ \ p_{Ai}^f = p_{Bi}^f \\ 1.1,\ if\ \ POS(p_{Ai}^f) = POS(p_{Bi}^f) \\ -0.3,\ if\ \ \left(len(p_{Ai}^f) = 0\ AND\ len(p_{Bi}^f) <> 0\right) \\ \qquad\qquad OR\ \left(len(p_{Ai}^f) <> 0\ AND\ len(p_{Bi}^f) = 0\right) \\ -0.6,\ otherwise \end{cases}, \tag{4}$$

$$Sw(p_{Ai}^c, p_{Bi}^c) = \begin{cases} 1.5,\ if\ \ p_{Ai}^c = p_{Bi}^c \\ 1.2,\ if\ \ POS(p_{Ai}^c) = POS(p_{Bi}^c)\ and \\ \qquad \bigcup_{\substack{fea_{k1} \in E \\ 0.5*|F| < |E| < |F|}} fea_{k1}(p_{Ai}^c) = fea_{k1}(p_{Bi}^c) \\ 1.1,\ if\ \ POS(p_{Ai}^c) = POS(p_{Bi}^c)\ and \\ \qquad \bigcup_{\substack{fea_{kj} \in E \\ 0.5*|F| \geq |E|}} fea_{k1}(p_{Ai}^c) = fea_{k1}(p_{Bi}^c) \\ 1.0,\ if\ \ POS(p_{Ai}^c) = POS(p_{Bi}^c) \\ 0.8,\ if\ \ POS(p_{Ai}^c) \neq POS(p_{Bi}^c)\ and\ POS(p_{Ai}^c) \in \{n, r\}\ and \\ \qquad POS(p_{Bi}^c) \in \{n, r\} \\ 0.6,\ when\ the\ words\ before\ p_{Ai}^c\ and\ p_{Bi}^c\ is\ function\ words,\ and \\ \qquad they\ are\ not\ equal,\ and\ p_{Ai}^c = p_{Bi}^c \\ 0.4,\ when\ the\ words\ before\ p_{Ai}^c\ and\ p_{Bi}^c\ is\ function\ words,\ and \\ \qquad they\ are\ not\ equal,\ and\ POS(p_{Ai}^c) = POS(p_{Bi}^c) \\ -1.5,\ otherwise \end{cases}. \tag{5}$$

$Sp(p_{Ai},p_{Bi})$ is the similarity score between phrases p_{Ai} and p_{Bi}; p^c_{Ai}, p^c_{Bi} are the content words in phrases Ai and Bi respectively; and p^f_{Ai}, p^f_{Bi} are the function words in phrases A_i and B_i, respectively; and $len(p_{Ai})$ and $len(p_{Bi})$ are the total number of words contained in phrase p_{Ai} and p_{Bi}, respectively.

We set the weights in equations 4 and 5 based on the results of many experiments. We think that the function word and content word have the equal function in the comparison of sentences, so they have the same similarity score, i.e. 1.5. In equation 4 (for function words), if the parts-of-speech of the function words in A_i and B_i are equal, we think we can simply exchange the function word in the example sentence with the source function word, which will not affect the translation sequence. In this case, we give the higher similarity score of 1.1. If there is a function word in A_i, and no function word in the corresponding location in B_i, we think the structures of both A_i and B_i are not equal, so we assign a negative similarity. Otherwise, the function words of A_i and B_i are totally different, so the lower negative weight is given. Equation 5 is used to calculate the content word similarity. All content words have their own semantic features, which can be used to calculate their similarity. If the parts-of-speech of the content word in Ai and B_i are equal, and if most of their features are equal, then we give the higher similarity weight, 1.2; otherwise, their identical features are less than half of the whole feature set F, and we think they belong to different categories, so we assign a weight of 1.1. If their features are totally unequal and their POSs are equal, we think the difference between A_i and B_i is semantic, so the weight is 1.0. If the parts-of-speech of the content words of A_i and B_i are not equal and belong to (n,r), we think this difference doesn't affect the translation sequence, so the weight is 0.8. When the content words of A_i and B_i are equal and the function words before them are not equal, we think this may affect the translation result, so a 0.6 weight is given. If the POSs of the content words in A_i and B_i are equal and the function words before them are not equal, we think their similarity is low, so the weight is 0.4. Otherwise, they are totally different. Because the content word plays the main function in determining meaning of the sentence, we give a weight of -1.5.

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

The example base was created by Yu Shiwen of Beijing University and more Cantonese sentence pairs have been has added. Now, there are about 9000 Cantonese and English sentence pairs, and all the sentences have been annotated with parts-of-speech. The average sentence length for Cantonese is 11 characters and for English is 14 words. Moreover, many

sub-dictionaries of nouns, verbs, adjectives, pronouns, classifiers, and prepositions, etc. are employed. There are many specific features that are helpful for sentence comparison in each of these dictionaries.

For the parsed Cantonese sentence "*S=[他/r]NP[是/v[一個/q 學生/n]NP]VP(He is a student)*", the example sentence could be "*S=[她/r]NP[是/v[一個/q 工人/n]NP]VP (She is a worker)*".

3.5 Target Construction

This stage involves using the Cantonese and English phrase structure relations of the example translation as a template to build the target English sentence. The SST of the source Cantonese sentence contains the following types of nodes:

1) Bilingual corresponding Node (BN): it provides a correspondence between the example English sentence tree and translation template tree (see Figure 4).

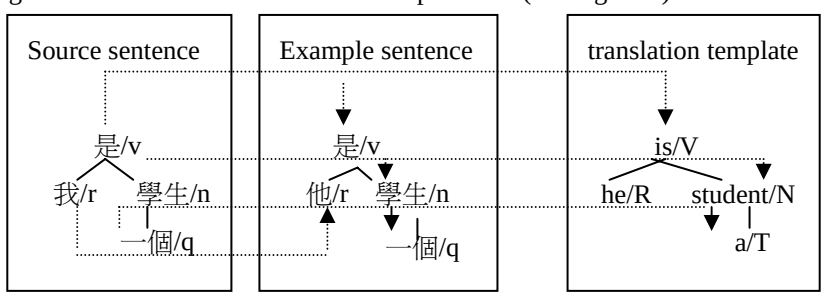

Figure 4. An example of a BN in the SST.

The nodes "*是(be)*" , "*學生(student)*" , and "*一(a)個*" belong to BN .

2) Single corresponding Node (SN): this type of node only has a corresponding node in the example English sentence tree and has no corresponding node in the translation template tree. An example is the node "*我(I)*" in the above source sentence.

3) Non-corresponding Node (NN): this type of node provides no correspondence between the example English sentence tree and translation template tree (see Figure 5). There are two types of NNSs:

   a) NN_c: the word depicted by this node is a content word. See the node "*女兒(daughter)*" in the following example.

   b) NN_f: the word depicted by this node is a function word. See the node "*和(and)*" in the following example.

4) Tense Node (TN): this type of node can determine the tense of a target English sentence. Table 2 shows Cantonese words that can represent the tense of the corresponding English sentence.

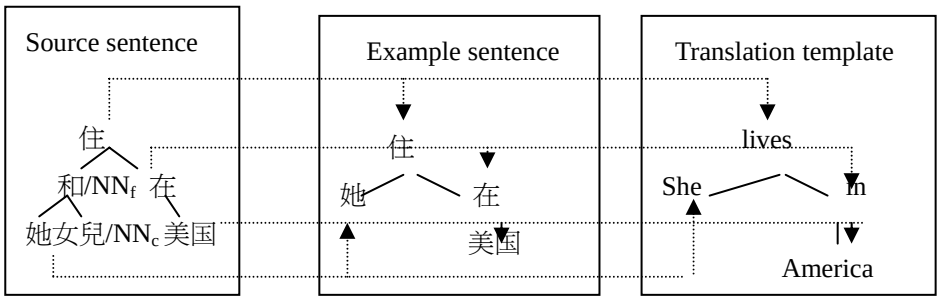

Figure 5. An example of an NN in the SST.

Table 2. The correspondence between English sentence tense and Cantonese words.

English sentence tense	Corresponding Cantonese words
The present continuous	正(just), 正在(in progress of), 即時(at present), 即刻(immediately), 在進行(in progress)...
The present perfect	已(already), 已經(already), 經已(already), 曾經(ever) ...
The past indefinite	過(over), 了(end), 過去(past), 以往(previously), 以前(ago), 从前(aforetime), 上次(last time), 昨日(yesterday) ...
The future indefinite	會(be able to), 將(shall), 就要(going to), 終将(eventually), 將會(will be able to), 即將(be about to), 就會(will be able to), 就快(soon), 就來(come soon), 快要(soon), 明日(tomorrow), 明年(next year)...

5) Type, Voice, and Mood Node (TVMN): this type of node can determine the voice and mood of a target English sentence. Table 3 shows Cantonese words that can represent the tense of the corresponding English sentence.

For the above different types of nodes in the SST, the system applies different replacement rules to translate the phrases stored in these nodes.

Table 3. The correspondence between English sentence types and Cantonese words.

The type of English sentence	Corresponding Cantonese words
The interrogative sentence	嗎?, 什麼? (what), 呢?, 哪(which), 哪些(which kind of), 哪樣(which kind of), 哪裡(where), 是否(whether), 怎麼(how), 怎樣(what about), 怎可(why)
The imperative sentence	v+...+呵!, v+...+吧!, v+...+罷!, 禁止(forbid), 不要(don't), 不准(disapprove), 别(do not), 不許(disallow)
The exclamatory sentence	啊 !(oh), 吧 !, 唉 !(alas), 呀 !(oh!), 哇 , 呵 , 多麼 +...+!(how+...+!), 啦!,...
The negative sentence	不(not), 沒(no), 不許(disallow), 不要(not), 不准(not), 别(not), 不可(cannot), 不能(cannot), 不得(need not), 不顧(in spite of), 别要(must not), ...
The passive voice sentence	被(be), 遭(by), 遭人(by someone), 遭到(be), 遭受(be), 受到(by) .....

The replacement rules in RB_2 are formulated as follows:

Rule ::= fore-condition | replacement-action;
fore-condition ::= condition$_1$|*condition*$_2$|...|*condition*$_n$;
replacement-action ::= action$_1$,*action*$_2$,...,*action*$_m$.

For the node BN, m=0; i.e., the system does not need any replacement action because the source word has the corresponding target word in the translation template.

For the node SN,

replacement-action ::= look(ew), look(sw), repl(E-ew, E-sw) .

Here, *look* is the action of looking up the bilingual dictionary; *repl* is the action of replacing the translation template; *ew* and *sw* are the Cantonese words in the example sentence and source sentence, respectively; *E-ew* and *E-sw* are the English words corresponding to *ew* and *sw*, respectively.

For the node NN,

replacement-action ::= look(sw),loca(sw), inst(E-sw).

Here, *loca* is the action of determining where to insert *E-sw* in the translation template*; inst* is the action if inserting *E-sw* in the translation template.

For the node TN,

replacement-action ::= look(sw$_v$*), chan(E-sw*$_v$*).*

Here, *sw*$_v$ is the current verb in the source sentence, and *chan* is the action of changing *E-sw*$_v$, for example, *E-sw+*... "*ing*" for the present continuous tense, *E-sw+* "*ed* " for the past tense, *E-sw* + "*will*" + *sw for* the future tense, and so on.

For the node TVMN,

replacement-action ::= recv(E-sw$_v$*), chan(tran-tmplate).*

Here, *recv* is the action of recovering the verb of the template, *chan(tran-tmplate)* is the action of changing the voice of the translation template, such as "do" + *subj+verb* , "*will"+ subj+verb* , " *have*" + *subj+verb* for query sentence, or " *do not" +verb*, "*did not"+ verb* for a negative sentence.

The process of target construction can be described as follows (see Figure 6 for an example):

1) Recovering the words in the translation template: Because the criterion of similarity matching is based on content words, and because in a Cantonese sentence, the function

words determine the word form change of its corresponding English sentence, when the system gets an example sentence from the example base, the chance of having an example sentence with a different tense and voice from that of the source sentence is quite high. So the system first deletes the tense and voice of the translation template, and then adds the tense and voice corresponding to the source sentence.

For example,

Translation template: he worked in the factory. ⟶ *he work in the factory.*

2) The replacement rules are applied to change the translation template and generate the target sentence.

3) Experimental results

The LangCompMT05 system was realized using MS Visual C++ for Windows. Users can easily interact with the system to perform translation. Table 4 lists some experiential results. They indicate that the accuracy of the system is 80.6% (see Table 5). The test sentences were created by the authors. Four translation experts manually scored the system's translation results. The score range was from 0 to 100, and we got the accuracy of the system by averaging the scores. The average translation time per sentence was 36 seconds.

Most of the translation errors are due to the following cases:

1) The preposition and noun in the sentences are replaced with error words. The corrected translation for "*在桌上*" is "on the desk", not "in the desk".
2) Some Cantonese phrasal words has no corresponding English words. "*急急腳*", for example, is a special Cantonese phrasal word. An insufficient knowledge base is the cause of most of the problems in natural language processing.
3) Segmentation errors also cause the translation errors. For example, "*是/非常/常/混淆(Is extremely confused)*", "*她/是/非常/漂亮/的 (She is very pretty)*".
4) POS errors also cause the translation errors. POS tagging is mainly statistic-based, and it selects categories that often occur in the corpus. For example, "*書/n 在/p 桌/n 上/u (The book is in the desk)*", "*他/r 上/u 山/n (He is climbing up the mountain)*". This type of error can be solved by means of syntactic analysis.

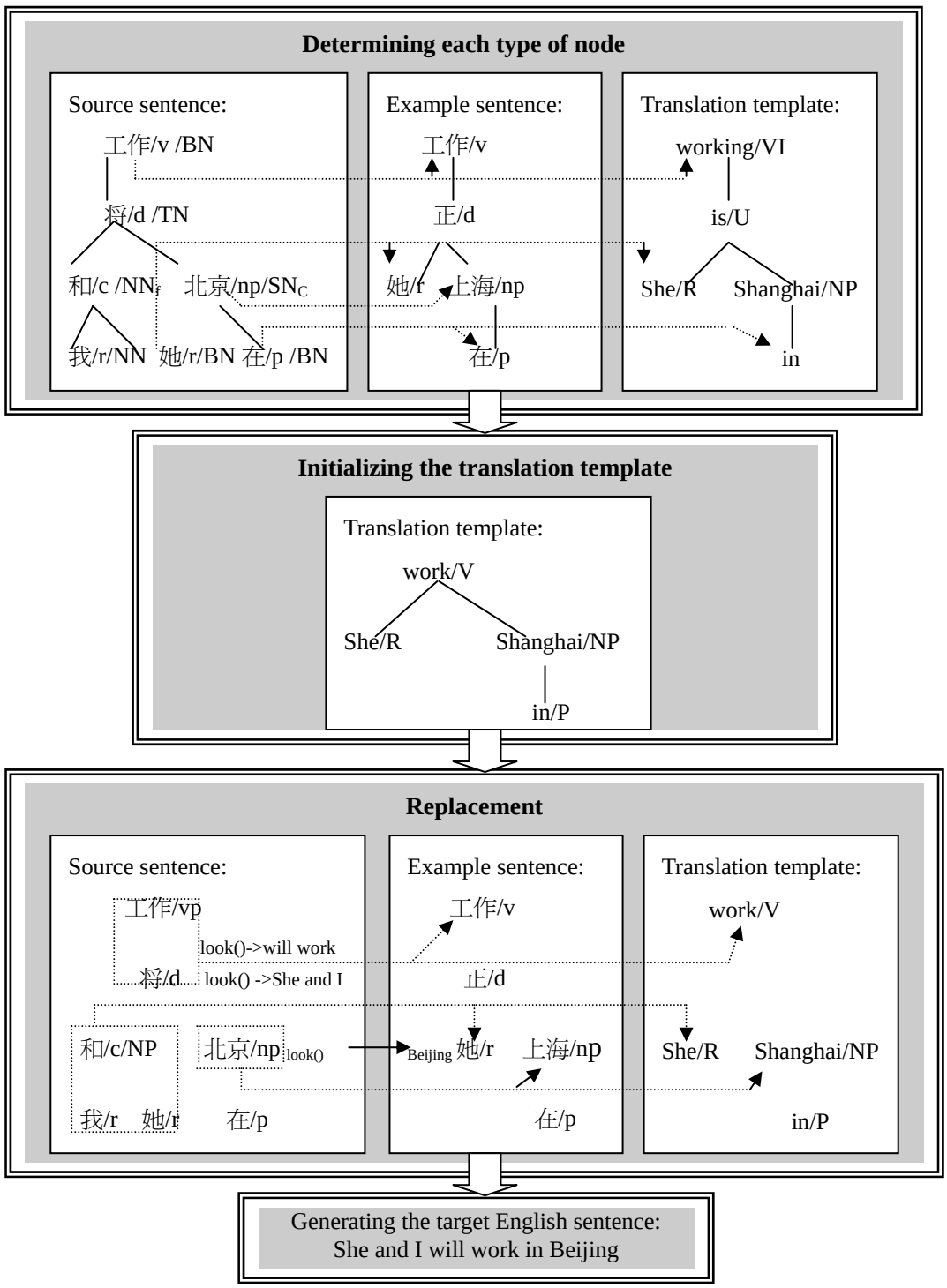

Figure 6. An example of target construction

Table 4. The experimental results

Test target	Input sentence	Selected example sentence and template	Target sentence
Testing 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 is playing football.)	The boy with his hands in his pockets is 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 pockets is playing football.)	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 hands in his pockets is playing football.)	The boy with his feet in the desk is reading a book.
	4. 他騎單車返工.	她乘巴士返工. (She goes to work by bus.)	He goes to work by bike.
Testing sentence tense change	1. 她明天將離開這裡.	我昨天離開這裡的. (I left here yesterday.)	She will leave here tomorrow.
	2. 她已讀書了.	她正在讀書. (She is reading the book.)	She has read the book.
	3. 他讀書了.	她正在讀書. (She is reading the book.)	He reads the book.
Testing plural nouns	1. 她有兩把刀.	她有一把刀. (She has a knife.)	She has two knives.
	2. 我有三個孩子.	她有一個孩子. (She has a child.)	We have three children.
Testing the irregular verbs for past tense	1. 更多業內人士讀了这个規定.	學生們借了你的茶壺. (Students borrowed your teapot.)	More professional people have 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.

Testing the coherence between the subject and verb	1. 香港公證会正式獨立.	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. 物价因應市場反應而增減	人們因應季節變化而換裝. (People change their clothes according to the season.)	The price changes according to the market reaction.

Table 5. The experimental results.

Source sentence type		Number of Test sentences	Translation accuracy (%)
Descriptive sentence	Positive	100	81.0%
	Negative	80	82.2%
	Passive	50	81.6%
	Present tense	50	84.0%
	Present continuous tense	35	83.6%
	Present perfect tense	90	79.9%
	Future indefinite tense	40	82.9%
Interrogative sentence	Present tense	65	78.9%
	Present continuous tense	70	80.6%
	Present perfect tense	60	80.8%
	Future indefinite tense	50	75.5%
Imperative sentence	Positive	80	79.7%
	Negative	45	76.8%
Exclamatory sentence		50	81.9%
Total		865	80.6%

4. Conclusion and Future Work

We have proposed an integrated method for Cantonese-English machine translation that makes use of morphological knowledge, syntax analysis, translation examples, and target-generation-based rules. The principles and algorithms used in this MT system have been

well tested. The source sentence is segmented first, then it is tagged and parsed it, and the SST of the source sentence formed for its structural representation. Finally, using the computational linguistic method, an example sentence is selected from the EB; its corresponding English translation sentence is used as the translation template, and the target sentence (English) is generated based on rules.

Machine translation especially in the Cantonese-English domain is quite a difficulty task. Based on our research on the LangCompMT05 system, we have proposed an integrated MT method that is mainly based on an example-based machine translation method, and we believe that this integrated method is feasible for solving many translation problems. With the computational method, we find that it is possible to acquire bilingual knowledge from a small-scale, representable EB. We have proposed a number of algorithms, such as a Cantonese segmentation algorithm, similarity calculation algorithm, and a target sentence construction algorithm. We have created databases, which contain many Cantonese words and related information. For example, our Cantonese dictionary contains part-of-speech and word frequency information. The EB stores many Cantonese-English sentence pairs that have been segmented and tagged with POSs. The bilingual dictionary stores the Cantonese words and corresponding English words. This information source will be valuable for future development of other NLP systems.

References

Brown, R. D., "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, 1997, pp. 23-25.

Carl, M., "Inducing Translation Templates for Example-based Machine Translation," In *proceedings of Machine Translation Summit VII99*, 1999, pp. 250-258.

Carpuat, M., and D. Wu, "Word Sense Disambiguation vs. Statistical Machine Translation," *43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*. Ann Arbor, MI: Jun 2005, pp. 58-75.

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

Chen, K., and J. You, "A Study on Word Similarity using Context Vector Models," *International Journal of Computational Linguistics and Chinese Language Processing*, 7(2), 2002, pp. 37-58.

Church, K., "Aligning Parallel Texts: Do methods Developed for English-French generalization Asia Language?" Technical Reported from Tsinghua University, 1994.

Fung, P., and K. W. Chen, "K-vec: A New Approach for Aligning Parallel Texts," *COLING-94*, pp.1096-1104.

FuRusE, O., and H. Iida, "An Example-Based Method for Transfer-Driven MT," *TMI-92*, pp. 139-148

Hou, M. , J. J. Sun, and Z. X. Chen, "Ambiguities in Automatic Chinese Word-Segmentation," In *Proceedings of 3rd national conference on computing linguistics*, 2001, pp. 81-87.

Kit, C., K. Pan, and H. Chen, "Learning case-based knowledge for disambiguating Chinese word segmentation: A preliminary study," In *COLING2002 workshop:SIGHAN-1*, 2002, pp. 33-39.

Kit, C., H. Pan, and J. J. Webster, "Example-based machine translation: A new paradigm," *Translation and Information Technology*, ed. By S.W. Chan, translation department, Chinese University of HK Press, Hong Kong, 2000, pp. 57-78.

Knight, K., and D. Marcu, "Machine Translation in the Year 2004," In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 18-23,2005, pp. 45-50.

Li, W., Q. Lu, and R. Xu, "Similarity Based Chinese Synonym Collocation Extraction," *International Journal of Computational Linguistics and Chinese Language Processing*, 10(1), March 2005, pp. 123-144.

Liu, Y., Q. K. Tan, and X. Shen, *Contemporary Chinese Language Word Segmentation Specification for Information Processing and Automatic Word Segmentation Methods*, Tsinghua University Press, Beijing, 1994.

Marcu, D., "Towards a Unified Approach to Memory- and Statistical-Based Machine ranslation," In *Proceedings of ACL-2001*, Toulouse, France, July 2001, pp.59-70.

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

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

Mosleh, H. A. A., and E. K. Tang, "Example-based Machine Translation Based on the Synchronous SSTC Annotation Schema," In *Proceedings of Machine Translation Summit VII'99*, 1999, pp. 244-249.

Nie, J. Y., M.-L. Hannan, and W. Y. Jin, "Unknown Word Detection and Segmentation of Chinese Using Statistical and Heuristic Knowledge," *Communications of COLIPS*, 5(1&2), 1995, pp. 47-57.

Sergei, N., "Two Approaches to Matching in EBMT," *TMI-93*, 1993, pp. 47-57.

Somers, H. L., "Example-based machine translation," Eds. by R. Dale, H. Moisl and H. Somers, New York:, pp. 611-627.

Tsujii, J., "Future Directions of Machine Translation," In *Proceedings of 11th International Conference on Computational Linguistics*, Bonn,pp. 80-86.

Wu, Y. and J. Liu, "A Cantonese-English Machine Translation System PolyU-MT-99," In *Proceedings of Machine Translation Summit VII 99*, Singapore, 1999, pp. 481-486.

Zhou, L.N., J. Liu, and S. W. Yu, "Similarity Comparison between Chinese Sentences," In *Proceedings of ROLING'97*, Taiwan, 1997, pp. 277-281.

Zhou, L.N., J. Liu, and S. W. Yu, "Study and implementation of combined techniques for automatic extraction of word translation pairs: An analysis of the contributions of word heuristics to a statistical method," *International Journal on Computer Processing of Oriental Languages*, 11(4), 1998, pp. 339-351.

Zhang, M., S. Li, T. J. Zhao, and M. Zhou, "A Word-Based Approach for Measuring the Similarity between two Chinese Sentence," In *Proceedings of national conference of 3rd Computational Linguistics*, Beijing, 1995, pp. 152-158.

Zens, R., H. Ney, T. Watanabe, and T. Sumita, "Reordering constraints for phrase-based statistical machine translation," In *Proceedings of COLING-2004*, Geneva,Switzerland4, pp. 23-29.