自然言語処理 120－4
(1997. 7. 24)

スーパー関数に基づく多言語機械翻訳

任 福 継

広島市立大学情報科学部

〒731-31 広島市安佐南区大塚東 3-4-1

TEL/FAX (082) 830-1584
Email: ren@its.hiroshima-cu.ac.jp

あらまし　高精度の機械翻訳システムを実現するために、ここ数十年にわたって、構文解析、意味解析、さらには文脈情報を利用するといった方法による数多くの機械翻訳システムが実用化されてきた。このような従来の研究では、流暢な訳文の作成並びに翻訳作業の自動化もしくは半自動化を目的としていたが、現段階では多くの問題が未解決のままで残されている。

本論文では、インターネットの急速な普及という情報社会の現状を鑑み、流暢な訳文の作成より、外国語情報をすばやく母国語で拾い読みする（Browse）ことを機械翻訳の目的とする考えに基づき、構文解析や意味解析などは行わず、単に言語と言語との間の構造関係（本文ではスーパー関数と呼ぶ）を用いる多国語機械翻訳方法を提案するものである。

キーワード　スーパー関数，機械翻訳，翻訳エンジン，日本語，中国語，英語

Super-Function Based Machine Translation

Fuji Ren

Faculty of Information Sciences
Hiroshima City University

TEL/FAX (082) 830-1584
Email: ren@its.hiroshima-cu.ac.jp

Abstract　In this paper, we present a new method called Super-Function Based MT(SFBMT) to improve the GBMT. We use a Super-Function(SF) in the translation engine to enhance the translation quality and to reduce the glossary quantity. The Super-Function is a function that shows the correspondence between original language sentence patterns and target language sentence patterns. We used TTB to store SF and to match SF with the input sentence. An experiment on the Japanese-English-Chinese textbook has been simulated. The 61 SFs are acquired and the result shows that this method is efficient.

key words　Super-Function, Machine Translation, Translation Engine, Japanese-Chinese-English

1 Introduction

Over the past decade, the number and diversity of experimrnts in Rule-Based Machine Translation (RBMT), Knowledge-Based Machine Translation(KBMT), Example-Based Machine Translation (EBMT) have grown significantly[1-9]. Most of these motheds are aimed at building Automatic High-Quality Translation System (AHQTS). For example, nobody would think that the following A-groups are better than B-groups.

Original Sentence (Chinese): (C-1) Ta CHI jiaozi
(C-2) Ta CHI yuo

Target Sentence (English):
A-group: (E-1a) He eats Chinese dumpings
(E-2a) He eats medicine.
B-group: (E-1b) He eats Chinese dumpings
(E-2b) He takes medicine.

Target Sentence (Japanese):
A-group: (J-1a) 彼はチャオズを食べる.
(J-2a) 彼は薬を食べる.
B-group: (J-1b) 彼はチャオズを食べる.
(J-2a) 彼は薬を飲む.

However, if youe goal is to just browse sets of documents written in foreign languages using your mother language, say in a Web, do you care that the "take medicine" is changed to "eat medicine"? We think that most users could rather that the system is fast, inexpensive, easy to control and easy to update. This means that we can't and don't neeed highly fluent translations. Some researches have indicated that evaluate Machine Translation effectiveness make the point that theoretically and practically productive systems will reside in applications that exploit the complementary strengths ofthe machine and the human[5,13].

Recently, Zajac and Michelle described a Glossary-Based MT Engine in a Multilingual Analyst's Workstation Architecture[11,12]. The CRL Temple project has developed an open multilingual architecture and software support for rapid development of extensible Machine Translation functionalities. Currently, the Temple prototype provides automatic raw English translations from documents in several languages (Spanish, Arabic, Japanese and Russian). Translations are produced using a Glossary-Based Machine Translation engine. Analysts and translators can edit the raw translation using a multilingual editor. Source documents and their translations are managed using the Tipster Document Manager developed at CRL which is also used as the architectural basis for integrating the system's components. One important outcome of the Temple project is the development of an architecture to support reuse of NLP tools and resources: (a) Tools that are acquired from an external source, such as morphological analyzers, generators, or taggers, can be integrated in the system with a minimum of programming effort. (b) Heterogeneous linguistic resources are parsed and mapped to a common multilingual representation. A Glossary-Based Machine Translation (GBMT) engine which provides an automatic translation for each language pair is one of the major components of the Temple prototype.

GBMT is used to provide an English gloss of a foreign document. A GBMT system uses a bilingual phrasal dictionary (glossary) to produce a phrase-by-phrase translation. Translation (based on phrase pattern matching) is fast and accurate regarding the content of the document and browsed documents can be translated almost in real-time[11,12]. A GBMT system for a language pair is also extremely simple, cheap and fast to be developed. Moreover, all language resources used by the system are entirely under the control of the user.

However, current implementations of the GBMT system lack good translation accuracy and readability. For example, because the order of words is not considered, X1 of X2 should be translated X2 of X1 for a Japanese document. Moreover, although X1 について (ABOUT X1) has been taken in the bilingual glossary, if we need to translation X2 について, the X2 について (ABOUT X2) must be taken in the bilingual glossary.

In this paper, we present a new method called Super-Function Based MT(SFBMT) to improve the GBMT. We use a Super-Function in the translation engine to enhance the translation quality and to reduce the size of the required glossary. The Super-Function(SF) is a function that shows the correspondence between original language sentence (word, phrase, sentence, paragraph, text) patterns and target language sentence (word, phrase, sentence, paragraph, text) patterns. At the currently, we consider the SF only for phrase and sentence.

2 Overview of GBMT

GBMT systes were first developed at Carnegie Mellon University as a part of the Pangloss MT project[9,14-16]. In that effort, a sizeable Spanish-English glossary-based MT system was implemented. The Temple project has bulit upon this experience and extended the GBMT approach to other languages: Japanese, Arabic, and Russian.

The GBMT engine uses a bilingual glossary and a bilingual dictionary to produce a translation of phrases in a source text. The input to the engine is a flat tree structure where the root represents the entire text, the intermediate nodes are sentence nodes, and the leaves of the tree are analyzed lexical tokens that also contain the translation of each lexical token. The GBMT engine is parametrized by a bilingual glossary. The bilingual glossary is essentially a phrasal dictionary: a glossary entry contains a source phrase pattern, a set of corresponding target phrase patterns, and correspondences between variables in the source and in the target patterns. A GBMT system produces phrase-by-phrase translation of the source text, falling back on a word-by-word translation when no phrase from the glossary matches the input. Thus, the size of the glossary and the flexibility of the pattern language are crucial for the production of good translations.

The GBMT engine processes source tree structures in four steps[11]:

1. Glossary phrases are matched within sentence sub-trees,

2. Target phrases pattern are added in the tree for each source phrase match,

3. Morphological information is transferred from source tokens to target tokens, and

4. Agreement binding information is generated for each source phrase.

3 SF Definition

3.1 SF Definition

[Definition 1]

A Super-Function(SF) is a function that shows the correspondence between original language sentence (word, phrase, sentence, paragraph, text) patterns and target language sentence (word, phrase, sentence, paragraph, text) patterns.

[FORMAL DESCRIPTION]

[O_STRing] < <VARiable>+ <O_STRing>* >+
[O_STRing] :=> SF(T_STRing,VARiabl) (1)

Notation: [ ] means optional(i.e.,0 or 1); + means 1 or more; * means 0 or more; O means original language; T means target language.
Here, STRing means a natural language (original language or target language) character string; VAEiable could be a word or, a phrase or, a sentence or, a paragragh. VAEiable could be also a SF.

For discussion, we sometimes note SF (1) into (2) or (3).

SF_O(O_STRing,O_VARiable) :=>
SF_T(T_STRing,T_VARiabl) (2)

f(X1,X2,......,Xn) (3)

Here, Xi(i=1,..n) is a VARiable, n means the number of variable of SF f.
Examples: (notation: C means Chinese, J means Japanese, E means English.)

f1:<C_VARiable>:=><E_VARiable> (4)

ex. Lianheguo :=> the United Nations (4')

f2:<C_VARiable> ming bu xuchuan :=>

<E_VARiable> have a well-deserved reputation (5)

ex. ta ming bu xuchuan :=>
he has a well-deserved reputation (5')

f3: <J_VARiable>のみならず <J_VARiable>2
:=> not only <E_VARiable>1 but <E_VARiable>2
(6)

f4: <J_VARiable>は重要である :=>
<E_VARiable> is important. (7)

ex. 富のみならず健康は重要である:=>
Not noly wealth but health is important. (7')

NOTE: We can see that SF can be a complex function. In other words, a variable in SF could be a SF. For example, the variable of SF f4 could be a SF f3.

f5: WoZhengzaimouqiunimen <C_VARiable>1 yue <C_VARiable>2 ri zai <C_VARiable>3 shangkandengzhaopin <C_VARiable>4 degongzuo :=> I am looking for a position as <E_VARiable>4 which you describe in your advertisement in <E_VARiable>3 of <E_VARiable>1 <E_VARiable>2
(8)

ex. Wo Zhengzai mouqiu nimen 5yue 18ri zai Guangming Daily shang kandengzhaopin zhulimishudegongzuo :=> I am looking for a position as an assistant secretary which you describe in your advertisement in Guangming Daily of May 18.
(8')

f6: <J_VARiable>1 要約 <J_VARiable>2 まえがき <J_VARiable>3 ...... むすび <J_VARiable>4 謝辞 <J_VARiable>5 参考文献 <J_VARiable>6 :=>
<E_VARiable>1 Abstract <E_VARiable>2 Introduction <E_VARiable>3 ...... Conclusion <E_VARi able>4 Acknowledgments <E_VARi able>5 Refrernces <E_VARiable>6 (9)

f1 is the same as a bilingual dictionary, and f6 is about whole text, so the SF, such as f1 and f6, will not be discussed in this paper.

4 SFBMT Engines

This section describes the structure of a SF, how it is mapped to the text and how the translation is produced.

4.1 Example of SF

As a preliminary study, we examined a group of scientific papers and abstracted the following SFs which consist of combinations of the pattern sets A, B, C and D (below). By using the SF (10) below we can translate such sentences which are constructed of combinations of phrases from these four groups. This also suggests how to find SF from corpus.

SF_E(X1,X2,X3,X4) = There is X1 information [X2] X3 of X4
:=> SF_C(X1,X2,X3,X4) = [X2] you guanyu X4 de X3 de X1 baodao
:=> SF_J(X1,X2,X3,X4) = [X2] Y4 の Y3 に関して X1 情報がある
(10)

A: There is some information
There is enough information
There is sufficient information
There is a great deal of information
There is a lot of information
There is ample information
There is precise information
There is detailed information
There is reliable information
There is valuable information

B: (in literature)
(at present)
(nowadays)

C: about the use of

on the application of
concerning the observation of
regarding the fine structure of
bearing on the action of

D: atomic energy
mitochondria
new technology to printing and dying
......

4.2 The process of SFBMT

The SFBMT uses a bilingual dictionary and Super-Functions to produce a translation of a source text. The input sentence is frist subjected to morphological analysis, then matched with the source sentence and a SF. SFBMT produces sentence-by-sentence translation of the source text, falling back on phrase-by-phrase and word-by-word translation when no SF matches.

The process of SFBMT consists of three major parts.
(1) Morphological analysis
(2) SF matching
(3) Morphological tanslation

4.2.1 SF Matching

A SF is represented by a Node Table and an Edge Table. Matching a SF is simply matching each node of NTB and confirming each edge's kind. The node sometimes has more than one value and sometimes can be specified. The construction of NTB and ETB are described in Table 1 and Table 2. NTB is also called String TaBle(STB). ETB is called Variable TaBle (VTB) too.

Table 1: Node Table

No.	Language 1	......	Language m	
1				
2				
......				
n				

Table 2: Edge Table

No.	Kind	Location 1	......	Location m
1				
2				
......				
n				

Language i in NTB is a string of language i; Kind in ETB indicates the kind of variable. The location i in ETB indicates the location relationship of language i.

We use the following SF(11) and SF(12) to explain such node.

X1 HA X2 WO X3(V+TEIRU) :=>
X2 BE^ X3(V+ing) X2 (11)

The node BE^ means that it can be matched with all inflected form of verb be such as is, are.

X1 NI X2 GA IRU I ARU^ :=>
There BE^ X2 ON* X1 (12)

Here, the symbol I represents enumeration and ^ represents all inflected form. It means that the node "GA IRU I ARU^" can be matched by (1) GA IRU, (2) GA IMAU, (3) GA ARU and (4) GA ARIMASU.

The * indecates that the string before * may not be right according context. For example, translation (13) is right but translation (14) is wrong, and the right word is under.

There is a book ON the shelf. (13)

There is a dog ON (under) the tree. (14)

The TTB are sorted according as SF string, so the SF matching is very fast.

4.2.2 Morphological Agreement

Morphological translation is done by tables describing correspondence between categories and sets of features and values in the source language and in the target language. Each category and each morphological feature of a word in the source language is mapped to a catagory and a set of features in the target language[12].

In Chinese there is no change of verb for person and number. For example, for verb BE^ of a SF, is's agree with the subject to generate IS or ARE.

4.2.3 Dispose of Unknown Words

Usually, we suggest an unknow word as noun to match a SF, and generate a target language translation using the string of nuknow source language word. However, for a specific language, we can make some rules to decide if an unknow word is a verb or a noun. For example, in Japanese, if an unknow word locates before a case particle, then the unknow word is considered a noun; if a unknow word locates after the case particle "WO", then the unknow word is considered a verb.

5 Experiment

The SFBMT system has not been completed so far. However, we have done the simulational experiment on a Japanese-English-Chinese text whose title is in "Japanese in thirty hours".The SFs are semi-automatically acquired. We first got 89 SFs from the 50 chapters of this book, then we compiled the 89 SFs into 55 SFs. In additional, we added 6 RSFs. So finally, the SFs used to the experiment was 61. The following show some SFs we got from the textbook.

<26-J> X1 (P1) GA X2 WO X3(V+MASYOU).
<26-E> Let X1(Obj) X3 X2.
Ex.
Japanese: WATASITACHI GA HAKO WO AKEMASYOU.
English: Let us open the box.

The following is the dictionary information for the sentence.

WTASITACHI{}:=>we {pronoun, frist-person; object_form->us;possessive_form->our;...}
HAKO{}:=>box{common_noun, container; +s;......}
AKERU{}:=>1/2:open{verb-i; +s,+ed,+ed,+ing; .....}

Here,P1 indicstes the variable X1 ought be a frist person pronoun, and Obj means object form of a pronoun.

<28-J> X1 (-P1) GA X2 WO X3(V+MASYOU).
<28-E> X1 will X3 X2.
Ex.
Japanese: KARE GA HAKO WO AKEMASYOU.
English: He will open the box.

Here, -P indicates the variable X1 is not a frist person pronoun.

We can see form the SF<26> and SF<28> that the sentences "Let" and "Will" can be well translated by using the variable's features.

<39-J> X1 [NO NAKA]DE X2 GA ICHIBAN X3(a).
<39-E> IN X1 X2 BE^ X3(a+est).
Ex.
Japanese: KONOHEYA NO NAKADE WATASI GA ICHIBAN OOKIIDESU.
English: In this room I am the biggest.

<40-J> X1 HA X2 YORI X3(a).
<40-E> X1 BE^ X3(a+er) than X2.
Ex.
Japanese: KARE HA WATASI YORI OOKIIDESU.
English: He is bigger than I.

Here, the "a+est" and "a+er" represent the superlative degree of a adjective and degrees of comparison of a adjective, respectively.

<41-J> X1 GA X2 (V|V+TA) X3 HA ......
<41-E> (the) X3 which X1 X2 BE^.......

The apostrophe"......." represents any pattern, and that before the apostrophe just is considered as one element when matching the remained parts of the sentence.

Ex.
Japanese: WATASI GA MIMASITA UCHI HA TAIHEN YOI UVHIDESU.
English: The house which I looked at was a very good house.

In the example, overline represents the apostrophe parts.

<42-J> X1 HA X2 GA X3(V|V+TA) X4 DESU^.
<42-E> X1 BE^ X4 which X2 X3(V|V+P).
Ex.
Japanese: KORE HA WATASI GA KAKIMASITA TEGAMI DESU.
English: This is a letter which I wrote.

Using such SFs, we can got not very fluency but usually good translation.

6 Conclusions and Future Works

In this paper, we intrduced the outline of SFBMT project which would used to improve the GBMT. We use a Super-Function in the translation engine to enhance the translation quality and to reduce the glossary quantity. We have definited the SF and dicided the format of SF. We used TTB to store SF and to match SF with the input sentence. A experiment on the Japanese-English-Chinese textbook has been simulated. The 61 SFs are acquired and the result shows that this method is efficient, at least for the text book.

The advantages of SFBMT are clear:

(1) By introduving the SF variables, the glossary quantities will be reduced.

(2) By the SF, the translation quality will be improved, such as word order, conventional expressions

(3) Because the detail syntactic analyzing and semantic analyzing didn't be used, the translation speed is very quick.

It is easy to extend to any specific domain, as add the SF of specific domain.

SFBMT does not provide a function to deal with such as particle "the, a" and "in, on", because we don't think it is a very big problem in browsing documents in mother language.

The future work is:

(1) To automatically acquire SF from corpus.

(2) To use a Finite State Technique in matching between the SF and a sentence.

(3) To apply the SFBMT into Web.

Acknowledgments

The work reported here started when the author worked at CRL as a visiting professor. The author wish to thank Sergei Nirenburg, Jim Cowie, Rema Zajac, Mark Casper and Zhiging Guan for their help and support. We should also thank the Ministry of Education of Japan to fund this project under Joint Research Grant-in-Aid for International Scientific Research JCE-TC(09044179).

References

[1]Kitakami, M., and Matsumoto,Y. 1995, A Machine Translation System based Translation Rules Acquired from Parallel Corpora, Proc. RNpANLP, pp.27-44.
[2]Sata,S.,1995,MBT2: A Method for Combining Fragments of Examples in Example-based Translation, Artificial Intelligence, Vol.75,pp.31-49.
[3]Onyshkevych, Boyan and Nirenburg, Sergei,1996, A Lexicon for Knowledge-Based MT, in Machine Translation vol 10:1-2.
[4]Carbonell,J., T. Mitamura and E. Nyberg, 1992, The KANT Perspective: A Critique of Pure Transfer,Proc of the Fourth International Conference on Methodological Issues in Machine Translation.
[5]Church, K. and Eduard Hovy, 1993, Good Applications for Crummy Machine Translation, in Machine Translation, vol.8, no.4.
[6]Dorr,Bonnie, 1993, Machine Translation: A View from the Lexicon. Cambridge MA: MIT Press.
[7]Farwell,David,Louise Guthrie,and Yorick Wilks,1993, Automatically Creating Lexical Entries for ULTRA, a Multilingual MT System, Machine Translation vol.8:3,pp.127-146.
[8]Nirenburg, Sergei and Kenneth Goodman,1990,

Treatment of Meaning in MT System, Proc. of the Third International Conference on Theoretical and Methodlogical Issues in Machine Translation of Natural Language.

[9]Nirenburg, Sergei, 1995, Bar Hillel and Machine Translation: Then and Now, Proc. of BISFAI-95.

[10]Nirenburg, Sergei,, editor,1996, Recent Papers from the Mikrokosmos and Corelli Projects, New Mexico State University, Computing Research Laboratory.

[11]Remi Zajac, 1996, A Multilingual Translator's Workstation for Information Access, International Conference on Natural Language Processing and Industrial Application, Canada.

[12]Remi Zajac and Michelle Vanni, 1996, Glossary-Based MT Engines in a Multilingual Analysts's Workstation Architecture, to appear in Machinr Translation.

[13]Johnson,R.L. and P. Whitelock,1987,Machine Translation as an Expert Task. in S. Nirenburg(ed.), Theoretical and Methodlogical Issues in Machine Translation, Cambridge.Combridge University Press,pp.136-144.

[14]Nirenburg, Sergei, (editor) 1995, The PANGLOSS Mark III Machin Translation System. CMU-CMT-95-145. A Joint Technical report by NMSU CRL.

[15]Nirenburg, Sergei et al, 1993, Multi-purpose Development and Operations Environments for Natural Language Applications, Proc. of the 3rd Conference on Applied Natural Language Processing,Italy.

[16]Frederking,R., D. Granners, P. Cousseau, and S. Nirenburg, 1993, An MAT Tool and Its Effectiveness, Proc. of the DARPA Human Language Technology Workshop, Princeton, NJ.

Appendix 1: Tool for MaKing SF(MKSF)

we describe a tool for making SF here. The function of MKSF is described as following.

OL:	****	xxx X1	****	xxx X2	...xxx... Xi	****.
TL:	*****	xxx Y1	*****	xxx Y2	...xxx... Yj	******.

***: string of original language or target language.

xxx: a variable

Xi(i=1,2...n): original language variable

Yj(j=1,2...n): target languag variable

When the above sentence(s) is decided to make a SF, and we suppose that the order of the target languag variables which correspond with original language variables X1,X2,...is Yj(->X1),Y1(->X2),Y2(->X2), then:

Step 1: call the sentence(s) (OL and TL)

Step 2: using a mouse to touch X1,X2,Xi
in proper order
RETURN

Step 3: using a mouse to touch Yj,Y1,Y2
successively
RETURN

For example:

English: There is detailed information nowadays
X1 X2
regarding the fine structure of mitochondria.
X3 X4

Japanese: 現在のところ ミトコンドリア の
Y2 Y4
正確な構造に関して 詳しい 情報がある
Y3 Y1

Chinese: 現在 有関於 線粒体 的 精密結構 的
Z2 Z4 Z3
詳細 報道
Z1

We noly need to touch X1,X2,X3,X4,RETURN, Y1,Y2,Y3,Y4,RETURN, [and Z1,Z2,Z3,Z4] successively to make such a SF. Namely, The NT and ET could be auto-made by above operation using the MKSF.