Functional Approach for Patent Translation

Hideto Ikeda
Department of Computer Science
Ritsumeikan University
Japan
hikeda@is.ritsumei.ac.jp

Nguyen Thanh Hung
Department of Computer Science
Ritsumeikan University
Japan
hungnt@is.ritsumei.ac.jp

Ze Zhong Li
Department of Computer Science
Ritsumeikan University
Japan
lizezhonglaile@163.com

Abstract

This paper introduces a new approach of patent translation. The basic technology is functional language model proposed by the same authors. Using a long claim sentence that applied to America and China of Japanese patent originally, the translation process is demonstrated. Through this process, special linguistic issues in each language are clarified in translation of patents.

Keywords: Natural Language Proceeding, Machine translation, Patent translation, Functional Model of sentences

1 .. Introduction*

Although patent translation is hot topic and urgent issue for the recent globalization of the world, almost translations are done by manual work, because the quality of machine translation systems is not enough for using translated sentences directly. It is important to review conventional approaches of machine translation (MT) and establish a new approach to overcome the problems, especially in patent translation.

Conventional approach of MT for patent translation can be classified as follows:

- Rule-based approach
- Statistical Approach
- Example-based approach

Among them, the rule-based approach analyzes sentences grammatically on the base of morpheme dictionary and part-of-speech (POS) based rules. The quality was not enough, because we cannot neglect various exceptions of registered rules. The morpheme-based analysis is also problematic for long sentences.

Statistical Approach takes ways to making language model statistically. One of statistical model is Bayesian network[1] that is based on probability of word appearances after a specified word. Although there have been proposed various statistical models, the altitude of less emphasis in grammatical structure of sentences drives grammatically illness of translated sentences. This approach, however, was useful for gathering many candidates of words, phrases and sentences to develop various corpuses.

Example-based approach has taken very important roles for automation of patent translation. The patent office of each country provides a service to publish patents in English as one of foreign languages by using example-based approach. This approach encourages alignment technologies to find a pair of word-, phrase- and clause-patterns from a pair of sentences in a parallel corpus. The result of alignment is useful for patent translation. Some

[1] http://en.wikipedia.org/wiki/Bayesian_network

Proceedings of the 4th Workshop on Patent Translation, pages 64-69,
MT Summit XIII, Xiamen, China, September 2011

patent offices, [Japio 2010, kipo 2011] have collected several hundred patterns from the patents and improved the quality of translated patents, but still not enough.

Almost real products of patent translation adopt a combination of these approaches to improve the quality of translation.

This paper introduces a new approach of MT and application to patent sentences to show how to translate patent sentences and some linguistic issues specially appeared in patent sentences.

2 Example of a Patent Translation

2.1 Sample sentence of a patent

One of special characteristics of patent sentences is the length of sentence, especially sentence for claim because each claim has to be described one sentence. We shall show one example of claim sentence.

(Original sentence into English)

"A first aspect of the present invention is directed to an information processing system, including a first information processing apparatus and a second information processing apparatus, wherein the first information processing apparatus includes: a first object storage portion in which at least one object having at least one data node and at least one application node that are hierarchical can be stored; a first instruction receiving portion that receives an object instruction, which is an instruction having an object identifier for identifying an object and is an instruction for the object, from the second information processing apparatus; a first object non-existence information acquiring portion that acquires information to an effect that the object identified with the object identifier contained in the object instruction does not exist, in a case where the object does not exist; a first processing portion that processes the object identified with the object identifier contained in the object instruction according to the object instruction, in a case where the object exists; and a first transmitting portion that transmits result information, which is information acquired by the first object non-existence information acquiring portion or information on a result of the process executed by the first processing portion, to the second information processing apparatus; and the second information processing apparatus includes: a second non-existence case process information storage portion in which a pair of an object identifier and a non-existence case process identifier for identifying a process that is executed in a case where an object identified with the object identifier does not exist can be stored; a second receiving portion that receives the result information transmitted by the first information processing apparatus; a second processing portion that executes the process identified with the non-existence case process identifier corresponding to the object identifier of the object, in a case where the result information received by the second receiving portion is information to an effect that the object does not exist; and a second output portion that outputs the result information received by the second receiving portion, in a case where the result information is not information to an effect that the object does not exist."

2.2 Functional decomposition of the English sentence

The key flame of the English sentence is as follows;

A first aspect of the present invention is directed to an information processing system, including a first information processing apparatus and a second information processing apparatus, wherein …

By using our functional approach [Ikeda, et al 2011], the sentence can be expressed as the following sequence of functions:

S0=A first-aspect-of-the-present-invention-is-directed-to-_-, including-_- and_,-wherein-_([N1],[N2],[N3],[S4]);
N1= an information processing system;
N2= a first information processing apparatus;
N3= a second information processing apparatus;

Although *S4* is a sentence followed word "wherein", it can also be expressed by the following sequence of functions:

S4= _-includes-_- and-_-includes:-_ ([N5],[N6],[N7],[N8]);
N5= the- first-information-processing apparatus;
N6=_;-_;-_;-_;-and-_([N9],[N10],[N11],[N12],[N13]);
N7= the second information processing apparatus;
N8=_;-_;-_;-and-_([N14],[N15],[N16],[N17]);

N6 and N8 are lists of noun phrases. Since each element of these lists has similar structure, we shall show just one example of N9:

N9=_- in- which- ([N18],[S19]);
N18= a first object storage portion;
S19= _-can-be-stored([N20]);
N20= _-having-_ ([N21],[N22]);
N21=at- least-one-_([N23]);
N22=_-that- are hierarchical([N24]);
N23=object;
N24=_- and-_([N25],[N26]);
N25= at- least-one-_([N27]);
N26= at- least-one-_([N28]);
N27= data node;
N28= application node

2.3 Problems of Phrase Alignment in English and Chinese

Translation of "relative pronoun" of English into Chinese has many mistakes. Although a Chinese word with the same role of English "relative pronoun" is particle "的". For example, a translation of English sentence "The book that you bought is really interesting." could be "你买的书真有意思". The previous phrase of "的" in Chinese is a sentence "你买 (I bought)". It is possible to locate a sentence in front of "的", but it is not adequate to locate long sentence, especially if it includes another "的". For example in the patent sentence in Chinese, it is difficult to translate English sentences including a relative pronoun into Chinese

In such a case, we have to use 2 or more sentences to express one English sentence. For example, the following English noun phrase:

> "*a first object storage portion in which at least one object having at least one data node and at least one application node that are hierarchical can be stored;*

is translated in to Chinese as follows:

> "*第一对象存储部，可存储一个以上的对象，该对象具有数据的一个以上结点和应用程序的一个以上结点，且结点被层次化*"

Is this Chinese phrase as a noun phrase? Ordinary grammar of Chinese cannot accept this sentence as a noun phrase. This type of translation, however, is done many cases in patent translation. How we can solve the problem?

3. Sets of Sentences

3.1 Phrase set of Syntax of Sentence

It is well-known that a sentence can be expressed as tree of some phrases, called a syntax tree as

> ((*A (first (aspect of (the (present invention))))) is directed to ((an ((information processing) system)), including)(a (first ((information processing) apparatus))) and (a ((second (information processing) apparatus))))))*"

Let *Sy(s)* be the set of all phrases of appeared in the syntax tree of sentence s and we have $Sy(s) \subset Ph(s)$. We collect all phrases of sentences appeared in set of sentences S as follows:

$$Ph(S) = \cup \{ Ph(s) \mid s \in S \}$$

We can define the set of the set of all phrases appeared in syntax tree of sentences S as follows:

$$Sy(S) = \cup \{ Sn(s) \mid s \in S \}$$

and we have

$$Sy(S) \subset Ph(S) \text{ for any sentence set } S.$$

3.2 Phrase Pattern Set

In a phrase set *Sy(S)* of sentence set *S*, we have phrases that share a sub-phrase. For example, we have many sentences that share

> *A first aspect of the present invention is directed to …*

in patent documents(, this is a typical expression for a claim of patent.). This type of sub-string is called *a phrase pattern* and expressed as a strung function as follows:

> *A-first-aspect-of-the-present-invention-is-directed-to([N]);*

Then we can express the first sentence as follows:

> *"A first aspect of the present invention is directed to an information processing system"*
> *= A-first-aspect-of-the-present-invention-is-directed-to([an information processing system])*

A phrase function has a type that is POS of enumerated function. The POS is just 4 types of *N* for noun, *C* for complement, *P* for predicate and *S* for sentence in our function. We shall call such kind patterns that are created by syntax phrases *Sy(S)* for sentence set *S* as the phrase patterns and denoted by *PhP(S)*. Phrase pattern can be contained a word itself. If *PhS* is a subset of *PhP(S)*, we can define a set of sentences *Sn(PhS)*.

3.3 Sentence sets

Let ***S(L)*** be the set of correct sentences in language ***L***. There is a subset *F* of *S(L)* that is consisted of inadequate sentences to be translated, e.g., too-long sentences and ambiguous sentences even if the sentence is not illegal grammatically. This type of sentences should be classified into set F in this paper. There may be a set *D(L)* of sentences in *S(L)* − *F* having the following conditions:

- (Reversible translation condition)
  There is a translation $g : D(L) \rightarrow D(L')$ for any pair of languages *L* and *L'*, that g(s) and g(s') has the same meaning and same syntactical structure.
- (Minimum set)
  There is only one sentence with the same meaning, that is
  If *g(s) = g(s')* in D(L') then *s = s' in D(L) for any s and s' in D(L)*.
- (Sufficiency)
  For any sentence s in *S(L)* − *F, there is a sentence s' in D(L) so that s and s' have the same meaning, but not necessary to have the same syntactical structure. This sentence s' is*

called canonical form of s in this paper.

The set of D is called a canonical set of sentence in language L. This definition of a canonical sentence set drives a function from $S(L) - F$ to D and denoted by *c*. There are two or more canonical sets for a language. For example, there are different styles, e.g., oral conversation style, official document style and childrenese.

3.4 Natural sentence set

In *S(L) － F, there is a set of sentences that is widely accepted* among native speakers. Usually, corpus sentences are selected from the set of such sentences. This paper refers such kind sentence set to a set of all natural sentences in language *L* and denotes *C(L)*.

3.5 What is translation of sentences?

By using the above sets of sentences, we can define an activity translation as follows: Translation from language *S(L)* to *S(L')* is a function t satisfies the following conditions:

- *If t(s) = s' and s ∈ Nat(L) then s' ∈ Nat(L'),*
- *For any s is in S(L), there exists s' in Nat(L') so that*
  g(c(s)) = c(t(s))

In the example of the above sentence, we have

s = "第一对象存储部，可存储一个以上的对象，该对象具有数据的一个以上结点和应用程序的一个以上结点，且结点被层次化" ） ∈ Nat("N")
s'= "A first object storage portion can store at least one object having at least one data node and at least one application node that are hierarchical." ∈ Nat("E")
and t(s)=s',

because

c(s)="第一对象存储部，可存储一个以上的对象，该对象具有一个以上的数据结点和一个以上的应用程序结点，且结点被层次化) ∈ D
g(c(s)) = s' ∈ D' ∩ Nat("E")

and *c(s)* and *s'* have the same syntactical structure as table 1.

Table 1: Syntactical Structure

S0=_- *can-store-_([N1],[N2]);*	S0=_，*可存储_ ([N1],[N2]);*
N1= "A first object storage portion";	*N1= "第一对象存储部";*
N2=_-that- are hierarchical([N3]);	*N3=_具有_([N4],[N5]);*
N3=_-having-_([N4],[N5]);	*N2=_,且结点被层次化([N3]);*
N4= at-least-one-_([N6]);	*N4= 一个以上的_([N6]);*
N5=_-and-_([N7],[N8]);	*N5=_和_([N7],[N8]);*
N6=object;	*N6=对象;*
N7= at-least-one-_([N9]);	*N7= 一个以上的_([N9]);*
N8= at-least-one-_([N10]);	*N8= 一个以上的_([N10]);*
N9=data-object;	*N9=数据结点;*
N10=application-object;	*N10=应用程序结点;*

4. Translation of Patent Documents

We shall show the overview of our translation system based on the functional approach. The keys are the function dictionary and parsing algorithm.

4.1 Statistics of Patent Data for NTCIR-9

The structure of function dictionary includes about 37,468 functions from the training data of Japanese-English patent sentences supplied by NTCIR-9 project as follows:

● 389,146	Sentences
● 28,708	Nouns
● 2845	Verbs patterns
● 86	Conjunction patterns

- 1564 Sentence Final Expressions
- 417 Adverbs

Then the function dictionary excluding nouns is less than 10,000 functions. It is not difficult to construct functional dictionaries. Also Japio and Kipo collected about several hundred individual patterns from their patens, it is enough to collect just 50,000 phrase patterns excluding nouns.

4.2 Parsing algorithm of patent sentences

The parsing can be done by peering algorithm proposed by the authors [Ikeda et al, 2011]. The peering algorithm takes top-down analysis and solve the diversity program of computing, especially long sentences such as patent claim statements.

4.3 Translation of English Patent into Chinese

If we prepare the Chinese function set corresponding to the set of English functions mentioned in Section 2.2 as follows:

S0=本第一发明的_，其具备_和_，其中，
_([N1],[N2],[N3],[S4]);
N1=信息处理系统;
N2=第一信息处理装置;
N3=第二信息处理装置;
S4=_具备：_;_具备：_; ([N5],[N6],[N7],[N8]);
N5=上述第一信息处理装置;
N6=_；_；_；_；_([N9],[N10],[N11],[N12],[N13]);
N7=上述第二信息处理装置;；
N8=_;_；_：_；_([N14],[N15],[N16],[N17]);
N9=_,_([N18],[S19]);
N18=第一对象存储部;
S19=可存储_([N20]);
N20= _具有_ ([N21],[N22]);
N21=一个以上的_([N23]);
N22=_被层次化([N24]);
N23=该对象;
N24=_和_([N25],[N26]);
N25=一个以上的_([N27]);
N26=一个以上的_([N28]);
N27= 数据的结点;
N28=应用程序结点;

Then we have translated the English sentence into Chinese as follows:

“本第一发明的信息处理系统，其具备第一信息处理装置和第二信息处理装置，其中，上述第一信息处理装置具备：第一对象存储部，可存储一个以上的对象，该对象具有数据的一个以上结点和应用程序的一个以上结点，且结点被层次化；第一指令接收部，从上述第二信息处理装置接收对象指令，该对象指令是具有用于识别对象的对象标识符的指令，是对该对象的指令；第一对象不存在信息取得部，当由上述对象指令所具有的对象标识符进行识别的对象不存在时，取得该对象不存在的信息；第一处理部，当由上述对象指令所具有的对象标识符进行识别的对象存在时，根据上述对象指令来处理对象；第一发送部，将上述第一对象不存在信息取得部取得的信息即结果信息，或者上述第一处理部处理的结果的信息即结果信息，发送给上述第二信息处理装置；上述第二信息处理装置具备：第二不存在时处理信息存储部，可将对象标识符和不存在时处理标识符成对地存储，该不存在时处理标识符用于识别由该对象标识符识别的对象不存在时进行的处理；第二接收部，接收上述第一信息处理装置发送的结果信息；第二处理部，当上述第二接收部接收的结果信息为对象不存在的信息时，执行由对应于该对象的对象标识符的不存在时处理标识符所识别的处理；第二输出部，当上述第二接收部接收的结果信息不是对象不存在的信息时，输出该结果信息。”

5. Supporting patent document writing

The functional approach is also useful in writing patent documents. The detail is introduced in the following paper [Ikeda, WTIM’ 2011]. The process is done by Left-right input approach to write sentences using phrase patterns registered in the function dictionary. In this case, we need about 300 sentence patterns as phrase functions and about 3000 definition sentence patterns added to the common function dictionary.

6. Conclusion and Future Works

This paper introduced a new approach for patent document translation and writing. This approach has a possibility to solve long-term issues of the quality of machine translation. In order to implement this approach, construction of the function dictionary is necessary. But we have never finished it yet. This is our future work.

References

Alagin(Advanced Language Information Forum) . 2009. http://www.alagin.jp/purpose-e.html

Barkley Aligner. 2009. *A word alignment software package for machine translation*. http:// code. google. com/p / berkeleyaligner/

Babych, B., Hartley, A. 2003. *Improving Machine-Translation Quality with Automatic named EntityRecognition*. Proc. EACL-EAMT, Budapest.

Brown, P. F., Pietra, S. A. D, Pietra, V. L. D., and Mercer, R. L. 1993. *The mathematics of statistical* *machine translation: Parameter estimation*. Computational Linguistics, 19(2):263–311, June.

Chiang D. 2007. *Hierarchical Phrase-Based Translation.* Computational Linguistics, Volume 33, Number 2, Association for Computational Linguistics.

GIZA++. 2003. *GIZA++ statistical translation models toolkit*. http://code.google.com/p/giza-pp/

Ikeda, Hideto. Nguyen Thanh Hung and Ze Zhong Li. 2011. *Functional Language Modeling in Machine translation system for Chinese, Japan, Korean, Vietnamese and English.* Submitted to MT SUMMIT 2011.

Ikeda, Hideto. Nguyen Thanh Hung Nam, Nguyen Xuan and Ze Zhong Li. 2011. Multilingual Sentence Input System by Functional Language Model. WTIM' 2011.

Japio(Japanese Patent Office) Year's Book in 2010.

KIPO(Korean Intellectual Property Office) : http://www.kipo.go.kr/kpo/user.tdf?a=user.english .html.HtmlApp&c=91000&catmenu=ek02_01_01

MacCarthy, J. T. 2001. *MacCarthy's Desk Encycropedia of Intellectual Property*. Second Edition. BNA books.

Mel'čuk. 2003. *Levels of Dependency in Linguistic Description: Concepts and Problems*. In V. Agel, L. Eichinnger, H.-W. Eroms, P. Hellwig, H. J. Herringer, H. Lobin (eds): Dependency and Valency. An International Handbook of Contemporary Research, vol. 1, Berlin - New York, W. de Gruyter, 188-229.

Nagao, M. 1984. *A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in Artificial and Human Intelligence*, A. Elithorn and R. Banerji (eds.) North- Holland, pp. 173-180.

NiCT(National Institute of Information and Communications Technology). 2010. *Nict-EDR.*

NTCIR-9 Home page: http://research.nii.ac.jp/ntcir/ntcir-9/index.html

Och, F. J., and Ney, H. 2004. *The alignment template approach to statistical machine translation*. Computational Linguistics, 30(4):417–449.