MT Summit XI Workshop

Workshop on Patent Translation

September 11, 2007
Copenhagen, Denmark

September 11, 2007

Workshop on Patent Translation

Prof. Shoichi Yokoyama (Yamagata University)

Patent documents are one of the major application areas of machine translation. The workshop aims to foster research and development of the technology for patent translation by providing a forum in which researchers and practitioners can exchange their ideas, approaches, perspectives, and experiences from their work in progress. Double submission of papers to both the workshop and the main conference will also be welcomed. The workshop will consist of 1-3 invited talk(s), presentation of submitted papers, and one panel discussion.

Topics of interests include, but are not limited to:

- Analysis and classification for patent documents,
- MT and translation aids for patent documents,
- Contrastive studies for multilingual patent documents,
- Language resources for patent translation,
- Information extraction from patent documents,
- Evaluation techniques for patent translation,
- Multilingual patent classification and retrieval.

September 11, 2007

MT Summit XI Workshop on Patent Translation

List of program committee

Workshop Co-Chairs

Jun'ichi Tsujii	Univ. of Tokyo, Japan
Shoichi Yokoyama	Yamagata Univ., Japan

Program Committee

Terumasa Ehara	Tokyo Univ. of Science, Suwa, Japan
Laurie Gerber	Language Weaver, Inc., USA
Chikara Hashimoto	Yamagata Univ., Japan
Munpyo Hong	Sungkyunkwan University, Korea
Hiroyuki Kaji	Shizuoka Univ., Japan
Akira Kumano	Toshiba Corporation, Japan
Sadao Kurohashi	Kyoto Univ., Japan
Shinichiro Miyazawa	Shumei Univ., Japan
Hiroshi Nakagawa	Univ. of Tokyo, Japan
Naoya Oku	Japan Patent Information Organization, Japan
Tadaaki Oshio	Japan Patent Information Organization, Japan
Birgit Pichat	Lingtech A/S, Denmark
Svetlana Sheremetyeva	Lanaconsult, Denmark
Sayori Shimohata	Oki Electric Industry Co., Ltd., Japan
Eiichiro Sumita	ATR, Japan
Akira Ushioda	Fujitsu Laboratories, Ltd., Japan
Takehito Utsuro	Univ. of Tsukuba, Japan

September 11, 2007

MT Summit XI Workshop on Patent Translation

Program

9:00- 9:15	**Opening Remarks**		
9:15-10:30	**Invited Talk**	(Chair: S. Yokoyama)	
	WIPO's activities in patent translation and terminology	Rachel Chrem (WIPO)	
10:30-10:45	**Coffee Break**		
10:45-12:30	**Session I**	(Chair: S. Sheremetyeva)	
	English-Korean Patent Translation System: From To-EK/PAT	Oh-Woong Kwon Sung-Kwon Choi Ki-Young Lee Yoo-Hyung Roh Young-Gil Kim	1
	Phrase Alignment for Integration of SMT and RBMT Resources	Akira Ushioda	8
	Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation	Terumasa Ehara	13
12:30-13:30	**Lunch**		
13:30-15:00	**Talk by Users**	(Chair: T. Ehara)	
	to be announced	Wolfgang Taeger (EPO) Takahiko Tohyama (JPO)	
15:00-15:30	**Coffee Break**		
15:30- 17:00	**Session II**	(Chair: A. Ushioda)	
	Patent Documentation -- Comparison of Two MT Strategies	Lene Offersgaard Claus Povlsen	19
	Error Correcting System for Analysis of Japanese Patent Sentences	Shoichi Yokoyama Shigehiro Kennendai	24
	On Portability of Resources for Quick Ramp up of Multilingual MT for Patent Claims	Svetlana Sheremetyeva	28
	Wrap-up	(Chair:)	

Author Index

English-Korean Patent Translation System: FromTo-EK/PAT

Oh-Woog Kwon, Sung-Kwon Choi, Ki-Young Lee, Yoon-Hyung Roh, Young-Gil Kim

Natural Language Processing Team, Electronics and Telecommunications Research Institute
161 Gajeong-dong, Yuseong-gu,
Daejeon, Korea, 305-350
{ohwoog, choisk, leeky, yhnoh, kimyk}@etri.re.kr

Munpyo Hong
Dept. Of German Language & Literature Sungkyunkwan Univ.
53 Myeongnyun-dong 3-ga, Jongno-gu, Seoul, Korea, 110-745
skkhmp@skku.edu

Abstract

This paper addresses a method for customizing an English-Korean machine translation system from general domain to patent domain. The customizing method includes the followings: (1) extracting and constructing large bilingual terminology and the patent-specific translation patterns, (2) adapting the probabilities of POS tagger trained from general domain to the patent domain, (3) syntactically analyzing long and complex sentences by recognizing coordinate structures, and (4) selecting a proper target word using patent-specific bilingual dictionary and collocation knowledge extracted from patent corpus.

The translation accuracy of the customized English-Korean patent translation system is 82.43% on the average in 5 patent categories (machinery, electronics, chemistry, medicine and computer) according to the evaluation of 7 professional patent translators. A patent MT system for electronics domain was installed and started an on-line MT service in IPAC (International Patent Assistance Center) under MOCIE (Ministry of Commerce, Industry and Energy) in Korea. In 2007, KIPO (Korean Intellectual Property Office) is expected to launch its English-Korean MT service for whole patent domain.

1. Introduction

Given the growing number of foreign language patents filed in the multiple countries, it is feasible that users want to read the patent documents translated to their native language. Such users' demand has become a hot research issue in the MT community. Also because NLP techniques associated with specificity of patent domain have promise for improving the translation quality, patent translation is recently attracting many researchers and MT-related companies.

It is well known that sentence style and dominant translation for a word vary with domains. Therefore, if the domain to be translated is fixed to patents, bilingual dictionary adaptation to the patent domain and customizing natural language analyzers to the linguistic specificity of patent style are effective ways to improve the translation quality of MT system. There have been studies concerned specifically with patent MT using these domain-specific advantages (Shinmori et al., 2003; Hong et al., 2005; Kaji, 2005; Shimihata, 2005).

Though intensive research has been made on patent MT for the domain-specific advantages, there still remain many issues to be tackled. In this paper, we focus on the several issues: (1) new terminology construction, (2) patent-specific probabilities of POS tagger, (3) long and complex sentence analysis, and (4) target word selection.

This paper addresses the customization of an English-Korean MT system for patent translation. The English-Korean patent MT system "FromTo-EK/PAT" described in this paper is based on an English-Korean MT system developed for the web translation in a general domain. English-Korean patent MT system belongs to basically the pattern-based methodology for machine translation. It has the formalism that does English sentence analysis in which English patent-specific patterns are used, matches the English patent pattern with its Korean patent pattern, and then generates a Korean sentence from it. English-Korean patent MT system consists of an English morphological analysis module based on lexicalized HMM, an English syntactic analysis module by pattern-based full parsing, a pattern-based transfer, and a Korean morphological generation.

Section 2 describes the issues of customizing a MT system to the patent domain. In section 3 we will introduce the customization process according to the issues described in section 2. The experimental work is presented in section 4. Lastly, in section 5, we present some conclusions.

2. Issues for Customizing MT System to Patent Domain

It is important to customize translation knowledge and translation modules for adapting the existing general MT system to translation of patent documents. The customization for the translation knowledge is able to be divided into two steps: (1) tuning general translation knowledge to patent-specific translation knowledge, and (2) efficiently constructing the unknown words and the translation patterns found in patent documents. The patent customization of existing translation knowledge is closely related with the customization of the translation knowledge of module. For example, the customization of the module of target word selection is decided by the customization of existing English-Korean bilingual dictionary. The POS tagging knowledge trained from general domain also have an influence on the customization of the POS tagging module. In this respect we consider the method extracting unknown words from

patent documents and the method customizing translation modules to patent.

What is firstly necessary for applying a general MT system to patent is to extract the large-scale terms found newly in patent documents and construct their translation knowledge such as the target words. We have built an English-Korean bilingual dictionary by use of exiting Korean-English bilingual dictionary of a Korean-English patent MT system developed in 2005, in order to cut cost and time for building an English-Korean bilingual dictionary. The unknown words could be constructed at maximum effect with little cost and little time by the method, where we preferred selecting the high-frequently and positively necessary words for the English-Korean translation to constructing all unknown words appearing in patent documents.

In relation to POS taggers with good performance and broad coverage, they have recently become available (Brants, 2000; Pla et al., 2004), but have not been trained for patent documents. This means that there is room for doubt that the general POS taggers keep their performance in the patent domain. We can easily find an example to degrade the performance, only looking through any patent document. The example is the word "said": the word is mainly used as a past verb (VBD) in general domain, but is almost used as a adjective (JJ) in patent domain. The words like "said" are retrained from a tagged patent corpus. It is however very difficult to construct the tagged patent corpus because we have no tagged patent corpus. In this paper, we will describe how to adapt the general-purpose POS tagger to the patent domain by using raw patent corpus.

Compared with general documents, one characteristic of patent documents is to use the abnormally long and complex sentences (Kando, 2000), which makes it difficult to apply a parser for general domain to patent domain. A usual method for treating long sentences is to segment a long sentence into several segments and to analyze each segment respectively. However, in case a long sentence is formed by coordination structure, simple segmentation can cause syntactic analysis errors if the coordination structure is not firstly recognized. For this, we will present a method for recognizing the coordination structure in patent documents to enhance parsing efficiency and performance.

Target word selection in English-Korean machine translation is very important factor in that it has a direct influence on the machine translation quality. Particularly, in the case of general domain documents such as web pages, the target word selection problems of English ambiguous words occur very frequently. In general domain documents, many frequently used English words can be translated to various Korean words depending on the contexts. However, in English-Korean patent machine translation, most of words used in patent documents belong to technical terms. These technical terms have relatively low ambiguities of target word selection. Some English words used in patent domain also have a tendency to be translated to specific Korean word according to International Patent Classification (IPC) codes. Although patent documents include many technical terms, target word selection problem still remains an obstacle which should be solved to improve the performance of machine translation system. We customized English-Korean dictionary for patent machine translation to resolve the translation ambiguity of English ambiguous words appearing in patent documents. So, some English ambiguous words contain dominant Korean target word according to specific IPC code. For target word selection ambiguities which did not resolved by dominant Korean target word of translation dictionary, we tried to disambiguate the possible senses of English words by use of other knowledge like sense vectors and Korean bi-gram context information.

3. Customizing Methods

3.1 Construction of Patent Terminology

Terminology construction for English-Korean patent MT system described in this paper is similar to the methods of Kaji(2005), Shimohata(2005), and Kim(2005) in respect of using the existing dictionary and the existing patent corpus, but our method is different in that it contains a step inverting the existing Korean-English bilingual terminology. Extraction and construction of terminology might be represented in Figure 1.

As shown in Figure 1, the patent terminology can be built by two steps. The first step is the step to convert the existing Korean-English terms into the English-Korean terms, to delete the terms overlapped with the terms in the existing English-Korean bilingual dictionary, and to construct the English-Korean bilingual terms semi-automatically. Among inverted English-Korean bilingual terms, if English terms are the nominal phrases including a prepositional phrase, a gerund, and a relative clause, they are deleted. These nominal phrases were constructed for lack of an English compound word suitable to a Korean compound word in Korean-English patent translation. If such nominal phrases are entered in the English-Korean dictionary, the structural errors such as attachment of prepositional phrase or analysis of coordination structure in parsing might be produced. For example, if "method for 1+1 line protecting switching" as an English term equivalent to Korean term "1+1 선로 보호 절체 방법" is made an entry of English-Korean dictionary, it may give rise to the incorrect analysis of coordination structure "(NP (NN device) (CC and) (NN method for 1+1 line protecting switching))" in analysis of a English phrase such as "device and method for 1+1 line protecting switching".

Each English term in the English-Korean terms constructed by the first step may have different Korean target words. To select a dominant one among different Korean target words, we sorted Korean target words automatically according to their frequency occurring in Korean patent documents and made a selection of dominant target word manually. Through this work we could create 801,046 English-Korean terms from 3,052655 Korean-English terms.

The second step is to extract the unknown words from 1,001,419 English patent documents applied to the U.S. Patent Office from 2001 to 2005 and remove the overlapped entries. We extracted about ten million English unknown words from this step, but manually constructed 1,039,189 English-Korean bilingual

terminology with high coverage by using the method 'Setting Lexical Goals' Hong(2005) presented.

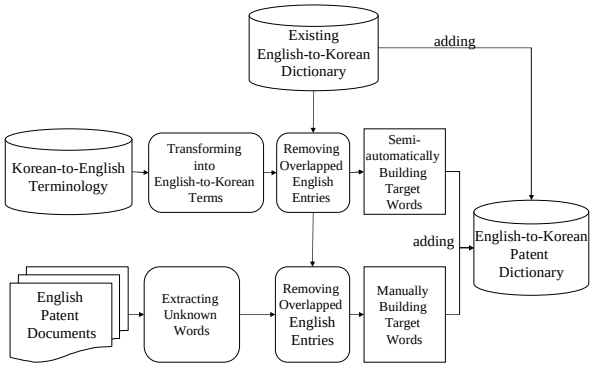


Figure 1: Customization process for building English-Korean patent terminology

3.2 A Domain Adaptation Method for POS Tagger

Three items were tuned for customizing a broad coverage POS tagger based on HMM to patent domain. They are as follows:

- For customization of surface form, a tokenization module and/or a morphological analyzer were modified for tokenizing and/or analyzing the peculiar surface forms found in the specific domain.
- For customization of lexical information, lexical probabilities (output probabilities) were tuned for holding domain-specific lexical information.
- For customization of context information, contextual probabilities (transition probabilities) were controlled for holding the domain-specific contextual information.

In the first step 'customization of surface form', the tokenization module was modified to tokenize and/or chunk very complex symbol words, a chemical formula, a mathematical formula, programming codes, and so on. We improved our morphological analyzer to assign the estimated part-of-speeches to a compound word connected with hyphen or slash. The estimated part-of-speeches are estimated by the part-of-speeches of their components.

Our English POS tagger uses a lexicalized HMM (Pla et al., 2004). The process of our POS tagger consists of finding the sequence of POS tags of maximum probability, that is:

$$\overline{T} = \arg\max_{t_1 \cdots t_n} \left(\prod_{1 \cdots n} P(t_i \mid t_{i-1}, t_{i-2}) \cdot P(w_i \mid t_i) \right) \quad (1)$$

for given sequence of words $w_1, \ldots, w_n$ of length n. $t_1, \ldots, t_n$ are elements of the tagset, the additional tags t_{-1}, t_0, and t_{n+1} are beginning-of-sequence and end-of-sequence markers. In this equation, lexical probability is $P(w_i|t_i)$, and contextual probability is $P(t_i|t_{i-1},t_{i-2})$. The lexical and contextual probabilities are estimated from tagged corpus.

The best simple strategy for the second and third customization phase is to re-estimate lexical and contextual probabilities from very large tagged patent corpus. However, there is not a tagged patent corpus and it is also very difficult to construct it. For customizing the lexical and contextual probabilities, we used a raw patent corpus consisting of about one million U.S. patent documents. First, we tagged automatically the words of the raw corpus with our POS tagger and estimated lexical probability $P'(w_i|t_i)$ and contextual probability $P'(t_i|t_{i-1},t_{i-2})$ from the machine-tagged patent corpus. Next, we extracted the high-frequent lexemes having $abs(P(w_i|t_i)-P'(w_i|t_i))$ greater than arbitrary threshold value and the high-frequent contextual n-grams having $P(t_i|t_{i-1},t_{i-2})$ less than arbitrary threshold value. The extracted lexical and contextual n-grams are tuned by the three human experts for two months. For customization of our POS tagger, we tuned about 6,000 lexemes and about 1,500 tri-grams.

It is difficult that a expert perceives the exact meaning of the output probability, because lexical probability, $P(w_i|t_i)$, corresponds to the output probability in which the word w_i is generated given POS t_i. But, the expert could easily decide whether a word w_i is used as POS t_p more frequently than POS t_q in the patents, or not. In this view point, the expert can more easily and correctly tune $P(t_i|w_i)$ than $P(w_i|t_i)$ for each extracted word w_i. To customize lexical probabilities to patent domain, the experts adjusted $P(t_i|w_i)$ examining the POS tagged sample sentences. Then, we calculated $P(w_i|t_i)$ by using the tuned $P(t_i|w_i)$ as follows:

$$P(w_i \mid t_i) = P(t_i \mid w_i) \times f(w_i) / f(t_i) \quad (2)$$

For customization of the context information, the experts selected correct n-grams from the extracted n-grams. To estimate the selected context probabilities $P(t_i|t_{i-1},t_{i-2})$, we first find $P'(t_p|t_{p-1},t_{p-2})$ that is the nearest probability to $P'(t_i|t_{i-1},t_{i-2})$. Then we calculated $P(t_i|t_{i-1},t_{i-2})$ as follows:

$$P(t_i \mid t_{i-1}, t_{i-2}) = P'(t_i \mid t_{i-1}, t_{i-2}) \times \frac{P(t_q \mid t_{q-1}, t_{q-2})}{P'(t_q \mid t_{q-1}, t_{q-2})} \quad (3)$$

The representative tri-grams among the extracted n-gram are "NN CD VBZ" and "NNS CD VBP". They mean that a cardinal number comes before a verb in patent documents, while a cardinal number basically comes before a noun in general documents. In the patent documents, a cardinal number after a noun denotes almost always a reference mark for a diagram or a box in a figure. For example, in the sentence "Another management chip connected to *pad 117 controls* the parallel port 102b and the serial ports 104c and 104d.", the cardinal number "117" points out the box corresponding to the pad apparatus in a figure.

3.3 Syntactic Analysis for Patent Document

Two most important ones among peculiar syntactic characteristics of patent documents are the frequent use of patent-specific patterns and the abnormally long sentences (Shinmori et al., 2003). Considering these characteristics as central features, I will describe the main contents of syntax analysis for patent documents in detail.

3.3.1 Application of patent-specific patterns

We applied patent-specific patterns before parsing to reduce a parsing complexity. A general form of the patent-specific patterns is composed of some lexical words and some syntactic nodes as shown in a sample of below pattern.

1) *The method for VP , wherein S*

For the recognition of the patterns, lexical words are firstly matched, and the ranges between the lexical words are recognized as tentative syntactic nodes. Assuming that above pattern is applied to a example sentence 2), "the method for" is matched, the word strings between "for" and "," are recognized as a verbal phrase(VP) and the matching of next lexical symbols ", wherein" is attempted.

2) "*The method for controlling the flow in the micro system according to claim 1, wherein the stimulation is a voltage.*"

Actually, we conduct simple condition check to know whether the word strings can be VP or not. If the pattern matches wholly with the input sentence, a parsing with all the tentative nodes is attempted. If all nodes are successfully parsed into the corresponding syntactic nodes in the translation pattern, the syntactic pattern is recognized finally. As a result, the actual parsing ranges are reduced to parsing of two clauses such as "controlling the flow in the micro system according to claim 1" and "the stimulation is a voltage".

3.3.2 Recognizing coordinate construction

The usual method for treating long sentences is to segment a long sentence into several segments by use of syntactic clues or some other conditions (Kim et al., 2001). However, the segmentation method is applicable only in case that segments resulting from segmentation don't have any hierarchical relation between each other. In case of sentences formed by coordination of syntactic nodes such as NP, VP, that-clause, etc., if a sentence is segmented between coordinate constituent nodes, segmentation can cause syntactic analysis errors, because a segment can be dependent on some other node in parse tree.

For example, in the example sentence 3), the sentence can be segmented at the positions such as ", collecting" or ", driving". But verb phrases starting at those positions are objects of the verb "comprising", so such dependency relation is broken by segmentation.

3) A method of operating a transaction system which comprises a plurality of currency acceptors, the method comprising installing the acceptors in host machines, performing individual transactions using the machines, collecting performance data from the acceptors, performing a statistical analysis on the performance data from the acceptors, deriving re-configuration data for at least one acceptor as a result of the statistical analysis and re-configuring said at least one acceptor on the basis of the re-configuration data.

Therefore, we need to recognize coordination structures first before segmentation. Sadao K. and Makoto N. (1994) detected conjunctive structures in a general domain using dynamic programming. Compared with coordinate structures in the general domain, a typical feature of coordination structures in patent documents is that the coordinate structures have a lot of coordinate constituent nodes like VPs in the example sentence 3). Sometimes, each node has very complex structure, which makes the recognition of coordination structure very difficult. So, we have introduced a method of recognizing coordination structure using similarity table. The similarity table is a table which stores similarities between all the possible nodes constituting candidate coordinate structures. All starting positions of possible nodes constituting the candidates of coordination structures are recognized by syntactic clue such as NP or verb followed by "comprise, include, have, etc.". The similarity between nodes is calculated by syntactic similarity and some other factors. Once the similarity table is constructed, all the candidates of coordination structures are searched and their weights are calculated by the similarity table. Finally, the coordinate structure with maximum weight becomes a final result. The sentence is simplified because the recognized coordination construction is chunked to one node. The example sentence 3) is reduced to " A method of operating a transaction system which comprises a plurality of currency acceptors, the method comprising VP."

3.3 Customization for Target Word Selection

We approached target word selection problems in patent machine translation in two ways considering knowledge and engine. For adapting English-Korean bilingual terms to patent domain, we first defined 5 patent categories such as mechanics, chemicals, medicals, electronics and computers and mapped all IPC codes to 5 patent categories. Next, we reconstructed translation dictionary putting the dominant translation word according to 5 patent categories. For this reconstruction process, we made a collection of each 5 patent corpus using a mapping table between IPC codes and 5 categories. And then, we extracted English ambiguous words with high frequency. For these extracted English words, human patent translator registered dominant Korean word by hands considering each category. Our patent machine translation system receives IPC code of an input patent document as a parameter and decides proper Korean target word by it.

For the ambiguous English words which did not resolved by dominant Korean word of translation dictionary, we made a target word selection module using context knowledge constructed from corpus. We extracted context information from English-Korean comparable corpus. The context information was converted to sense vectors. The sense means Korean translation word for the ambiguous English word. The sense vectors were used to disambiguate the possible senses of ambiguous English words (Lee et al., 2006). Sense vector is defined by the following formula:

$$SV = (w(c_1), w(c_2), w(c_3), ..., w(c_n)) \quad (4)$$

where $w(c_k)$ is a weighting function for co-occurring word c_k. And $w(c_k)$ can be calculated by the following formula:

$$w(c_k) = \Pr(s = s_i | w = c_k) \quad (5)$$

where s_i is an i-th sense (a group of target words sharing same semantic code) of source word. When $w(c_k)$ is 1, it means that if co-occurring word c_k appears with ambiguous word, the probability that the sense of ambiguous word will be s_i is 1.

In the test phase, the test vector for ambiguous word in input sentence is constructed and has same dimension as the sense vector of the corresponding ambiguous word. The elements of test vector are 0 or 1, where 0 indicates that corresponding co-occurring word c_k does not appear in the input sentence and 1 represents that corresponding co-occurring word c_k appears in the input sentence. The similarity between test vector constructed from input sentence and each sense vector of the ambiguous word is calculated using following formula:

$$sim(v,w) = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2 \sum_{i=1}^{N} w_i^2}} \quad (6)$$

Also, we extracted Korean bi-gram information from Korean monolingual corpus. Korean bi-gram information is used to decide the most proper Korean translation word in final generation phase of our system.

4. Experiments and Evaluation

4.1 Translation Accuracy Evaluation

In this section, we describe the evaluation about translation quality of English-Korean patent MT system. We used the following test sentences, evaluation method and evaluation criterion for translation quality:

- Test sentences: translation accuracy was assessed with 100 test sentences for each one of 5 patent categories (machinery, electronics, chemistry, medicine and computer). Among 100 sentences for each patent category, about 54 sentences were selected from the "detailed description" section of patents, 24 were extracted from the "claim" section, the rest from the "description of the drawing" and the "background of the invention" section. The average length of a sentence was 28.33 words.
- Evaluation criterion:

Score	Criterion
4	The meaning of a sentence is perfectly conveyed
3.5	The meaning of a sentence is almost perfectly conveyed except for some minor errors (e.g. wrong article, stylistic errors)
3	The meaning of a sentence is almost conveyed (e.g. some errors in target word selection)
2.5	A simple sentence in a complex sentence is correctly translated
2	A sentence is translated phrase-wise
1	Only some words are translated
0	No translation

Table 1: Scoring criteria for translation accuracy

- Evaluation method:
  - 7 professional translators evaluated the results. Ruling out the highest and the lowest score, the rest 5 scores were used for translation accuracy evaluation. The translation accuracy was defined as follows:

    : translation accuracy(%) = $\sum_{i=1}^{n}(\sum_{j=1}^{5}(score_j/4))/5)/n \times 100.0$, where n is the number of test sentences and $score_j$ is the score evaluated by the j-th professional translator.

Table 2 shows that the translation accuracy of English-Korean patent MT system was 82.43% on the average. Among the patent fields, the translation of the machinery field was best, while the translation of the medicine field scored worst. The reason for the best scoring of the machinery field is that patent-specific patterns were applied to most of sentences. The medicine field contained, as expected, many unknown words and incorrect target word selection. The number of the sentences that were rated equal to or higher than 3 points was 438. It means that about 87.60% of all translations were understandable.

Patent field	Average length of a sentence	Translation accuracy	Translation accuracy higher than 3 scores
machinery	30.34 words	83.50%	85.00%
electronics	29.42 words	82.20%	88.00%
chemistry	29.67 words	82.20%	91.00%
medicine	26.75 words	81.63%	86.00%
computer	25.49 words	82.63%	88.00%
average	28.33 words	82.43%	87.60%

Table 2: Translation accuracy for each patent field

4.2 Evaluation for Customization

We evaluated the performance the modules specialized to the patent domain, compared with the performance of our general-purpose modules. For the evaluation, we used 100 sentences of the electronics category among the whole translation evaluation test set.

Table 3 shows the word accuracy and sentence accuracy of two taggers: the POS tagger specialized to the patent domain (PatTagger) and our general-purpose POS tagger (GPTagger). From these results we can draw the following conclusions. First, the PatTagger reduced significantly the error tagging about 91% with respect to the GPTagger. Second, PatTagger improved the sentence accuracy with 41% compared with GPT3agger.

	GPTagger	PatTagger
Word tagging accuracy	95.85%	99.62%
Sentence tagging accuracy	50.00%	91.00%

Table 3: Comparison of the tagging accuracy between GPTagger and PatTager

Table 4 shows the performance improvement factors of PatTagger and the improved word accuracy according to the factors. The improvement factors of PatTagger are three customization phases mentioned in the section 3.2 and terminology construction mentioned in the section 3.1. The terminology construction is to add unknown words and their part-of-speeches into morphological analysis dictionary. The performance improvement of word supplement is very low because our POS tagger handles unknown words using suffix analysis as proposed in Brants(2000). From the results of table 4, the customization of lexical and context information is surely needed in order to specialize a general-purpose POS tagger based on HMM to a specific domain.

Table 5 shows the evaluation result by the customization of syntactic analyzer. In Table 5, the syntactic analysis accuracy is calculated by the ratio of the number of correctly analyzed sentences to the number of total sentences. We consider a sentence as correct when the

syntactic analysis result of the sentence has a trivial error that don't affect the translation result.

Table 6 shows the experimental results of target word selection of the customized MT system and the non-customized MT system. The percentage of unknown word is decreased in customized MT system by registering unknown words to translation dictionary consistently. We can see that how the unknown word can affect target word selection problems. At the same time, the customization of transfer module considering characteristics of patent domain can improve the performance of target word selection.

Table 7 is the result to compare the translation accuracy before customization with that after customization in the electronic patent document. In Table 7, the difference of translation accuracy between before customization and after customization in electronic patent document was 27.95%. This means that the customization process described in this paper made an important role to enhance the translation quality of English-Korean MT system on patent documents.

The performance improvement factor	The # of tagging error correction	The correction rate	The improvement of word tagging accuracy
Customization of surface form analysis	6	5.41 %	0.20%
Customization of the lexical information	81	72.97 %	2.75%
Customization of the context information	22	19.82 %	0.75%
Construction of Terminology	2	1.80 %	0.07%
Total	111	100.00 %	3.77%

Table 4: The performance improvement of PatTagger and the improvement of its word tagging accuracy.

	Syntactic analysis accuracy
General-purpose syntactic analyzer	69.0%
Customized syntactic analyzer	85.0%
ERR (Error Reduction Rate)	51.6%

Table 5: Evaluation of customization of syntactic analyzer

	Accuracy of target word selection for noun	Percentage of unknown word
Non-customized MT System	71.7%	16.3%
Customized MT System	92.4%	1.5%

Table 6: Result of target word selection for noun

Patent field	Translation accuracy before customization	Translation accuracy after customization
electronics	54.25%	82.20%

Table 7: Comparison of translation accuracy before customization with that after customization in electronic patent document

4. Conclusion

In this paper we described a method for customizing English-Korean machine translation system from general domain into patent domain. First, we described the construction method of the large English-Korean bilingual dictionary using the existing Korean-English bilingual dictionary and extracting unknown words from about one million patents. Secondly, to adapt general-purpose POS tagger to the patent domain, we proposed the method for semi-automatically adjusting probabilities trained from general domain to patent context using raw English patent documents. Thirdly, the syntactic analyzer is proposed for segmenting and analyzing long and complex patent sentences by recognizing coordinate structures. Lastly, we proposed the target word selection using patent-specific bilingual dictionary and collocation knowledge extracted from raw patent corpus.

The English-Korean patent MT system "FromTo-EK/PAT described in this paper was developed under the auspices of the MIC (Ministry of Information and Communication, Korea) during 2005-2006. FromTo-EK/PAT was installed in IPAC (International Patent Assistance Center) under MOCIE (Ministry of Commerce, Industry and Energy) in Korea and provides the patent attorneys and the patent examiners with the on-line English-Korean machine translation service for electro-electric patent documents (http://www.ipac.or.kr). In 2007, KIPO (Korean Intellectual Property Office) is expected to launch its English-Korean MT service for whole patent domain.

Acknowledgements

This work was supported by the IT R&D program of MIC/IITA, Domain Customization Machine Translation Technology Development for Korean, Chinese, and English.

Bibliographical References

Brants T. (2000). TnT – a statistical part-of-speech tagger. In Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000) (pp. 224--231).

Hong M.P., Kim Y.G., Kim C.H., Yang S.I., Seo Y.A., Ryu C. & Park S.K. (2005). Customizing a Korean-English MT System for Patent Translation. MT Summit X (pp. 181—187).

Kaji H. (2005). Domain Dependence of Lexical Translation: A Case Study of Patent Abstract. MT Summit X Workshop on Patent Translation.

Kando N. (2000) What Shall we Evaluate? Preliminary Discussion for the NTCIR Patent IR Challenge (PIC) Based on the Brainstorming with the Specialized Intermediaries in Patent Searching and Patent Attorneys. In Proceedings of the ACM SIGIR 2000 Workshop on Patent Retrieval in conjunction with The 23rd Annual

International ACM SIGIR Conference on Research and Development in Information Retrieval. Athens. Greece

Kim Y.K., Yang S.I., Hong M.P., Kim C.H., Seo Y.A., Ryu C., Park S.K. & Park S.Y. (2005). Terminology Construction Workflow for Korean-English Patent MT. MT Summit X Workshop on Patent Translation.

Lee K.Y., Park S.K. & Kim H.W. (2006). A Method for English-Korean Target Word Selection Using Multiple Knowledge Sources. IEICE TRANS. FUNDAMENTALS, Vol.E89-A, No.6.

Sadao K., Makoto N. (1994). A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structure. Computational Linguistics 20(4): 507-534.

Shinmori A., Okumura M., Marukawa Y. & Iwayama M. (2003). Patent Claim Processing for Readability - Structure Analysis and Term Explanation, ACL-2003 Workshop on Patent Corpus Processing.

Sung-Dong Kim, Byoung-Tak Zhang, and Yung Taek Kim. (2001). Learning-based Intrasentence Segmentation for Efficient Translation of Long Sentences. Machine Translation, 16(3):151-174.

Shimohata S. (2005). Finding Translation Candidates from Patent Corpus. MT Summit X Workshop on Patent Translation.

Pla F. & Molina A. (2004). Improving Part-of-speech Tagging Using Lexicalized HMMs. Natural Language Engineering 10(2) (pp. 167-189).

Phrase Alignment for Integration of SMT and RBMT Resources

Akira Ushioda

Software and Solution Laboratories
Fujitsu Laboratories Ltd.
4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki 211-8588
Japan
ushioda@jp.fujitsu.com

Abstract

A novel approach is presented for extracting syntactically motivated phrase alignments. In this method we can incorporate conventional resources such as dictionaries and grammar rules into a statistical optimization framework for phrase alignment. The method extracts bilingual phrases by incrementally merging adjacent words or phrases on both source and target language sides in accordance with a global statistical metric. Phrase alignments are extracted from parallel patent documents using this method. The extracted phrases used as training corpus for a phrase-based SMT showed better cross-domain portability over conventional SMT framework.

1. Introduction

In the phrase-based SMT framework (Marcu & Wong, 2002; Och & Ney, 2004; Chiang, 2005), extraction of phrase pairs is a key issue. Currently the standard method of extracting bilingual phrases is to use a heuristics called diag-and (Koehn et. al., 2003). In this method starting with the intersection of word alignments of both translation directions additional alignment points are added according to a number of heuristics and all the phrase pairs which are consistent with the word alignments are collected.

Although this method is effective by itself it is very difficult to incorporate syntactic information in a straight manner because phrases extracted by this method have basically little syntactic significance. Especially if we intend to combine strength of conventional rule-based approach with that of SMT, it is essential that phrases, or translation units, carry syntactic significance such as being a constituent (Yamada & Knight, 2001).

Another drawback of the conventional method is that the phrase extraction process is deterministic and no quantitative evaluation is applied. Furthermore if the initial word alignments have errors, these errors propagate to the phrase alignment process. In doing so the burden of statistical optimization is imposed on the final decoding process.

We propose in this paper a novel phrase alignment method in which we can incorporate conventional resources such as dictionaries and grammar rules into a statistical optimization framework for phrase alignment. For a statistical optimization to be in effect, it is preferable that the initial word alignments are numerical, not zero/one. Let's take a simplified example to obtain an intuition behind the proposed method. Consider the following Japanese-English parallel sentences.

(1a) ジョンは 白球を 投げた
(John- Nominative) (white ball-Accusative) (threw)

(1b) John threw white balls

Figure 1 shows the degrees of correspondence (scores) between each Japanese word/phrase and English word/phrase. The score in each cell is just an illustrative figure. In Figure 1(a) each Japanese word in (1a) is arranged in a row and each English word in (1b) is arranged in a column. The dominant cells are shadowed and they are considered to show a clear correspondence. For a pair of languages with similar word order, the corresponding cells tend to align diagonally, but for languages like Japanese and English which have quite different word order, the corresponding cells are scattered. Nonetheless, when we look at local correspondences like words within a phrase, the corresponding cells come to next to each other. In this representation, when we obtain

1	2	3	
98	1	1	1 John
0	2	98	2 threw
0	98	2	3 white
0	97	3	4 balls
ジョンは (John-Nom)	白球を(white ball-Acc)	投げた (threw)	

(a)

1	2	3	
98	1	1	1 John
0	2	98	2 threw
0	195	5	3 white balls
ジョンは	白球を	投げた	

(b)

1	2	3	
98	1	1	1 John
0	100	100	2 threw white
0	97	3	3 balls
ジョンは	白球を	投げた	

(c)

Figure 1: Phrase alignment example

a one-to-one correspondence in which one and only one dominant cell appears in each row and column, we can judge that we obtained a phrase alignment. It is rarely the case that we obtain a one-to-one correspondence at the initial stage (1-a). However when we repeat merging a pair of adjacent words (or phrases) on Japanese side and English side, and adding the score of merged rows or columns, then we will eventually arrive at a one-to-one correspondence, in a worst case leaving only one row and one column. In the example of Figure 1, when we merge two adjacent English words "white" and "balls", we reach a one-to-one correspondence (Figure 1-b). This is because "白球を" and "white balls" constitute a pair of phrases with no excess or deficiency on either side. On the other hand, when we merge "threw" and "white", the matrix goes away from the one-to-one correspondence. We present in the next section a formal framework of the proposed method.

2. Phrase Alignment Method

Although our objective in this work is to extract alignments of phrases which are linguistically motivated, there might be cases in which a phrase in one language in a pair constitutes a constituent while the corresponding phrase in the other language does not. Therefore the basic strategy we adopt here is to try to extract bilingual phrases whose source language side at least constitutes a constituent. As for the target language side, a preference is given to constituent constructs.

The phrase alignment method we propose here extracts bilingual phrases by incrementally merging adjacent words or phrases on both source and target language sides in accordance with a global statistical metric along with constraints and preferences composed by combining statistical information, dictionary information, and optionally grammatical rules.

2.1 Without Syntactic Information

We begin by describing the proposed phrase alignment method in the case of incorporating no syntactic information. Figure 2 shows the framework of the phrase aligner. In the case of incorporating no syntactic information, *Syntactic Component* in the figure plays no role. We take here an example of translating from Japanese to English, but the framework presented here basically works for any language pair as long as conventional rule-based approach is applicable.

As a preparation step, word alignments are obtained from a bilingual corpus by GIZA++ (Och & Ney, 2000) for both directions (source to target and target to source), and the intersection $A = A1 \cap A2$ of the two sets of alignments are taken. Then for each English word $\boldsymbol{e}$ and Japanese word $\boldsymbol{j}$, the frequency $N(e)$ of e in A and the co-occurrence frequency $N(e, j)$ of e and j in A are calculated. Furthermore, using a discrimination function $\delta(e, j)$ which determines whether e and j are a translation of each other with respect to a predefined bilingual dictionary, word based empirical translation probability is obtained as follows.

$$(2)\ Pc(j|e) = (N(e,j)\,\delta(e,j)) / (N(e) + \Sigma_t \delta(e,t))$$

$\delta(e, j)$ takes a value of 1 when (e, j) appears in the bilingual dictionary, and 0 otherwise.

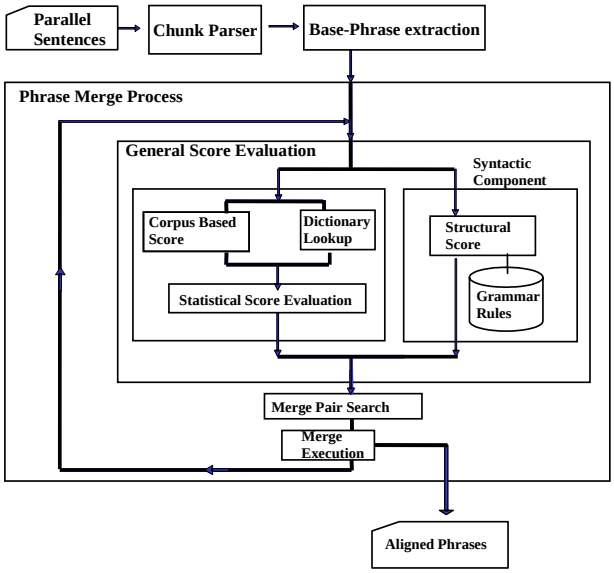


Figure 2: Framework of Phrase Aligner

An input to the phrase aligner is a pair **(J, E)** of Japanese and English sentence. The pair (**J**, **E**) is first chunk-parsed to extract base phrases, such as minimum noun phrases and phrasal verbs on both sides. Let $\boldsymbol{J = j_1, j_2, \ldots, j_M}$ be a series of Japanese chunks. These chunks are the minimum units for composing a final phrase alignment on Japanese side. Let $\mathbf{E} = w_1, w_2, \ldots, w_N$ be a series of English words. We now consider the probability that the translation of word w_i appears in chunk $\boldsymbol{j}_j$ in the given sentence pair using the empirical translation probabilities $Pc(j|e)$. From the assumption that the translation of word w_i always appears somewhere in the Japanese sentence,

$$(3)\ \Sigma_j \Sigma_t P(t|w_i)\ P(t \text{ appears in } \boldsymbol{j}_j) = 1$$

, where t is the translation candidate of w_i, $P(t|w_i)$ is the probability that w_i is translated to t in the given sentence pair, and $P(t \text{ appears in } \boldsymbol{j}_j)$ is the probability that t appears in $\boldsymbol{j}_j$. Since the sentence pair is given and fixed here, $P(t \text{ appears in } \boldsymbol{j}_j)$ is zero if $\boldsymbol{j}_j$ doesn't contain t as a substring and one if it does. Precisely speaking, there is a possibility that t appears not as a translation of w_i even if $\boldsymbol{j}_j$ contains t as a substring, but we define $P(t \text{ appears in } \boldsymbol{j}_j)$ as stated above. We also make an assumption that the translation probability $P(t|w_i)$ in the given sentence pair is proportional to the empirical translation probability defined in (2). That is,

$$(4)\ P(t|w_i) = \lambda Pc(t|w_i)$$

for some constant λ. From (3) and (4), the probability that the translation of word w_i appears in chunk $\boldsymbol{j}_j$ is given as follows.

$$(5)\ P(\boldsymbol{j}_j|w_i) = \Sigma_t P(t|w_i) P(t \text{ appears in } \boldsymbol{j}_j)$$
$$= \Sigma_t P(t \text{ appears in } \boldsymbol{j}_j) * Pc(t|w_i) / \Sigma_j C_{ij}$$
$$= C_{ij} / \Sigma_j C_{ij}$$

, where

(6) $C_{ij} = \Sigma_t Pc(t|w_i) P(t \text{ appears in } \mathbf{j}_j)$

is called a *bilingual phrase matrix* which represents the relative likelihood that the translation of word w_i appears in chunk $\mathbf{j}_j$ in contrast to other Japanese chunks. Note that the values of C_{ij} can be calculated given the parallel sentence pair and the empirical translation probability.

Similarly for Japanese phrases, we can calculate the probability $P(w_i | \mathbf{j}_j)$ that the translation of $\mathbf{j}_j$ is represented as w_i as follows.

(7) $P(w_i | \mathbf{j}_j) = C_{ij} / \Sigma_i C_{ij}$

Next we consider the degree of uncertainty as to in which Japanese chunk the translation of w_i appears. For example, if $P(\mathbf{j}_j | w_i) = 1$ then it is certain that the translation of w_i appears in $\mathbf{j}_j$, that is, the entropy of the probability distribution $P(\cdot | w_i)$ is zero. The entropy $H(i)$ of the probability distribution $P(\cdot | w_i)$ in general is given as follows.

(8) $H(i) = -\Sigma_j P(\mathbf{j}_j | w_i) \log_2 P(\mathbf{j}_j | w_i)$

Since $\lim_{X \to 0} X \log_2 X = 0$, we define $H(i) = 0$ when $P(\mathbf{j}_j | w_i) = 0$ for all j.

In the proposed method, a statistical metric based on the entropy (8) is used for judging which adjacent phrases are to be merged. We calculate the change in the evaluation metric resulting from the merge just in the same way as we calculate the information gain (the reduction of entropy) of a decision tree when the dataset is divided according to some attribute, with the only difference that in a decision tree a dataset is incrementally divided, whereas in our method rows and columns are merged. We treat each row and each column of the bilingual phrase matrix as a dataset. The entire entropy, or uncertainty, of mapping English phrases to Japanese phrases is then given by:

(9) $H = \Sigma_i [\Sigma_j C_{ij}] H(i) / \Sigma_i \Sigma_j C_{ij}$

The entropy of mapping Japanese phrases to English phrases is obtained in the same way.

(10) $H_t = \Sigma_j [\Sigma_i C_{ij}] H(j) / \Sigma_i \Sigma_j C_{ij}$

Finally we define the total statistical metric, or evaluation score, as the mean value of the two.

$H_{tot} = (H + H_t)/2$

The merging process is terminated when the evaluation score H_{tot} takes a minimum value. When the final value of the bilingual phrase matrix is obtained, then for each non-zero element C_{ij} the corresponding English phrase in the i-th row and the Japanese phrase in the j-th column are extracted and paired as an aligned phrase pair. Whether rows are merged or columns are merged at each merging step is determined by the evaluation score. Since the merging process is easily trapped by the local minimum with a greedy search, a beam search is employed while keeping multiple candidates (instances of bilingual phrase matrices). The typical beam size employed is between 300 and 1000.

One of the advantages of the proposed method is that we can directly incorporate dictionary information into the scheme, which is quite effective for alleviating data sparseness problem especially in the case of small training corpus. Another distinctive feature of the method is that once word alignments are obtained and the empirical translation probability $Pc(j|e)$ is calculated together with the dictionary information, the word alignments are discarded. This is how this method avoids deterministic phrase alignment, and keeps a possibility of recovering from the word alignment errors.

2.2 With Syntactic Information

The proposed framework also has a capability of incorporating syntactic constraints and preferences in the process of merging. For example, suppose that there are two competing merging candidates; one is to merge (i-th row, i+1-th row) and the other is to merge (k-th column, k+1-th column). Then if there are no syntactic constraints or preferences, the merging candidate which has lower evaluation score is elected. But if there are syntactic constraints, the only merging candidate which satisfies the constraints is executed. When a syntactic preference is introduced, then the evaluation score is multiplied by some value which represents the degree of the strength of the preference. If we intend to extract only pairs of phrases which constitute a constituent, then we introduce a constraint which eliminates merging candidates that produce a phrase which crosses a constituent boundary.

Although our goal is to fully integrate complete set of CFG rules into the merging scheme, we are still in the process of constructing the syntactic rules, and in the present work we employed only a small set of preferences and constraints. Table 1 illustrates some of the syntactic constraints and preferences employed in the present work.

	Constraint	Preference
Japanese	・conjunctions and punctuations are merged with the preceding entities	・when the score ties, a merge which creates a constituent takes precedence
English	・conjunctions, prepositions and punctuations are merged with the following entities ・merging across base-phrase boundary is prohibited	・when the score ties, a merge which creates a constituent takes precedence. If the English preference conflicts with the Japanese precedence, the latter takes precedence.

Table 1: Syntactic constraints and preferences

3. Experiments on Parallel Patent Corpus

This section describes experiments with the proposed phrase alignment method on a parallel patent corpus.

We used the test collection of parallel patent corpus from the Patent Retrieval Task of the 3rd NTCIR Workshop (2002) for training word alignments. The corpus comprises of patent abstracts of Japan (1995-1999) and their English translation produced at Japan Patent Information Organization. We extracted 150 thousand sentence pairs from the PURPOSE part of the test collection of the year 1995. Each patent has its IPC category, from A through H. The description of the IPC categories is given in Table 2. In-house English and Japanese parsers are used to chunk sentences and to make a constituent judgment. We also used in-house bilingual dictionary with 860 thousand word entries.

For phrase alignment, we extracted 13,000 sentence pairs with English sentences of length smaller than 75 words, out of the sentence pairs in G-category of the above word alignment training set. The sentence length is constrained to reduce the computation load. Table 3 summarizes the training corpora used.

Category	A	B	C
Description	HUMAN NECESSITIES	PERFORMING OPERATIONS; TRANSPORTING	CHEMISTRY; METALLURGY
Category	D	E	F
Description	TEXTILES; PAPER	FIXED CONSTRUC-TIONS	MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING
Category	G	H	
Description	PHYSICS	ELECTRICITY	

Table 2: IPC Categories

Training	year	size(sent)	IPC CAT
Word Alignment	1995	150,000	A-H
Phrase Alignment	1995	13,000	G

Table 3: Training set description

Out of 13,000 sentence pairs 208 thousand unique phrase pairs are extracted. More than one set of phrase alignments can often be extracted from one pair of aligned sentences when the evaluation score reaches zero. Figure 3 shows examples of obtained phrase alignments. Japanese phrases acquired are mostly constituents, whereas many of English phrases are not, such as " by arranging", or "of infrared absorption ink". This is partly due to the fact that Japanese phrases are constructed out of base phrases, or chunks, whereas English phrases are constructed starting from individual words. Another reason is the fact that Japanese precedence rule takes precedence over English one as stated in Table 1.

The extracted phrase alignments were evaluated with an SMT engine. We used Pharaoh (Koehn, 2004) as the baseline. Although our goal is to use obtained phrase alignments as translation units of Rule-based/SMT hybrid systems, we haven't yet processed large amount of parallel corpora, and the decoding scheme which takes advantage of the constituent oriented phrase alignments is still under development. Therefore, instead of testing the phrase alignments as translation units, we tested the cross-domain portability of the obtained phrase alignments. One of the major merits of a syntactic constituent is its generalization capability. N-gram statistics extracted from large collection of data in a specific domain is a powerful resource within the same domain, but quite often fails to adopt to quite different domains. Constituents, or grammatical categories, on the other hand, cannot be tuned easily to a specific domain, but possess a generalization capability. In this experiment we trained Pharaoh using parallel sentences in one domain, namely IPC-G category, and tested the decoder in different domains. The training corpus we used is the 13,000 sentence pairs in IPC-G category listed in Table 3 for as a baseline setting.

We also used a set of aligned phrases extracted form the 13,000 sentence pairs for training Pharaoh (PhrAlign). The phrases are used alone and not mixed with the original parallel sentences. For testing, a set of 500 sentence pairs are randomly extracted from each IPC category of the year 1996. For development another set of 500 sentence pairs are extracted from the IPC-G category of the year 1996. Table 4 shows the result. PhrAlign outperforms Baseline in all the categories. Especially in category E, PhrAlign scores 1.49 points higher than Baseline, which is relative percentage of 16% increase from Baseline.

Since the training corpus is fairly small it is possible that the difference of the two cases decreases as the training data is increased, but this result suggests a generalizing capability of the syntactically oriented phrase alignments.

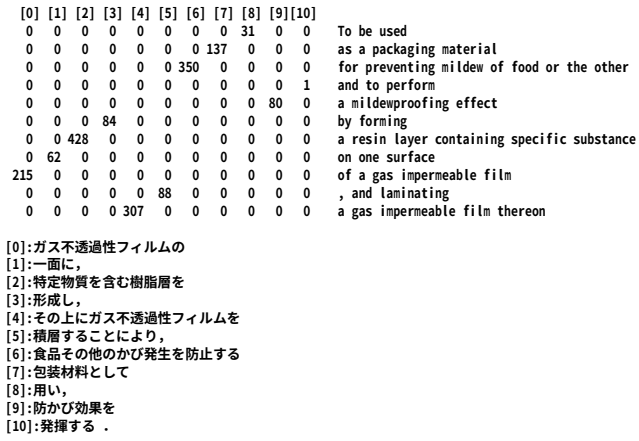
4. Conclusion

A novel approach is presented for extracting syntactically motivated phrase alignments. In this method we can incorporate conventional resources such as dictionaries and grammar rules into a statistical optimization framework for phrase alignment. The method extracts bilingual phrases by incrementally merging adjacent words or phrases on both source and target language sides in accordance with a global statistical metric along with constraints and preferences composed by combining statistical information, dictionary information, and also grammatical rules. Phrase alignments are extracted from parallel patent corpus using the method. The extracted phrases used as training corpus for a phrase-based SMT shows better cross-domain portability over conventional SMT framework.

References

Chiang, David (2005). A hierarchical phrase-based model for statistical machine translation. In Proceedings of the 43rd Annual Meeting of the ACL (pp263-270).

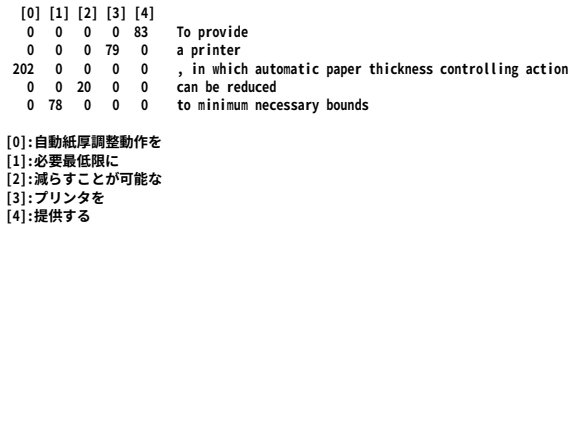
Koehn, Philipp, Franz Josef Och,, and Daniel Marcu (2003). Statistical Phrase-Based Translation. In Proceedings of HLT-NAACL.



```
[0] [1] [2] [3] [4] [5] [6] [7] [8] [9][10]
 0   0   0   0   0   0   0   0  31   0   0    To be used
 0   0   0   0   0   0   0 137   0   0   0    as a packaging material
 0   0   0   0   0   0 350   0   0   0   0    for preventing mildew of food or the other
 0   0   0   0   0   0   0   0   0   0   1    and to perform
 0   0   0   0   0   0   0   0   0  80   0    a mildewproofing effect
 0   0   0  84   0   0   0   0   0   0   0    by forming
 0   0 428   0   0   0   0   0   0   0   0    a resin layer containing specific substance
 0  62   0   0   0   0   0   0   0   0   0    on one surface
215   0   0   0   0   0   0   0   0   0   0    of a gas impermeable film
 0   0   0   0   0  88   0   0   0   0   0    , and laminating
 0   0   0   0 307   0   0   0   0   0   0    a gas impermeable film thereon

[0]:ガス不透過性フィルムの
[1]:一面に，
[2]:特定物質を含む樹脂層を
[3]:形成し，
[4]:その上にガス不透過性フィルムを
[5]:積層することにより，
[6]:食品その他のかび発生を防止する
[7]:包装材料として
[8]:用い，
[9]:防かび効果を
[10]:発揮する ．
```

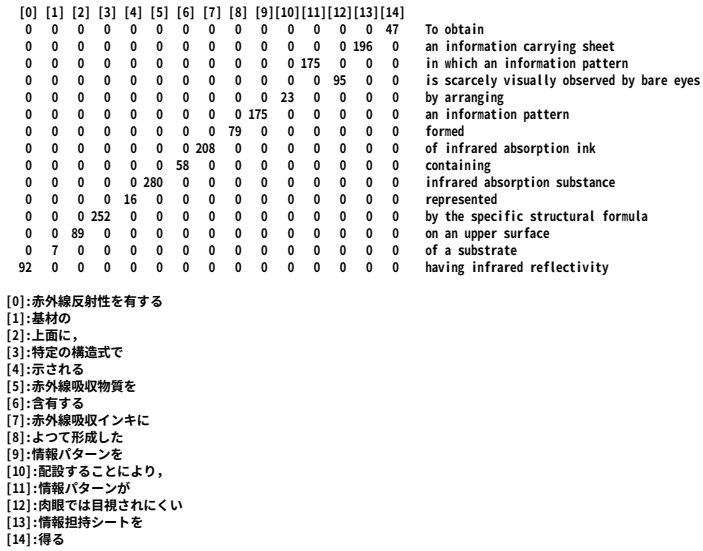
(a)



```
[0] [1] [2] [3] [4]
 0   0   0   0  83    To provide
 0   0   0  79   0    a printer
202   0   0   0   0    , in which automatic paper thickness controlling action
 0   0  20   0   0    can be reduced
 0  78   0   0   0    to minimum necessary bounds

[0]:自動紙厚調整動作を
[1]:必要最低限に
[2]:減らすことが可能な
[3]:プリンタを
[4]:提供する
```

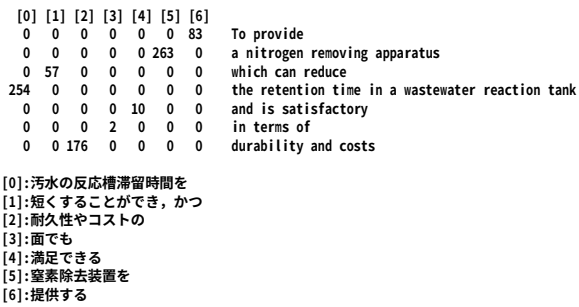
(b)



```
[0] [1] [2] [3] [4] [5] [6] [7] [8] [9][10][11][12][13][14]
 0   0   0   0   0   0   0   0   0   0   0   0   0   0  47    To obtain
 0   0   0   0   0   0   0   0   0   0   0   0   0 196   0    an information carrying sheet
 0   0   0   0   0   0   0   0   0   0   0 175   0   0   0    in which an information pattern
 0   0   0   0   0   0   0   0   0   0   0   0  95   0   0    is scarcely visually observed by bare eyes
 0   0   0   0   0   0   0   0   0   0  23   0   0   0   0    by arranging
 0   0   0   0   0   0   0   0   0 175   0   0   0   0   0    an information pattern
 0   0   0   0   0   0   0   0  79   0   0   0   0   0   0    formed
 0   0   0   0   0   0   0 208   0   0   0   0   0   0   0    of infrared absorption ink
 0   0   0   0   0   0  58   0   0   0   0   0   0   0   0    containing
 0   0   0   0   0 280   0   0   0   0   0   0   0   0   0    infrared absorption substance
 0   0   0   0  16   0   0   0   0   0   0   0   0   0   0    represented
 0   0   0 252   0   0   0   0   0   0   0   0   0   0   0    by the specific structural formula
 0   0  89   0   0   0   0   0   0   0   0   0   0   0   0    on an upper surface
 0   7   0   0   0   0   0   0   0   0   0   0   0   0   0    of a substrate
92   0   0   0   0   0   0   0   0   0   0   0   0   0   0    having infrared reflectivity

[0]:赤外線反射性を有する
[1]:基材の
[2]:上面に，
[3]:特定の構造式で
[4]:示される
[5]:赤外線吸収物質を
[6]:含有する
[7]:赤外線吸収インキに
[8]:よつて形成した
[9]:情報パターンを
[10]:配設することにより，
[11]:情報パターンが
[12]:肉眼では目視されにくい
[13]:情報担持シートを
[14]:得る
```

(c)



```
[0] [1] [2] [3] [4] [5] [6]
 0   0   0   0   0   0  83    To provide
 0   0   0   0   0 263   0    a nitrogen removing apparatus
 0  57   0   0   0   0   0    which can reduce
254   0   0   0   0   0   0    the retention time in a wastewater reaction tank
 0   0   0   0  10   0   0    and is satisfactory
 0   0   0   2   0   0   0    in terms of
 0   0 176   0   0   0   0    durability and costs

[0]:汚水の反応槽滞留時間を
[1]:短くすることができ，かつ
[2]:耐久性やコストの
[3]:面でも
[4]:満足できる
[5]:窒素除去装置を
[6]:提供する
```

(d)

Figure 3: Examples of obtained phrase alignments

IPC CAT	A	B	C	D	E	F	G	H
Baseline	7.94	11.43	10.24	7.42	9.29	11.38	14.66	12.03
PhrAlign	8.91	11.78	10.85	8.37	10.78	12.48	15.70	13.08

Table 4: Bleu score of baseline and the proposed method

Koehn, Philipp (2004). Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In 6th Conference of the Association for Machine Translation in the Americas, AMTA.

Marcu, Daniel and William Wong (2002). A Phrase-based Joint Probability Model for Statistical Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp.133-139).

NTCIR Workshop (2002). http://research.nii.ac.jp/ntcir/ntcir-ws3/work-en.html.

Och , Franz Josef and Hermann Ney (2000). Improved statistical alignment models. In Proceedings of the 38th Annual Meeting of the ACL (pp.440-447).

Och, Franz-Josef and Hermann Ney (2004). The alignment template approach to statistical machine translation. Computational Linguistics, 30(4), 417--450.

Yamada , Kenji and Kevin Knight (2001). A syntax-based statistical translation model. In Proceedings of the 39th Annual Meeting of the ACL (pp.523-530).

Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation

EHARA Terumasa*

* Department of Electronic Systems Engineering,
Tokyo University of Science, Suwa
5000-1, Toyohira, Chino-Shi, Nagano 391-0292, Japan
eharate@rs.suwa.tus.ac.jp

Abstract

Since sentences in patent texts are long, they are difficult to translate by a machine. Although statistical machine translation is one of the major streams of the field, long patent sentences are difficult to translate not using syntactic analysis. We propose the combination of a rule based method and a statistical method. It is a rule based machine translation (RMT) with a statistical based post editor (SPE). The evaluation by the NIST score shows RMT+SPE is more accurate than RMT only. Manual checks, however, show the outputs of RMT+SPE often have strange expressions in the target language. So we propose a new evaluation measure NMG (normalized mean grams). Although NMG is based on n-gram, it counts the number of words in the longest word sequence matches between the test sentence and the target language reference corpus. We use two reference corpora. One is the reference translation only the other is a large scaled target language corpus. In the former case, RMT+SPE wins in the later case, RMT wins.

1. Introduction

Sentences in patent texts are long. Figure 1 shows the frequency distribution of sentence length (characters) for Japanese patent text and Japanese newspaper text. The mean length of Japanese patent sentence[1] is 60 characters and of Japanese news sentence is 38 characters.

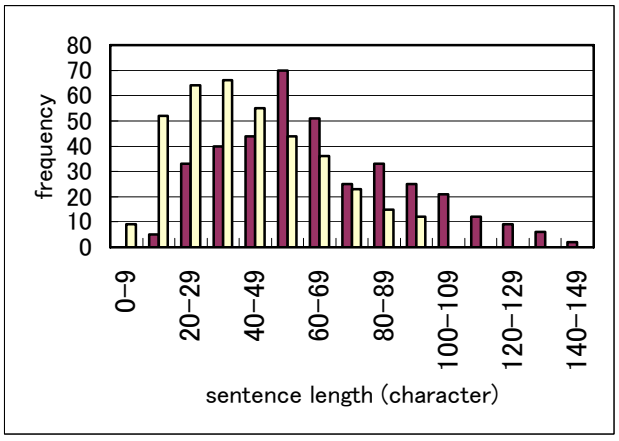


Figure 1: Frequency distribution of the sentence length of Japanese patent text and Japanese news text
dark bar: patent; light bar: news

Long sentences are difficult to translate by a machine, because these sentences often have complex syntactic structures. Although statistical machine translation is one of the major streams of the field, long patent sentences are difficult to translate not using syntactic analysis. Some papers show statistical machine translation gives high performance in translation word selection but it often gives syntactically strange outputs. So the combination of a rule based method and a statistical method was one candidate of high quality patent translation. Our system has a structure that combines a rule based machine translation (RMT) with a statistical based post editor (SPE).

There is some research about statistical post processing. (Langkilde and Knight, 1998) uses a statistical post processor in a language generation system. In this system, a symbolic language generator generates the word lattice and a statistical post processor extracts the most appropriate path from the lattice and outputs it. This post processing is controlled by n-gram based language model. (Senef et al, 2006) studies Chinese to English machine translation in the flight domain. They use a SPE system learned from artificially made parallel corpus composed of "bad" English and "good" English sentence pairs. Corpus size is 10,700 sentences. Sentence length is rather short. Mean sentence length of the corpus is 7.3 English words. Recently, (Simard et al, 2007) and (Dugast et al, 2007) used a similar strategy as ours. They are, however, concerning European languages.

In our patent translation case from Japanese to English, we have a parallel corpus. It is "Patent Abstract of Japan (PAJ)" corpus which is manually translated from the abstract part of "unexamined patent publication gazette (PPG)" of Japan[2]. An example of PPG and corresponding PAJ are shown in Appendix 1. So, we can collect "good" English as PAJ sentences and "bad" English as Japanese to English machine-translated results of original Japanese PPG sentences by the RMT.

2. System Architecture

Figure 2 shows the learning process of our statistical post editor. Translation model is learned from PAJ and machine translated results of PPG by RMT. We use GIZA++ as the translation model learner[3]. Language model is learned from PAJ using CMU-Cambridge's language model learner[4]. Figure 3 shows the translation process. Input Japanese patent sentences are translated by

[1] "Problem to be solved" part of "unexamined patent publication gazette of Japan".

[2] http://www.ipdl.inpit.go.jp/homepg_e.ipdl

[3] http://www.fjoch.com/GIZA++.html

[4] http://svr-www.eng.cam.ac.uk/%7Eprc14/toolkit.html

the RMT then they are fed to the SPE. We use the Isi-decoder[5] as the processor of the SPE.

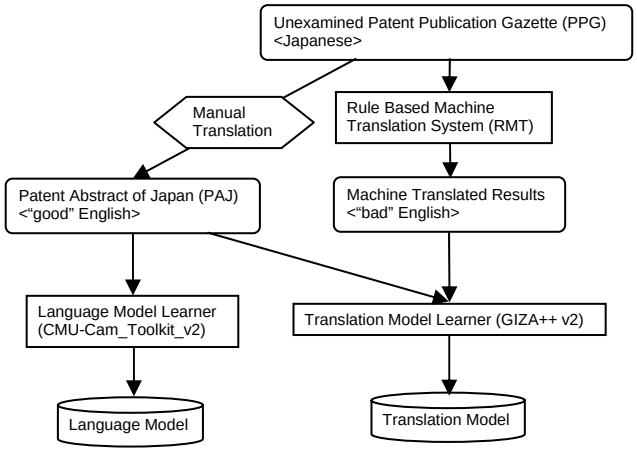


Figure 2: Learning process for the statistical post editor

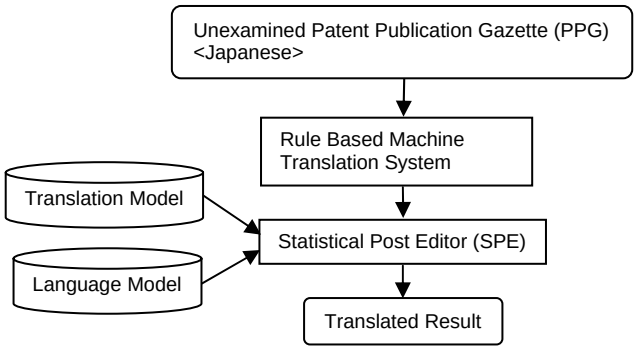


Figure 3: Translation process

3. Translation Experiments

3.1 Training Data and Test Data

We use Japanese and English parallel corpus of patent texts which are described in Chapter 2 as training and test data. They are "unexamined patent publication gazette (PPG)" of Japan as Japanese corpus and corresponding "patent abstract of Japan (PAJ)" as English corpus. We use 2003 year's data. We select only "problem to be solved" part from these corpora, because the first target of our research is to translate this part because it is less complex than the "solution" part.

First of all, we make text alignment between PPG and PAJ using the publication number. Next, we reject aligned texts which have a different number of sentences between PPG and PAJ. Since non-rejected aligned texts have the same number of sentences, we make sentence alignment between PPG and PAJ with the sentence number in the text.

Now, we call the PPG part of sentence aligned corpus as "src" (source sentence) and corresponding PAJ part as "ref" (reference translation). We also call rule based machine translation result of src as "rmt". From this ternary corpus, we make training and test data with the following process:

(1) When the numbers of words of sentences of either rmt or ref are over 90, the datum is rejected.
(2) When the ratio of the numbers of words in sentences of rmt and ref are less than 0.5 or more than 2.0, the datum is rejected.

Through above processes, we get a parallel corpus of src, rmt and ref. From 2003 year's PPG and PAJ original data which includes 337,026 text pairs, we can correct 316,570 sentence ternaries of src, rmt and ref. We use all of them to learn the language model and 92,855 ternaries to learn the translation model. We select 189 ternaries from the set of ternaries which is used for translation model as closed test data and another 189 ternaries from other than this set as open test data.

3.2 Translation Results and Preliminary Evaluation

Using the training data described above, the language model and translation model for SPE are learned. Then the translation system shown in Figure 3 is constructed. We call the output of RMT+SPE system "spe". We do closed and open test using the test data. We compare our results with base-line result that is the output of RMT only, that is "rmt" part of the ternary corpus. Some examples of test results are listed in Appendix 2. For the preliminary evaluation of the translation accuracy, we use the sentence level NIST score which needs reference translation(s). We use "ref" data as the reference. Therefore the number of reference is one. NIST scores are shown in Table 1.

Table 1: NIST scores
μ : mean; σ : standard deviation

test data	system	NIST	
		μ	σ
closed	rmt	4.274	1.329
	spe	5.198	1.769
open	rmt	4.423	1.262
	spe	4.871	1.498

The Kolmogorov-Smirnov test shows that all NIST scores belong to the normal distribution, with significant level 0.05. By the dependent t-test, spe provides significantly accurate translations than rmt, with significant level 0.01, both in closed and open test.

Manual check of the translation results by a human, however, reveals spe results often include syntactically strange expressions than rmt results. We guess that NIST is problematic to measure the translation accuracy, especially, the fluency as the target language. The BLEU case, (Callison-Burch et al., 2006) shows such problems.

3.3 A New Evaluation Measure NMG

To evaluate fluency measure, we need to use not only the small sized reference translation(s) but large sized target language corpus. We use US patent corpus as the target language corpus. Using this large sized reference corpus, we define a new evaluation measure of translation accuracy named NMG as follows.

(1) We consider that the test sentence C is constructed n words: $w_1, \cdots w_n$.

[5] http://www.isi.edu/publications/licensed-sw/rewrite-decoder/index.html

(2) For each w_i , we define $grams\ (w_i)$ as the maximum number of m that satisfies $w_i, \cdots w_{i+m} \in R$ where R is the set of all n-grams in the reference corpus.

(3) We define NMG score of C as

$$NMG(C) = \log_e(\sum_{i=1}^{n} grams(w_i)/n)$$

For example, if reference corpus includes the following four sentences:

i am a boy
you are a girl
he is a man
she is a woman

and when the test sentence C is

she is a girl

then, n=4 and

grams(she)=3
grams(is)=2
grams(a)=2
grams(girl)=1

Then

$$NMG(C) = \log_e((3+2+2+1)/4) = \log_e(2.00) = 0.69$$

3.4 Evaluation Using NMG

To evaluate RMT and RMT+SPE using NMG, we use two kinds of reference corpus. One is the same as the reference corpus which is used at the NIST score calculation. That is the corpus constructed by only one "ref" sentence which is in the PAJ. This reference corpus is named REF. The other is the corpus including 819,123 sentences extracted from the abstract part of the 157,596 US patent descriptions in the year 2000. This reference corpus is named ABS. We call NMG score using REF as NMG_REF and NMG score using ABS as NMG_ABS. When calculating NMG_ABS, we, however, ignore the following words as the stop words, because of reduction in the index file size.

the, a, of, ",", ".", and, to, is, in, an, for, with, by, which, from, at, on, be

We put the grams value of the above words as zero and, instead, we subtract the number of stop words from the word counts n.

The evaluation results using NMG are listed in Table 2.

Table 2: Evaluation Results using NMG
μ : mean; σ : standard deviation

test data	system	NMG_REF		NMG_ABS	
		μ	σ	μ	σ
closed	rmt	-0.1973	0.3802	0.7777	0.1798
	spe	0.1237	0.4839	0.7449	0.2005
open	rmt	-0.1463	0.3498	0.7795	0.1390
	spe	0.0533	0.3976	0.7159	0.1842

In NMG_REF case, spe wins rmt both in the closed and open test. In NMG_ABS case, rmt wins spe both in the closed and open test. These results suggest that spe has the advantage in "adequacy" and rmt has the advantage in "fluency". The Kolmogorov-Smirnov test shows that all NMG scores belong to the normal distribution, with significant level 0.05. By the dependent t-test, the differences between spe and rmt are significant with significant level 0.01 both in the closed and open test. Figure 4 shows the distribution of the difference of NGM_REF of spe and rmt in the open test. Figure 5 shows the distribution of the difference of NGM_ABS of spe and rmt in the open test.

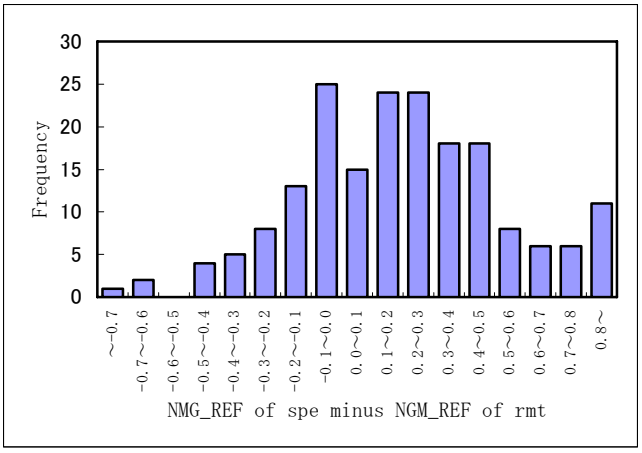


Figure 4: Distribution of the difference of NMG_REF

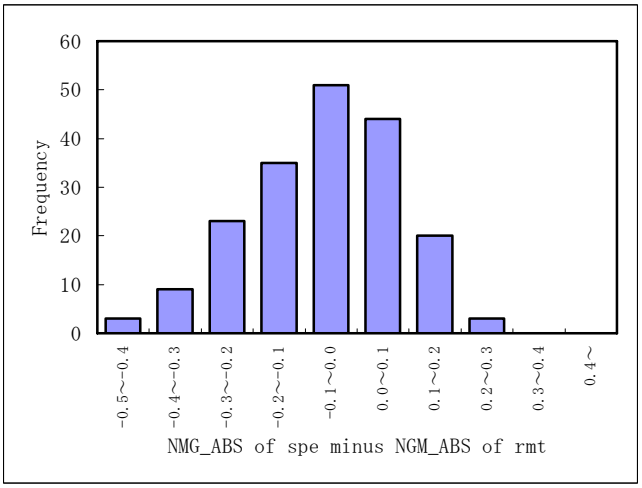


Figure 5: Distribution of the difference of NMG_ABS

3.4 Correlations between NIST and NMG_REF and between NMG_REF and NMG_ABS

Figure 6 shows the correlation between NIST score and NMG_REF score for the closed data. These data come from spe. Pearson's correlation coefficient between NIST and NMG_REF is 0.867. They are highly correlated.

Figure 7 shows the correlation between NMG_REF score and NMG_ABS score for closed test data of the spe system. Pearson's correlation coefficient between NMG_REF and NMG_ABS is 0.356. They are almost uncorrelated.

4. Related Works

Some researchers proposed translation accuracy evaluation measures using a large target language corpus (Callison-Burch & Flournoy, 2001; Akiba et al., 2002; Nomoto, 2003; Quirk, 2004; Corston-Oliver & Gamon, 2001; Kulesza & Shieber, 2004; Gamon et al., 2005). They use n-gram based perplexity type language models

and/or syntax/semantic based language models to evaluate translation accuracy. Syntax/semantic based model has the drawback that it needs lots of linguistic knowledge compared with n-gram based model. Our model is also based on n-gram, however, we do not use perplexity but the number of words of longest word sequence match. We do not find such an approach in previous works. (Miyashita et al., 2007) uses sentence match with the web corpus to evaluate fluency of the translation results, but it does not use word sequence match.

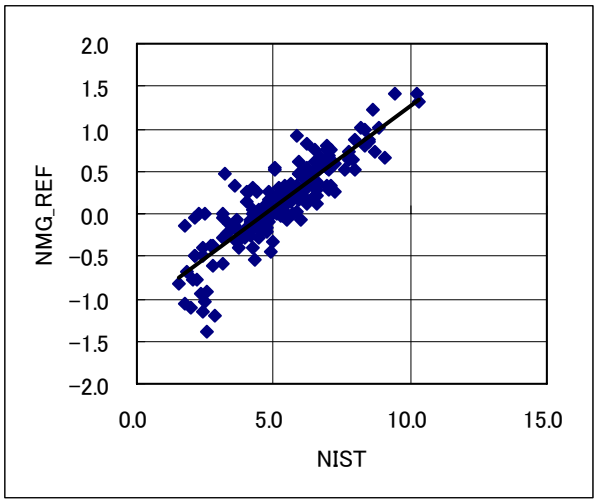


Figure 6: Correlation between NIST and NMG_REF

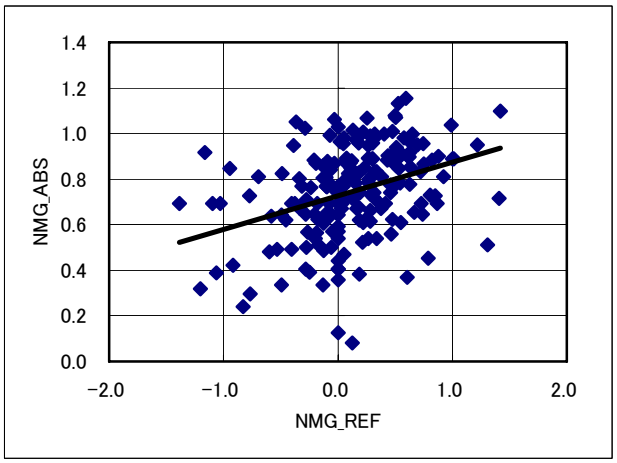


Figure 7:Correlation between NMG_REF and NMG_ABS

5. Conclusion

We proposed a rule based machine translation combined with statistical based post editing. In the evaluation process of our system, we proposed a new n-gram based measure NMG to evaluate translation accuracy. It uses word sequence match with reference translation(s) or large scaled target language corpus. From this evaluation result, we conclude the rule based part of the system has an advantage for structural transfer of a long and complex sentence, which is frequently seen in patent texts. On the other hand, the statistical part of the system has an advantage for lexical transfer of highly technical terms, which is also frequently seen in patent texts.

One of the future works is to compare NMG data to human evaluation results.

Acknowledgements

This work is done under the research in the AAMT/Japio Special Interest Group on Patent Translation. The author expresses sincere acknowledgements to the group members for their useful discussions. The corpora used in the work are provided by Japan Patent Information Organization (Japio). The author sincerely acknowledges Japio for their support.

References

Yasuhiro Akiba; Taro Watanabe and Eiichiro Sumita (2002): Using Language and Translation Models to Select the Best among Outputs from Multiple MT Systems, COLING2002.

Chris Callison-Burch and Raymond S. Flournoy (2001): A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines, MT Summit VIII, 2001.

Chris Callison-Burch; Miles Osborne and Philipp Koehn (2006): Re-evaluating the Role of BLEU in Machine Translation Research, EACL, 2006.

Simon Corston-Oliver; Michael Gamon and Chris Brockett (2001): A Machine Learning Approach to the Automatic Evaluation of Machine Translation, ACL2001.

Loïc Dugast; Jean Senellart and Philipp Koehn (2007): Statistical Post-Editing on SYSTRAN's Rule-Based Translation System, Proc. of the Second Workshop on Statistical Machine Translation, pp.220-223.

Michael Gamon; Anthony Aue and Martine Smets (2005): Sentence-level MT Evaluation without Reference Translations: Beyond language modeling, EAMT2005.

Alex Kulesza and Stuart M. Shieber (2004): A Learning Approach to Improving Sentence-Level MT Evaluation, TMI2004.

Irene Langkilde and Kevin Knight (1998): Generation that Exploits Corpus-Based Statistical Knowledge, ACL/COLING1998.

Kohei Miyashita; Seiichi Yamamoto; Keiji Yasuda and Masuzo Yanagida (2007): Quality Evaluation Method of Machine Translated Sentences by Comparing Text Retrieved from Web and Using Translation Model, Information Processing Society of Japan Special Interest Group Technical Reports, NL-177, pp.17-23, 2007 (in Japanese).

Tadashi Nomoto (2003): Predictive Models of Performance in Multi-Engine Machine Translation, MT Summit IX, 2003.

Christopher B. Quirk (2004): Training a Sentence-Level Machine Translation Confidence Measure, LREC2004.

Stephanie Seneff; Chao Wang and John Lee (2006): Combining Linguistic and Statistical Methods for Bi-directional English Chinese Translation in the Flight Domain, AMTA2006.

Michel Simard et al. (2007): Rule-based Translation with Statistical Phrase-based Post-editing, Proc. of the Second Workshop on Statistical Machine Translation, pp.203-206.

Appendix 1 Unexamined patent publication gazette and corresponding patent abstract of Japan[6]

公開特許公報フロントページ

(11)公開番号： 特開2000-253312
(43)公開日： 2000年09月14日

(51)Int.Cl.7 H04N 5/278
G09G 5/00 510
H04N 5/445

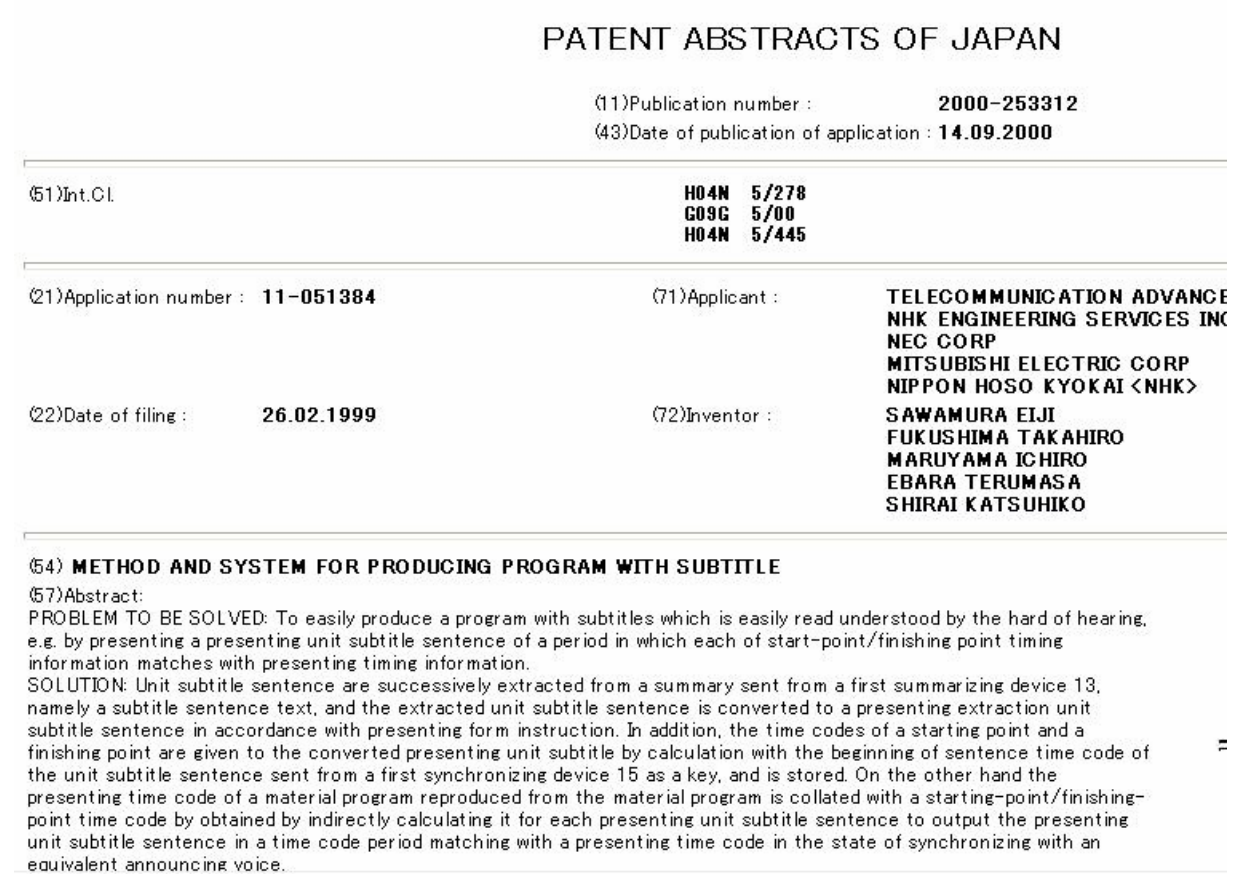
(21)出願番号： 特願平11-051384
(22)出願日： 1999年02月26日

(71)出願人： 通信・放送機構
財団法人エヌエイチケイエンジニアリン
日本電気株式会社
三菱電機株式会社
日本放送協会

(72)発明者： 沢村 英治
福島 孝博
丸山 一郎
江原 暉将
白井 克彦

(54) 字幕番組制作方法、及び字幕番組制作システム

(57)【要約】
【課題】例えば聴覚障害者にとって、読みやすくかつ理解しやすいことを考慮した種々の提示形式の字幕番組を容易に制作し得る字幕番組制作方法、及び字幕番組制作システムを提供することを課題とする。
【解決手段】字幕準備段階では、1又は2以上の単位字幕文が提示時間順に配列された字幕文テキストのなかから、提示対象となる1又は2以上の単位字幕文を提示時間順に順次抽出し、抽出された単位字幕文を、指示入力された字幕提示形式に従う提示単位字幕文に変換し、前記抽出された単位字幕文の文頭タイミング情報を参照して、前記提示形式変換後の提示単位字幕文毎に、始点／終点タイミング情報を付与して蓄積する一方、字幕提示段階では、提示単位字幕文毎に付与蓄積された始点／終点タイミング情報と、前記提示タイミング情報とを照合した照合結果に基づいて、始点／終点タイミング情報の各々が提示タイミング情報に合致する期間の提示単位字幕文を提示する。



PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-253312
(43)Date of publication of application : 14.09.2000

(51)Int.Cl. H04N 5/278
G09G 5/00
H04N 5/445

(21)Application number : 11-051384
(22)Date of filing : 26.02.1999

(71)Applicant : TELECOMMUNICATION ADVANCE
NHK ENGINEERING SERVICES INC
NEC CORP
MITSUBISHI ELECTRIC CORP
NIPPON HOSO KYOKAI <NHK>

(72)Inventor : SAWAMURA EIJI
FUKUSHIMA TAKAHIRO
MARUYAMA ICHIRO
EBARA TERUMASA
SHIRAI KATSUHIKO

(54) METHOD AND SYSTEM FOR PRODUCING PROGRAM WITH SUBTITLE

(57)Abstract:
PROBLEM TO BE SOLVED: To easily produce a program with subtitles which is easily read understood by the hard of hearing, e.g. by presenting a presenting unit subtitle sentence of a period in which each of start-point/finishing point timing information matches with presenting timing information.
SOLUTION: Unit subtitle sentence are successively extracted from a summary sent from a first summarizing device 13, namely a subtitle sentence text, and the extracted unit subtitle sentence is converted to a presenting extraction unit subtitle sentence in accordance with presenting form instruction. In addition, the time codes of a starting point and a finishing point are given to the converted presenting unit subtitle by calculation with the beginning of sentence time code of the unit subtitle sentence sent from a first synchronizing device 15 as a key, and is stored. On the other hand the presenting time code of a material program reproduced from the material program is collated with a starting-point/finishing-point time code by obtained by indirectly calculating it for each presenting unit subtitle sentence to output the presenting unit subtitle sentence in a time code period matching with a presenting time code in the state of synchronizing with an equivalent announcing voice.

[6] Quoted from Industrial Property Digital Library of Japan: http://www.ipdl.inpit.go.jp/homepg_e.ipdl

Appendix 2 Translation Examples and their NIST and NMG socres

closed test 2003000001
[src]
帯状土塊を破砕するとともに、培土作業を停止すると抱き込まれていた土塊が自動的に落下する構造の培土器を提供する
[ref]
to provide a ridger having a structure capable of crushing zonal clod and automatically dropping caught clod when ridging operation is stopped .
[rmt]
to provide structural Tstica that it wins to over one's side and the clod falls automatically when a zonal clod is crushed , and the Tstica work is stopped .
[spe]
to provide a ridger which is ridger to side and a clod falls down automati when a band-shaped screened is ignitionability and a drip-watering operation is stopped

	rmt	spe
NIST	3.4828	4.3577
NMG_REF	-0.4054	0.0000
NMG_ABS	0.3136	0.1251

open test 2003102102
[src]
電動機や制御回路が過負荷を受けるのを未然に防止するとともに、スムーズな乗り心地を実現させることのできる電気自動車の制御装置を提供する。
[ref]
to provide a control device for an electric vehicle , which prevents overload on an electric motor and a control circuit , and to realize smooth ride .
[rmt]
to prevent the electric motor and the controlling circuit from receiving the overload beforehand , and to provide the controller of the electric vehicle that can achieve smooth riding comfort .
[spe]
to prevent a electric motor and a control circuit from receiving an over load in advance , and to provide a transmission) of a motor-driven vehicle which can realize planarizing riding comfortableness

	rmt	spe
NIST	4.0614	4.8926
NMG_REF	-0.3429	0.2231
NMG_ABS	0.8708	0.7985

Patent documentation - comparison of two MT strategies

Lene Offersgaard Claus Povlsen

Center for Sprogteknologi
University of Copenhagen
Njalsgade 80, DK-2300 Copenhagen S
Denmark
{loff,claus}@cst.dk

Abstract

This paper focuses on two matters: A comparison of how two different MT strategies manage translating the text type of patent documentation and a survey of what is needed to transform a MT research prototype system to a translation application for patent texts. The two MT strategies is represented by PaTrans - a transfer and rule based system being used for more than 15 years by the translation agency Lingtech A/S and SpaTrans - a SMT system based on the Pharaoh framework. The SMT systems are characterised by shorter development time and low development cost compared to rulebased systems.

The distinctive text type of patents pose special demands for machine translation and these aspects are discussed based on linguistic observations with focus on the users point of view. Two main demands are automatic pre processing of the documents and implementation of a module which in a flexible and user-friendly manner offers the opportunity to extend the lexical coverage of the system. These demands and the comparison of the two MT strategies are discussed on the basis of proofread patents.

Introduction

Due to the characteristic features of patent documentation, this text type constitutes specific challenges for machine translation. This paper gives a brief description of patent documentation and how well two different MT systems are able to meet these challenges.

The first section gives an introductory description of the text type, patent documentation. Section two and three contain descriptions of the MT systems, a rule-based and an SMT-based system, respectively. The following sections introduce the evaluation procedure and report on the evaluation made on the two MT systems' translational results. The next section goes through the various error types that can be identified in the translational results. Some concluding remarks are given in the next section, summing up the observations that have been made with respect to comparing the translational results generated by the two MT systems. The final section outlines future plans on how to improve the translation quality of the SMT system.

Patent documentation – text typical feature

Since patent documents are official and juridical documents, they are kept in a departmental style meeting the following criteria:

- try to be as factual and impartial as possible
- let all information of given topic be expressed within one period.

The first criterion forms part of the reason why patent documentation texts have proven suitable for automatic translation. The demand of factual language usage promotes occurrences of many non-ambiguous technical terms. In addition, only the concrete and denotative meaning of words from the general word register are used. Even though patent texts are characterized by the absence of polysemiotic readings of the words used (facilitating the MT task), the whole idea or rationale behind writing a patent application makes certain demands that have to be met. The introduction of inventions lead per definition to coining of subject specific terms, designating the new concept in question. With respect to lexical coverage within the area of patent documentation the ratio of new terms will per definition be disproportionately high regardless of the size of the already system known terms.

In other words, an important design requirement of an MT-system tailored to patent documents is that it is capable of – one way or the other – treating system unknown words in flexible and user friendly way. Otherwise it would often result in poor translation results.

The second criterion entails the occurrence of very long sentences with many embedded subclauses and series of prepositional phrases. Again, in order to achieve high quality machine translation results the MT systems must be designed in such a way that treatment of very long sentences does not involve a profound decrease in translation quality. Another general feature embedded in the patent documentation text type is the frequent occurrences of entities such as references to other patents, dates, measure units and text internal references.

While the above mentioned characteristics cover patent documentation in general, other elements in domain subsets of patent documentation - related to the problem of system unknown terms - require specific treatment.

Focus in this context will be on the domain specific area of Chemistry. Not surprisingly this subset of patent documentation is dominated by the presence of many chemical formulae. The syntax of how chemical substances can be combined is well defined though they can be very complex, cf. the following examples:

$-CH_2CH_2N(R^{15})CH_2CH_2$
N-[3-[4-(6-fluor-1,2-benzisoxazol-3-yl)-1-piperidinyl]propyl]phthalimid
2-(3-(2-(ethoxy)ethylcarbonyloxy)propyl)ethyl

In some specific cases chemical formulae are language specific and need to be translated, but in general the formulae are language neutral and can be transferred/translated directly to the target language in question.

It is, however, crucial that MT systems in their design have encompassed a procedure for treating these non-verbal entities in order to obtain a reasonably high translational quality.

Comparison of two MT strategies

In the following a comparison of the use and performance of a statistical phrase-based MT system and a traditional rule-based MT system is made. The comparison focuses on linguistic aspects in the different kinds of error types which were identified in the output of the SMT system. First, we briefly describe the two systems to illustrate the very different nature of the two systems.

The PaTrans MT system

PaTrans is a rule based MT system designed for English-Danish translation of patent texts. PaTrans is a transfer based system directly descended from the Eurotra MT research prototype (EUROTRA, 1991). The transition from research prototype to a production MT system included extensions for optimisation, syntactic error recovery, grammatical coverage of patent document specific phenomena, integration of a part-of-speech tagger, document handling (with preservation of layout information), a rule based entity recogniser and implementation of an automatic post-editing tool (see Ørsnes et al, 1996; Povlsen & Bech, 2001 for a more detailed description).

In addition, in order to facilitate the manually conducted pre-editing task, various tools have been implemented, i.a. a term-coding tool and a tool that by making lookups in the existing term databases can identify system unknown words/terms in the source document.

The SpaTrans MT system

The SpaTrans system is developed in a research project financed by the Danish Research Council. The research concerned evaluation of the feasibility of developing SMT for the Danish language. The focus was on translation of patents from English to Danish. Two patent translation companies participated in the project acting the role of a potential future user and evaluated the potential of SpaTrans system.

The SpaTrans system is based on the phrasal SMT decoder Pharaoh (Koehn et al., 2003; Koehn 2004). The Pharaoh decoder is the translation engine and is placed in surroundings of pre and post processing components. The pre and post processing components are much simpler than the corresponding PaTrans components, but handle some of the same challenges, though they leave others unsolved for the time being. The possibility of using terminology databases and preservation of input document layout are not yet implemented in the SpaTrans system, while preservation of special characters, tokenisation and casing are handled. The SpaTrans system is based on a phrase table and a language model. The Pharaoh training software is used to train the phrase table. The training corpus consists of translated and sentence aligned patents. Experiments using europarl languages training material in combination with the patent texts lead to poorer results on development test set, so it was decided to do the training based only on patent texts at this stage. A similar observation is done by (Simard, 2007). The training corpus size can be seen in table 1. The training resulted in a phrase table with 2.3 mill phrases.

Corpus	English words	Danish words
Training	4.2 mill	4.5 mill
Language model	-	4.5 mill
Devel. test	19.464	17.465
Test	10.035	10.574

Table 1: Sizes for training and test corpus

The sentence length in the training material for Danish sentences is 25 words and for English sentences 28 words. The treatment of formulae and figures are not as elaborated in the SpaTrans system as in the PaTrans system, but regular expression substitutions are performed to solve the most widely used conversion problems between English and Danish figures and references.

The language model is trained (order 3) using srilm (Stolcke, 2002) based on the Danish part of the patent training corpus. Experiments based on human evaluation have shown that the use of the monotonic translation option is best suited for English-Danish translation. We are well aware that the quality of the translations by the SpaTrans system might improve if more training material could be added, but the issue here is mainly to investigate the potential in the use of SMT in Patent translations using domain text resources and to point out strengths and weaknesses. One important limitation of the SpaTrans system is that no terminology database is used. The input format of the decoder allows for applying information about predetermined translations of single words and multiword units. This facility can be used to apply specific terminology to the translation engine and before bringing the system in production use, this will be added to the pre processing module.

Evaluation

Analyses of the output of the two systems are based on BLEU metric (Papineni et al., 2002). There is much focus on evaluation of SMT and MT-systems and the used BLEU metric is only one simple way to measure quality. For a brief overview of other currently used evaluation metrics used for SMT and MT and recent experiences within the field, see Callison-Burch, 2007.

The BLEU metric gives one score for each test document. It has been argued that an increase/decrease in the value of the BLEU score does not guarantee a better/worse translation quality (Callison-Burch et al., 2006). But nevertheless the metric is widely used to measure development improvements in systems.

Given one or more reference translations the BLEU metric is normally used to score a text or a larger test corpus. The BLEU metric can also be used to calculate a score for each sentence in a test corpus, and these sentence based scores are in our evaluation used to focus on sentences with a low score, excluding very short sentences which by the definition of the algorithm will have a low score. Another aspect of the evaluation is the reference translation which is a product of post editing the PaTrans output by an experienced proofreader. This gives a large advantage to the PaTrans system, and this is reflected in the BLEU scores of the test material of the two systems, see table 2.

BLEU	Test patent A	Test patent B
PaTrans	**0.539**	**0.610**
SpaTrans reord.	0.439	0.399
SpaTrans mono.	**0.448**	**0.501**
Diff (PaTrans - SpaTrans mono.)	0.091	0.111

Table 2: BLEU scores for two test documents. Test patent A consists of 227 sentences and test patent B consists of 376 sentences.

Evaluation – one step further

About SpaTrans in general

A general observation concerning SMT systems is that the corpus used as training data per definition reflects the translation performance of the SMT system. As training data are collected within a specific text type about a domain specific subject, will in some cases involve that the SMT system suggests translations that are too narrow in their scope leading to poor evaluation results.

To give a translation example from the SpaTrans system[1]:

Further, these paints, properly formulated and applied, have the ability to remain effective for 5 years.

Has been translated into:
Yderligere, disse malinger, korrekt formuleret og påført, har evnen til at forblive der er effektiv i 5 år.

Literally translation:
Further, these paints, properly formulated and applied, have ability_the to to remain which is effective for 5 years.

The translation of 'effective' to 'der er effektiv' appears at first glance to be somewhat odd and it seems surprising that SpaTrans chose that translation. The Pharoah platform gives access to the word lattice generated during the translation process containing a list of the n-best translations that the system has considered. By adding an additional parameter in the command line, i.e. '–lattice' two files are generated. One that contains the word lattice and another that gives additional information about the states in the word lattice.

Opening the first file and looking up the n-best translations of 'effective', gives the following information:

```
(19638 (22478 "effektiv"              0.0117515))
(19638 (22485 "der er effektiv"       0.00482768))
(19638 (22472 "effektive ,"           0.000421076))
(19638 (22469 ", der effektivt"       0.00057635))
(19638 (22470 "er effektivt"          0.000124405))
(19638 (22471 "effektive med"         7.80866e-05))
-      -      -      -      -      -
-      -      -      -      -      -
```

The first number, 19638 refers to the particular state i.e. the token in the input sentence that is to be translated. The number 19638 contains the word coverage vector of 11111111111111100000, considering the transition probabilities between state 15 and 16 in the input sentence (i.e. between 'remain' and 'effective' in the source sentence). The number in the second column links to the translation of the next token in the input sentence.

As can be seen the best scores for translation of effective are the one in bold, i.e. *effektiv* with the score 0.0117515 and *der er effektiv* with the score 0.00482768. Seen from this isolated point of view it seems as if the model would select the translation 'effektiv' instead of 'der er effektiv'. If you, however, go through all the states and multiply all the probability transitions involved, it turns out that the best path (the least cost demanding path) is the one that has 'effective' translated as 'der er effektiv'. A quick look into the training data confirms that in patent documentation within the chemical subject domain, this translation will be the right one in most cases.

Reordering

Based on contrastive knowledge about English and Danish and various experiments conducted, it was decided that the parameter reordering value was set to the value of '-monotone', i.e. no reordering (see table 2). In terms of word order English and Danish are quite similar. One difference, however, can be seen in sentences in which adverbials (adverbs and prepositional phrases) have been topicalised, i.e. occurring in the first position of a sentence. While you in English preserve the SVO order, Danish swifts into a VSO order. In addition, since many adverbials in Danish cannot occur in position 1 of a sentence, you have to make a reorder to get the syntactically correct position translating from English into Danish. To give an example:

Indeed, marine antifouling paints based on organotin acrylate polymers have dominated the market for over 20 years.

The SpaTrans output:
Faktisk, marinbegroningshindrende malinger baseret på organotin- acrylatpolymerer har domineret markedet for over 20 år.

[1] It should be born in mind that the evaluation reference texts are the results of post-edited outputs from the PaTrans system which without any doubt in comparison with the Spatrans system favours PaTrans.

The post-edited version:
Begroningshæmmende skibsmalinger baseret på organotinacrylatpolymerer har faktisk domineret markedet i mere end 20 år.

Literally translated:
Marine antifouling paints based on organotin acrylate polymers have indeed dominated the market for over 20 years.

When sentences with topicalised adverbials are not reordered the resulting word order in Danish will be muddled-up. This is punished in the BLEU-evaluation and it contributes to the explanation of why the overall SpaTrans BLEU-scores are lower than the corresponding PaTrans BLEU scores.

Agreement

Another error pattern in the SpaTrans translation results is the frequent occurrence of agreement errors, such as in:

This constant erosion of the paint ...
Dette konstante erosion af maling ..

In Danish the noun, 'erosion' has the gender masculine and since the determiner 'dette' has the gender neuter, the translation has an agreement error. Seen from a BLEU score point of view these agreement errors are not crucial. Bearing in mind, however, that these errors are extremely frequent it helps explaining why PaTrans performs better than SpaTrans.

About PaTrans

Although the PaTrans system produces output of a quite high quality (reference to the BLEU-scores), some defective translation results are unavoidable especially in connection with automatic translation of patent documentation. The very high average sentence length requires the implementation of various robustness features ensuring that the system always produces a translation of an input sentence of whatever length. Whenever this failsoft component of the system takes over, it leads to activation of a more lean linguistic analysis of the input sentence which again leads to less precise translation results. The loss of information of morpho-syntax in these cases, for instance, results often in either mistaken or non-inflected word target translations.

Some concluding remarks

The core engines of SpaTrans and PaTrans perform approximately equal. If the problems concerning fronted adverbials and the agreement discrepancies mentioned above were solved then it would be likely that the two systems in terms of translation quality would perform approximately equal. This conclusion illustrates excellently the advantages of the SMT strategy. If you have access to parallel corpora of a high quality, it is possible to develop an SMT-system fast and at low cost that in terms of translation quality performs quite well.

As mentioned above, the step from the Eurotra research prototype to the PaTrans production system required both extensions and improvements of the system. These changes were made in order to tailor the system to process domain specific documents adhering to patent documentation text type.

In this context it would be relevant to call attention to two important PaTrans extensions. First a few comments about the implementation of the automatic entity recogniser. Patent documents contain per definition many entities of generic nature (chemical formulae, patent references etc.) which – seen from an SMT point of view - would require an almost infinite amount of training data to be included in the coverage of the system. Based on this assumption, it would be necessary to implement a preprocessor the functionality of which would be to identify these entities and mark them up so that the SMT system by handling these entities systematically can preserve its translation quality level.

Patent documentation contains per se new concepts and terms leading to a disproportionately high rate of system unknown words. In PaTrans these circumstances have been met by implementing an unknown word detection facility and in addition a user friendly term coding tool. Using SMT systems translating patent documentation would also require some kind of facility (e.g. a user dictionary) that would enable the user to extend the lexical coverage with system unknown terms before the translation process is activated. The need for a user dictionary has been recognized by the big SMT vendor Language Weaver since they have made it possible for the users to add existing term based and dictionaries to the phrase-tables. Language Weaver, however, point to the fact on their website, that in the world of statistical translation adding a user term base to the system could cause some disruption, since the language translation software is based on the probabilistic integrity of the phrase table. Language Weaver recommends alternatively that the user extends the SMT system coverage by including representative texts (containing the user coded terms) in the parallel corpus whenever an extension of the lexical coverage is needed.

SMT systems such as SpaTrans provide good translation results at low costs if good and many parallel data are available. Using SpaTrans in a commercial production context translating patent documentation would require that the functionality of the system is extended with an automatic entity recogniser and that the user of the system – one way or the other – is given the possibility of changing and extending the system lexical coverage in a flexible way.

Future work

In order to improve the BLEU-scores of SpaTrans the agreement problem reported above will be investigated. The experiments will follow two paths.

One will try to find out whether the general assumption – all other factors being equal – that more training data will improve the SMT outcome, as suggested in (Simard 2007). In this experiment both subject domain specific data and data from a general language corpus will be included, training both a general and a domain phrase table and combining these.

The other method will be to enrich the language model with additional linguistically based knowledge. This experiment will be conducted by tagging all the words in the corpus (which the language model is based on) with morpho-syntactic knowledge, by computing probability scores for sequences of these morpho-syntactic tags and finally by integrating these scores in the language model. This experiment will be made within the Moses framework[2]. At the workshop we will present the results from these two experiments.

Acknowledgements

The work reported here was partly financed by the Danish Research Council. We would like to thank Lingtech A/S and Plougmann & Vingtoft for providing us with training material and proofread patents. We would also like to thank the other participants in the SDMT-SMV project.

References

Callison-Burch, Chris and Fordyce, Cameron and Koehn, Philipp and Monz, Christof and Schroeder, Josh, "(Meta-) Evaluation of Machine Translation" in *Proceedings of the Second Workshop on Statistical Machine Translation*, June, 2007, Prague, Czech Republic, Association for Computational Linguistics, pp. 136-158.

EUROTRA (1991). Copeland, C., Durand, J., Krauwer, S. & Maegaard, B. (Eds.). S*tudies in Machine Translation and Natural Language Processing*, Vols. 1 and 2. Luxembourg: CEC.

Koehn, Philip Och, Franz and Marcu Daniel. Statistical phrase-based translation. In Procedings of the Human Language Technology and North American Association for Computational Linguistics Conference (pp. 127—133). Edmonton, Canada. Association for Computational Linguistics, 2003.

Koehn, Philip. Pharaoh: A beam-search decoder for phrase-based statistical machine translation models. In Procedings of AMTA, 2004.

Koehn, Philip, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello Bertoldi, Nicola Cowan, Brooke Shen, Wade, Moran, Christine, Zens, Richard Dyer, Chris Bojar, Ondrej Constantin, Alexandra and Herbst, Evan. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic, June 2007.

Maegaard, B. & Hansen, V. (1995). PaTrans: Machine Translation of Patent Texts, from Research to Practical Applications. In Engineering Proceedings of the Second Language Convention (pp.1—8).

Papineni, Kishore, Roukos, Salim, Ward, Todd and Zhu Wei-Jing. BLEU: A method for automatic evaluation of machine translation. In Procedings of ACL, 2002.

Povlsen, Claus, Bech A. Ape: Reducing the Monkey Business in Post-Editing by Automating the Task Intelligently. In Proceedings of MT Summit VIII (pp. 283—286). 2001, Santiago de Compostela, Spain.

Simard, Michel, Cyril Goutte & Pierre Isabelle. Statistical Phrase-based Post-editing. In Proceedings of NAACL HLT 2007 (pp. 508-515). Association for Computational Linguistics (ACL), 2007.

Stolcke, Andreas. SRILM - An Extensible Language Modeling Toolkit. In Proceedings of Intl. Conf. Spoken Language Processing, Denver, Colorado, September 2002.

Ørsnes, B., B. Music and B. Maegaard. PaTrans – A Patent Translation System. In Proceedings of COLING 1996 (pp. 1115-1118). Copenhagen. 1996.

[2] Moses is a open source drop-in replacement for the Pharaoh decoder. As a new facility Moses offers the possibility of using factored translation models. Factored translation models can be built based on surface forms, lemmas, part-of-speech and morphology.

Error Correcting System for Analysis of Japanese Patent Sentences

YOKOYAMA Shoichi and KENNENDAI* Shigehiro

Graduate School of Yamagata University (* now Dai-Nippon Printing Co.)
Jonan 4-3-16, Yonezawa, Yamagata 992-8510
JAPAN
yokoyama@yz.yamagata-u.ac.jp

Abstract

It is widely known that Japanese patent sentences, especially those regarding necessary conditions and details, have long and complicated structures. If these sentences are investigated by morphological analysis, most of the morphemes are correctly derived because each morpheme can be separated using the connective rules of parts of speech. However, the relation of modification is very difficult because the length of a sentence is large, and because the relationship is very complicated. Even humans researchers are often unable to extract the correct modification. The authors analyzed the automatic error correcting system for the modification analyzer. Initially, we extracted morphemes using the well-known standard morpheme analyzer "Chasen", and then extracted the modification relations for utilizing the standard software "Cabocha." The system automatically extracts the errors of Cabocha and indicates the corrections. We focused on the parallel phrases in Japanese, and estimated the result.

1. Introduction

It is widely known that Japanese patent sentences have long and complicated structures, with up to 200 Japanese characters (50 to 60 words), such that the modifications among phrases also have complicated and difficult structures. It is difficult for even native speakers to understand and clarify such structures.

If an individual wants to apply for a patent, they must retrieve the large-scale patent database in order to confirm whether or not there are similar patents. Correct and accurate retrieval requires automatic information extraction from the patent database.

Recently, the necessity of global application has increased due to rapid technological progress; thus, information should be shared immediately. A patent granted in one country should be valid in another country. If such system is realized, the request of machine translation for a patent will be increased. Therefore, the correct analysis of modification for patent sentences is necessary.

In this paper, we report a system that finds errors of automatic modification, and corrects these errors automatically. We describe the content of the system and the result of an evaluation (Kennendai, 2007).

2. Material and Background

The material is a DVD database in which all available patent gazettes of the Japanese patent office in 2003 are included (Patent, 2005). We have made a comparison of several Japanese patents and their English translations from the database. We previously reported that the modification errors in analyzing Japanese patent sentences reflect the translation result (Yokoyama, 2005). That is, if the modification is in error, the resulting translation also contains the erroneous modification.

If these errors are corrected, correct information about Japanese patent sentences can be obtained. The development of such a system will enable connection to a Japanese proofreading system.

2.1. Comparison of Modification between Japanese and English

The database stores the titles and abstracts of patents and their machine translations. We determined the existence of modification errors by comparing the machine translation data with the human translation data included in the patent database supported by the Japan Patent Office (Industrial Property).

2.2. Classification of Modification Errors

The content of a patent consists of bibliographical terms (publication number, date of publication of application, inventor, title of invention, etc), abstract and solution, range of the patent, detailed explanation, and a simple explanation of figures.

We previously classified the characteristic patterns of modifications occurring in patent sentences primarily written in the abstract and solution (Yokoyama, 2005). Based on this classification, we selected some patterns of modification errors. Analysis of modification is automatically performed by the "Chasen" modification software, which is commonly used by developing by the researchers at Nara Advanced Institute of Science and Technology (NAIST).

(a) Proper Representation in Patent Sentences

「本発明は～（中略）A である」 (This invention is A, which …) (A: noun)

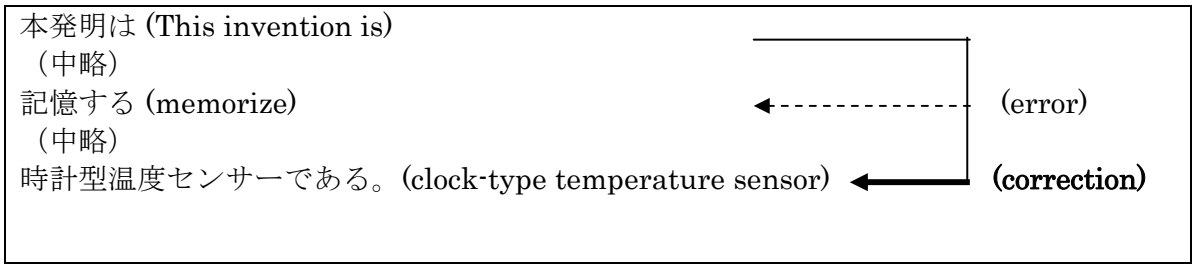


Fig. 1 An example of proper representation in patent sentences

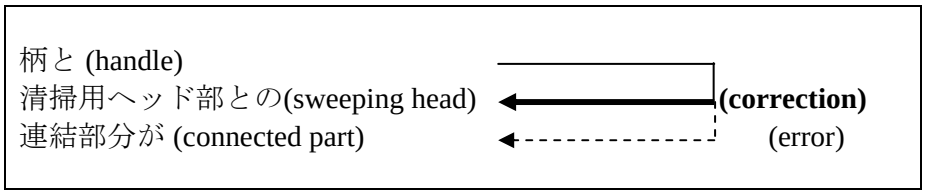


Fig. 2 An example of parallel structure in patent sentences

The above pattern is one of the most typical patterns in patent sentences. After the phrase "This invention is A, which," very long modifier(s) would follow. In Fig. 1, the subject "invention" is erroneously analyzed as the predicate modifying the verb "memorize". The correction should be made such that the subject should modify the last phrase "clock-type temperature sensor."

(b) Parallel Structure

「A と B との C が、」 (C of A and B is …) (A, B, and C are nouns.)

As shown in Fig. 2, the Japanese particle "to" ("and") is erroneously analyzed to modify the last noun. Correction is performed by the modification of the parallel property of nouns.

These corrections are usually made by human operators; however, we have developed a system which performs such corrections automatically. Other classifications are conjunctives, subject-verb agreement, modification between subordinate clauses, clause of noun modification, and parallel structure with noun and comma. These categories have not been implemented in the system because of the complexity of the procedure and/or algorithm.

3. System

The flowchart of the correction system, which automatically finds and corrects modification errors, is shown in Fig. 3. First, the patent sentence is input and analyzed by Cabocha. Using Cabocha, the system then finds erroneous candidates among the modifications, primarily through keyword and pattern matching. We also use a Japanese thesaurus (Ikehara, 1997); however, correction at this stage is not sufficiently effective because patent sentences often include many new and unknown words. If modification errors are found, they are then automatically corrected.

An example sentence is shown below. The correction of the sentence belongs to type (b) in the previous section, that is, the parallel structure followed by Japanese particle "to" ("and").

Example (partial) sentence

「製造設備、検査設備の各装置個別のデータ収集とデータ解析を下位のネットワーク上で可能とし、」(to make possible on the sub-network the collection of data and the analysis of data for each device in production facilities and inspection facilities)

0 1D 製造設備、 (production facilities)
1 2D 検査設備の (inspection facilities)
2 4D 各装置個別の (each device)
3 4D **<<3 7D>>** データ収集と (collection of data)
4 7D データ解析を (analysis of data)
5 6D 下位の (sub-)
6 7D ネットワーク上で (on … network)
7 8D 可能とし、 (to make possible)

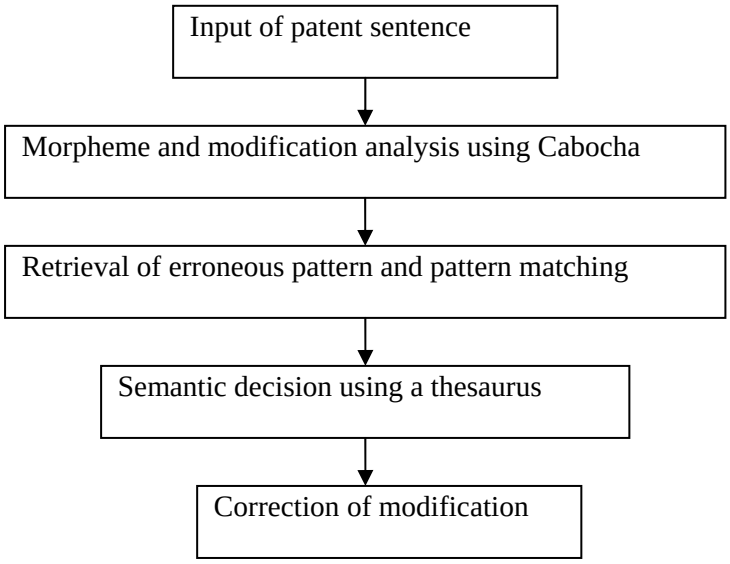


Fig.3 Flowchart of the system

The left-most number is the ordering number of the phrase, and the following number ("1D", "2D") is the modified phrase number. That is, "production facilities" correctly modifies "inspection facilities." However, phrase No. 3, "collection of data", erroneously modifies phrase No. 4, "analysis of data." Phrase No. 3 correctly modifies phrase No.7, "to make possible." This is the same pattern as shown in Fig. 2. The procedure for correction begins by finding the particle for parallel structure "to" ("and"). Next, the program retrieves the phrase "tono" (alternatively, "towo", "toni", or "to"). If such a structure is found, then the modification is corrected to the connection from "to" to "towo." In this case, the correction is successfully performed.

4. Evaluation

First, the result of classification of patent sentences by human operators is shown in Table 1.

Table 1 Classification of sample patent sentences

	Total 1228
Proper representation	19
Parallel structure	209
Conjunctives	92
Parallel structure with noun + comma	23
Unclassified error	85
Correct (no errors)	800

These 1228 patents are random files extracted from the DVD database (Patent, 2005). As shown in Table 1, the system deals with 19 proper representations and the parts of parallel structure (34 of 209) for the sentence form "C of A and B". All sample sentences were found and correctly modified and no correct modifications were modified erroneously.

Most of the 175 parallel structures, with 34 exceptions, have structure such as "A and D which C (verb) D." The above correction methodology cannot be applied to such structures. Among 209 parallel structures, the system can only deal with the structure "C of A and B," and cannot correct the similar structure "C of A and (A' and B')", in which B has the embedded parallel structure(s).

5. Concluding Remarks

This paper describes a system for finding and correcting modification errors. However, the system is only a simple prototype for error correction, and should be extended to address other types of errors, as shown in Table 1.

There are similar parallel structures written in the column at the parallel structure with nouns + comma in Table 1. This type of phrase has a complicated parallel structure (e.g., "meats, eggs, vegetables, spinach, eggplant, carrot,…") which sometimes includes a parallel structure with different levels. It is often difficult to clarify such detailed structures. The means to resolve such

errors is the use of a thesaurus for semantic interpretation. However, the range and depth of retrieval using a thesaurus is problematic. If the retrieval is too deep, the correct modification is erroneously modified; but, if it is too shallow, the error cannot be corrected satisfactorily.

The use of commas varies for each writer, and decisions on the error or correctness of usage can be difficult even for human operators. We will continue to examine the patterns of such sentences in the future.

If we can classify, detect, and adjust the modification structure of these sentences automatically, we will be able to contribute the improvement of automatic patent translation quality by correcting the modification structure. In addition, the same method can be applied to other type of Japanese sentences with complicated structures as well as patent sentences.

Acknowledgements

The authors thank Mr. Tadaaki OSHIO of the Japan Patent Information Organization (JAPIO) for elucidating the methodology of search and extraction for Japanese and English language patent data. This paper is based on the results of a discussion at the Special Interest Group in AAMT/JAPIO Research Committee.

Bibliographical References

IKEHARA Satoru et al. (eds.): Thesaurus for Japanese Vocabulary (Nihongo Goi Taikei) (in Japanese), Iwanami Publishing Co. (1997).

Industrial Property National Library in Japanese: http://www.ipdl.ncipi.go.jp/ homepg.ipdl

Industrial Property National Library in English: http://www.ipdl.ncipi.go.jp/ homepg_e.ipdl

KENNENDAI Shigehiro and YOKOYAMA Shoichi: A System Correcting Modification Errors in Patent Sentences (in Japanese), Proceeding of the 69th Meeting of the Information Processing Society Japan (IPSJ) (2007) 6Q-3, pp.2-427-8.

Modification Analyzer: “Cabocha”, Nara Institute of Science and Technology.

Morpheme Analyzer: “Chasen”, Nara Institute of Science and Technology.

Patent Database for the Special Interest Group in AAMT/Japio Research Committee, Japio (2005).

YOKOYAMA Shoichi and KANEDA Yuya: Classification of Modified Relationships in Japanese Patent Sentences, Proceedings of Workshop on Patent Translation in the 10th Machine Translation Summit (2005) pp.16-20.

On Portability of Resources for a Quick Ramp up of Multilingual MT of Patent Claims

Svetlana Sheremetyeva

LanA Consulting ApS
Jacobys Allé 23
DK-1806 Copenhagen Denmark
lanaconsult@mail.dk

Abstract

We describe a feasibility study on reusing the components of the unilingual authoring application AutoPat in a full-scale multilingual MT system APTrans, and explore to which extent MT knowledge can be ported from one language to another in the patent domain. We illustrate our findings on the example of English, Danish and French languages.

Introduction

Patents are a rich source of information about technological knowledge and a valuable tool in technology development. It is the area, which shows an increasing interest in high quality multilingual machine translation systems. To develop such systems requires rich knowledge resources (lexicons, grammar rules, world models), which nowadays must normally be painstakingly handcrafted from scratch for every language pair.

The idea to reduce development and maintenance costs, by sharing and reusing processing methods and knowledge has been in focus of researchers' attention for many years. For example, (Takeda, 1994) proposes portable knowledge sources for machine translation that consists of preference information on word sense, phrasal attachment, and word selection for translation. The basic idea of (Paul, 2001) is to devote efforts to the development of translation engines between the main linguistically different languages and to reuse the translation knowledge of these systems for translation into languages closely related to the target language. (Pinkham et al., 2001) describe the assembly of the French-English research MT system, which was constructed from a combination of pre-existing rule-based components and automatically created components.

A patent specific research in MT where the problem of portability is addressed by suggesting the constraint domain approach has been done for Russian to English by (Sheremetyeva and Nirenburg, 1999). Among the most recent attempts to reduce development cost by reusing pre-existing application components is a Japanese-English authoring patent system, which merges the English claim authoring system AutoPat (Sheremetyeva, 2003) and the Japanese machine translation application PC-Transfer (Neumann, 2005).

In this paper we present the results of further work on reusability of the AutoPat application. We describe a feasibility study on reusing the components of the existing unilingual authoring application in a full-scale multilingual MT system APTrans, and explore to which extent linguistic MT knowledge can be ported from one language to another in the patent domain. We illustrate our findings on the example of English, Danish and French in the frame of the APTrans architecture. Our discussion will mainly address the effort saving issues of augmenting the system with every new language-pair.

In what follows we shall first sketch the starting point of our research, the English patent claim authoring system AutoPat, we shall then describe the migration process from the unilingual AutoPat to the multilingual machine translation system APTrans followed by a worked out example for the three languages, - English, Danish and French. We shall also discuss other possibilities to use the APTrans architecture in machine translation.

AutoPat

AutoPat is a computer system for authoring patent claims in the English language. It consists of a technical knowledge elicitation module with an interactive user interface, lexicon, human input analysis module, content representation language, and generation module integrated with proofing tools (spelling, content and grammar checkers).

The knowledge base of the system includes a patent corpus-based English lexicon over a rich feature space, rules and knowledge representation language. AutoPat is a fully implemented product level application described in detail in (Sheremetyeva, 2003) and available at www.lanaconsult.com. We shall thus skip the AutoPat specification but rather concentrate on a re-engineering issue.

Development process: migration from unilingual authoring to multilingual MT.

Our goal is to find ways to speed up the development of a multilingual machine translation system, which can be specifically supported by domain constraints. Our multi-year R&D in the patent domain gave us a strong evidence of high lexical and structural similarity of patent claims in different languages. This inspired us to extrapolate "what is already there", - the knowledge base and program components of AutoPat, to another application, an MT system, and other languages.

Design

The first step in developing APTrans was to define a subset of the existing AutoPat components that will be the basis of the multilingual application and the extension it will need.

The modular architecture of AutoPat, which generates patent claims form content representations, suggested a transfer type MT architecture. All of the AutoPat components with the exception of the knowledge elicitation module can be reused for generation of the TL claim from the TL content representation. What is missing is an analyzing component, which could map raw claims into the AutoPat content representation format in a SL and a transfer module, which could convert a SL content representation into a TL content representation keeping the AutoPat format.

The knowledge base should be extended with multilingual MT lexicons, rules and heuristics. Other components that will definitely be needed are output post-editors for TLs.

To be a viable application that can be developed within a reasonable time a developer's environment for knowledge acquisition and maintenance should be an integral part of the application.

In our research the whole translation procedure was built "around" the existing AutoPat knowledge base and generator. We shall therefore first describe the reuse and customization of the lexicon, knowledge representation and generator and then show how the rest of the APTrans components were attached to them.

From the very start we programmed APTrans as a multilingual (not just bilingual) application, so that a new language can be easily integrated into the previously developed software.

Reuse and customization of existing components

Lexicon and feature space

The AutoPat lexicon (its vocabulary, entry format and feature space) is completely transferred to the APTrans application and used as a seed lexicon for lexical acquisition in other languages. We reused the approach to treat passive and active forms of verbs as different lexemes to simplify processing procedures.

Every entry following the English lexicon format is maximally defined as a tree of features:

SEM-CL [Language [POS [MORPH CASE_ROLE FILLER PATTERN],

where

SEM_Cl - semantic class; POS - part of speech; MORPH – morphological features, such as number, gender, etc., and domain relevant wordforms; CASE_ROLEs, - a set of lexeme case roles such as *agent, theme, place, instrument*, etc; FILLERs – lexical categories that can fill case-role slots of a lexeme; PATTERNs - linking features, that code both the knowledge about co-occurrences of lexemes with their case-roles and the knowledge about their linear order in the claim text.

Every node in the APTrans tree of features inherits its values from its ancestor. The mechanism of inheritance works in such a way that, in general, most values are inherited from the closest ancestor unless it is blocked or overwritten.

What is not trivial and probably only possible in such a restricted domain as ours is that there is a significant cross-linguistic parallelism (portability) in the values of two features, - CASE-ROLEs, and PATTERNs.

In other words, the set of case-roles for crosslingually equivalent predicates (verbs) and the order of their realization in the claim text are essentially invariant across languages. It means that in our tree of features there is not only a traditional "vertical" inheritance from parents to children, but for certain sibling nodes there is also a "horizontal" cross linguistic value inheritance which saves a lot of effort in non-English lexical acquisition.

Content representation language

AutoTrans reuses the AutoPat claim content representation language on both SL and TL sides of the translation process.

The format of the claim content representation as a set of predicate templates is given in Figure 1, where "label" is a unique identifier of the elementary predicate-argument structure, "predicate-class" is a label of a semantic class, "predicate" is a string corresponding to a predicate from the system lexicon, "case-roles" are "ranked" according to the frequency of their cooccurrence with a certain predicate in the training corpus, "status" is a semantic status of a case-role, such as *agent, theme, place, instrument*, etc., and "value" is a string which fills a case-role.

Sentence::={ template){template}*
template::={label predicate-class predicate ((case-role)(case-role))*}
case-role::= (rank status value)
value::= phrase{(phrase(word tag)*)}*

Figure 1. A claim content representation format.

Generation module

The AutoPat generation module, which takes a TL set of templates as input is what APTrans profits most of. It is fully reused from AutoPat for the English TL and, as our experiments show so far, requires only a slight updating for a non-English TL.

The whole concept of AutoPat generation, its rules and algorithms were originally worked out for Russian, and they actually code the legal requirements to the claim structure, which are essentially the same all over the world. This gave us the idea to port the generation knowledge to the English AutoPat, where it is now used without any essential changes.

We repeated our exercise in APTrans and ported the generation rules, this time, from English to Danish and French. For both languages only a few rules were updated, mainly to cover TL subject-predicate agreement.

In those cases where updating the English generation rules for Danish or French required too much effort we left them unchanged, thus "programming" mistakes in the translation output. We found it easier to correct these predictable mistakes at a later stage of processing, by running a TL posteditor on the generator output.

Analyzer

It was natural to think of the APTrans analyzer as the component to output its parse in the format of the content representation language.

Trying to reuse the knowledge we have already acquired for the English AutoPat we started with the analyzer for the English language and built it "on top" of the AutoPat disambiguating tagger. A bottom-up heuristic parser with a recursive pattern matching technique was then added to recursively chunk longer phrases preserving their inner structure. It also marks the head of every noun phrase and "learns" its "singular/plural" feature.

The last analyzing procedure determines the dependency relations between the chunks and predicates, and puts these chunks as fillers into case-role slots in predicate/argument structures, thus defining their semantic status (Sheremetyeva, 2003).

The reuse of the AutoPat generator has the advantage of simplifying the analysis task by making it possible to skip the problems of determining a) the syntactic relations between the predicate and its arguments within every individual predicate structure (*microsyntax*), and b) the syntactic hierarchy of predicate/argument structures in the input claim text (*macrosyntax).*

The generator, as was mentioned above, has the microsyntactic and macrosyntactic knowledge about the template hierarchy and the order of the phrases within predicate templates coded in its rules and lexicon.

To test the compatibility of the analyzer and the generator we modeled a "translation" experiment within one (English) language, thus avoiding (for now) lexical transfer problems. Raw English claims were input into the analyzer, and parsed. The parse was input into the generator. The modules proved to be compatible and the results of such "translation" showed a reasonably small number of failures, mainly due to the incompleteness of analysis rules.

We then tried to port the English analysis knowledge to the analyzers for Danish and French, the experiments show so far that a great deal of English analysis rules in our domain and approach can also be reused, though, of course, language specificity requires customization (e.g., location of adjectives in French noun phrases, lexical clues, etc.).

Transfer module

The APTrans transfer module takes the analyzer output, - a SL set of predicate templates as input and outputs a set of TL predicate templates whose slots are filled with presumably perfectly translated TL phrases/case-role fillers.

The APTrans transfer is in fact a combination of interlingual and syntactic transfer. The interlingual transfer finds TL equivalents[1] for every predicate and keeps the predicate template slot structure unchanged (invariant). The syntactic transfer is responsible for the translation of case-role strings.

A "real" translation procedure is thus reduced to the phrase level which, though not without problems, is still much simpler than machine translation of a full patent claim, especially when, which is often the case, it runs for a page or so.

[1] A base form of TL predicate from the lexicon substitutes a SL predicate gloss. The TL predicate gloss can be changed in the generator according to the generation rules.

Translation of phrases is done in two runs. First all lexical items in the SL case-role fillers are simply looked up in the lexicon and substituted by the base forms of their TL equivalents.

The second run applies syntactic transfer rules to the case-role strings. These rules are responsible for syntactic restructuring and agreement in TL language phrases. Besides the knowledge in the TL lexicon the rule condition part relies on the knowledge about the case-role, the type of phrase to which the lexeme belongs and the tag history. The tag history is the knowledge about the tag (e.g., part-of-speech) of the equivalent lexeme in the SL, which might be different from that in the SL.

The rules for phrase translation are of course language dependent, but here again a certain amount of portability is possible. We first tried our approach on the English/Danish pair, - the first pair of phrase translation rules was written for the English to Danish direction. These rules mostly coverer some Danish morphology phenomena [2] , and noun-article-adjective agreement in gender, definiteness and number.

In our experiments with the English to French translation we discovered that the left sides of agreement rules, which formulate the context for agreement, can in many cases be reused for the French language. The right sides of such rules, provided the reordering of adjectives is covered can to a certain extent be reused as well.

A worked example

Consider the following input claim text[3] in English to be translated in Danish and French:

A support for bearings comprising two connected half-shells provided with corresponding cavities adapted to form a seat for a bearing, characterized in that at least one of the cavities is shaped to form three radial raised portions for the contact of the bearing along corresponding imaginary lines parallel to the rotation axis of the bearing.

We illustrate this procedure on the example of translation a patent claim from English into French. The procedure for Danish is the same.

//A parsed output: English Predicate structures

Generic

(P1 Pgw "comprising"
1 Det1N2Prep3Np4 "A support for bearings "
2 Num5Pdc6Np7 "two connected half-shells ")

(P2p Pdw "provided"
1 Num5Pdc6Np7 "two connected half-shells "
2 Prepmn8Adjo9Np10 "with corresponding cavities")

(P3p Pdg "adapted"
1 Adjo9Np10 "open corresponding cavities "
3 Infm11Pgvi12Det1N13 "to form a seat ")

[2] For example, in the Danish language a definiteness of a noun is expressed morphologically: *a cup=en kop; the cup = koppen*

[3] For illustration we take a very short claim, normally the claim, still one sentence, can run for a page or so.

(P4 Pdgpr "for"
1 Det1N13 " a seat "
2 Det1N4 " a bearing ")

Difference

(P5p Pdvs "is shaped"
1 Qu14Qun15Detpl16Np10 "at least one of the cavities "
6 Infm11Pgvi12Num17Adjo18Pdo19Np20 "open to form three radial raised portions ")

(P6p Pdgpr "for"
1 Num17Adjo18Pdo19Np20 "three radial raised portions "
2 Detd16No21Prep22Detd16N4 "the contact of the bearing "
4 Prep23Adjo9Adjo24Np25 "along corresponding imaginary lines ")

(P7p Pdl "parallel"
1 Adjo9Adjo24Np25 "open corresponding imaginary lines
2 Prepmn11Detd16No26Prep22Detd16N4 "open to the rotation axis of the bearing ")

//French Predicate structures after BASE TRANSFER

Generic

(P1 W Pgw "comportant"
1Det1N35Prep17N6 "un soutien de roulement" 2Num41Pdc14Nfem19 "deux relié moitié-coquille ")

(P2p W Pdw "equipées"
1Num41Pdc14Nfem19 "deux relié moitié-coquille "
2 Prepmn42Adjo16Nfem7 "de correspondant cavité ")

(P3p G Pdg "adaptées"
1Adjo16Nfem7 "correspondant cavité close"
3 Prep39Pgv18Det2N34 " pour formant un siège")

(P4 G Pdg "pour"
1Det2N34 "un siège close"
2Det2N5 "un roulement")

Difference

(P5p V Pdv "formé"
1Qu4Qun23Detdm37Nfem7 "au moins un des le cavité close"
6 Prep39Pgv18Num38Adjo31Pdo32Nfem3 " pour formant trois radial augmentées partie ")

(P6p G Pdg "pour"
1Num38Adjo31Pdo32Nfem30 "trois radial augmentées partie
2Detdm36No15Prep22Detdm36N5 "le contact close de le roulement close"
4 Prep3Adjo16Adjo20Nfem21 " le long correspondant imaginaire ligne close")

(P7p L Pdl "parallèlles"
1Adjo16Adjo20Nfem21 "correspondant imaginaire ligne close"
2 Prepmn40Detdm36No33Prep22Detdm36N5 " à le axe de rotation de le roulement close")

//French Predicate structures after RULE TRANSFER

Generic

(P1 Pgw "comportant"
1 Det1N2Prep3Np4 "un soutien de roulements"
2 Num5Nfemp7Pdcp6 "deux moitié-coquilles reliées")

(P2p Pd "equipées"
1 Num5Nfemp7Pdcp6 "deux moitié-coquilles reliées"
2 Prepmn8Nfemp10Adjfmp9 "de cavités correspondantes")

(P3p Pdg "adaptées"
1 Nfemp10Adjfmp9 "cavités correspondantes"
3 Prep11Pgvi12Det1N13 "pour former un siège")

(P4 Pdg "pour"
1 Det1N13 "un siège"
2 Det1N4 "un roulement")

Difference

(P5p Pdv "formé"
1 Qu14Qunfm15Nfemp10 "au moins une des cavités"
6 Prep11Pgvi12Num17Nfemp20Pdo19Adjfmp18 "pour former trois parties augmentées radiales")

(P6p Pdg "pour"
1 Num17Nfemp20Pdo19Adjfmp18 "trois parties augmentées radiales"
2 Detdm16No21 "le contact"
4 Prep22Detdm16N4 "de le roulement")

(P7p Pdl "parallèlles"
1 Detdpl0Nfemp24Adjfmp23Adjfmp9 "des lignes imaginaires correspondantes"
2 Prepmn11Detdm16No25 "à le axe de rotation")

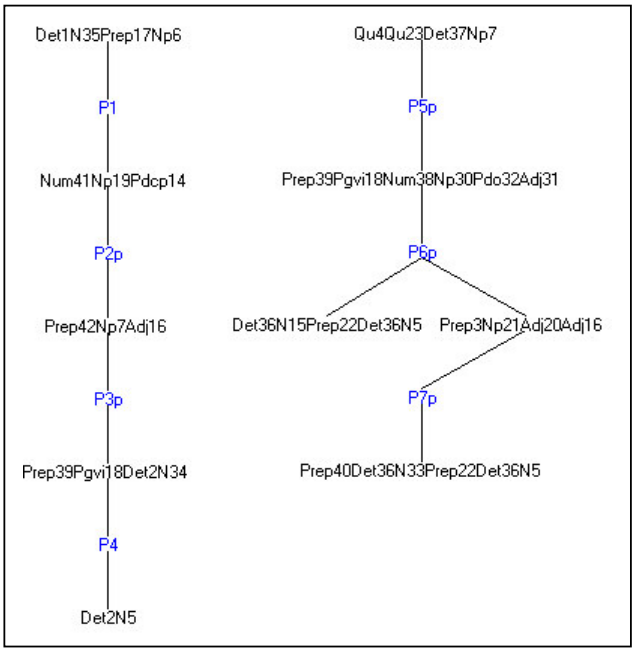


Figure 2. Trees built of the predicate templates by the generator.

French predicate structures after the RULE TRANSFER stage are input to the generator. All further operations are performed over strings of tags, which are substituted with the corresponding language phrases only after all the generation transformations are done. The input predicate templates are glued into trees following hard-coded language independent rules (See Figure 2). These trees following other set of generation rules, mainly universal, are linearised into a string of tags, which is further transformed to define the macrostructure and text cohesion of the TL French claims.

Figure 3 shows a screenshot of the professional user (e.g., translator) interface with the resulting APTrans translation from English into French and Danish. A trace of the postediting procedure is shown for every language.

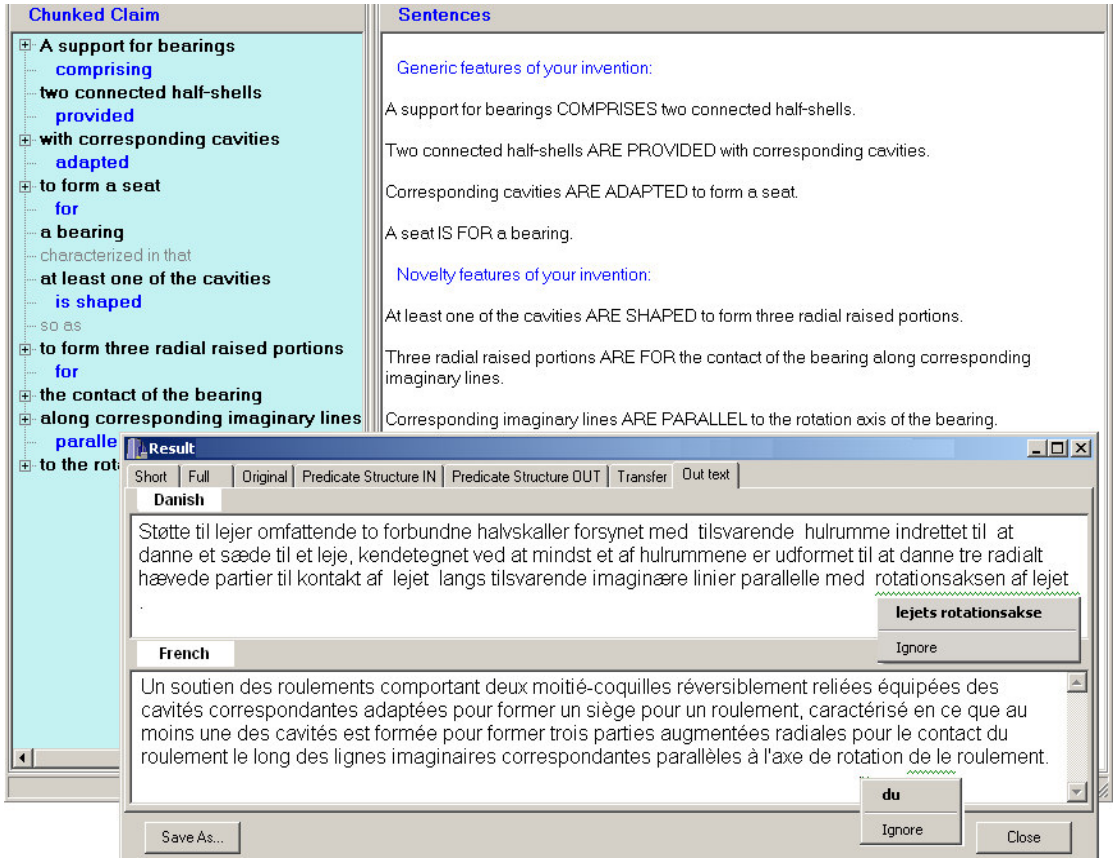


Figure 3. A screenshot of the APTrans user interface with an English claim translated in Danish and French.

The trace of the English claim analysis is shown to the user in the left pane of the background window.

The right pane of the background window shows simple sentences generated from the individual predicate templates. We kept this functionality from the original AutoPat generator for the user to check the correctness of the input claim analysis.

In case the simple sentences in the right pane are incorrect the user can interfere into the analysis procedure and through a special interface interactively correct the structure of the sentences thus correcting the analyzer output of predicate templates. This will result in a corrected translation.

Outsourcing MT

Reduction of the translation procedure to the machine translation on a phrase level opens another possibility for speeding up the multilingual translation development process: outsourcing phrase translation to a foreign MT system. We had a successful experience in trying this approach in a joint project on developing the Japanese-English patent authoring system[4], a patent claim generator in English from a Japanese-only interface. A Japanese user input the technical knowledge in his native language, which was further transformed by the system into a claim content representation in the AutoPat format with Japanese case-role fillers. The Japanese case-role fillers were separately translated from Japanese into English by the PC-Transfer MT system (see Neumann, 2005). The English strings were afterwards put back to the slots of predicate templates

[4] *The J-E patent system* ,Cross Language KK, Tokyo, Japan and LanA Consulting, Denmark, Copenhagen.

and input into the AutoPat Generator. As a result a full English translation of a Japanese claim was generated.

Performing MT by translating text segments smaller than sentence is getting into the focus of the MT research. (Bart et al., 2006) report on positive results achieved by reducing MT to a phrase level. In their experiment statistical techniques are used to decompose sentences into chunks, select the best translation of the chunks and recompose the translated chunks into a target language sentences.

Conclusions

In this paper we addressed the problem of saving on software development when building a family of NLP applications that share domain and task requirements. We illustrated the approach on the example of migrating from a system for authoring patent claims in English, AutoPat, to a multilingual machine translation system APTrans.

Though our research is a feasibility study we got a strong evidence that in the patent claim domain a noticeable economy of development effort could be achieved by porting linguistic machine translation knowledge from one language to another. We illustrated our findings on the example of English, Danish and French languages in the frame of the APTrans system architecture.

Due to the patent domain knowledge portability, as well as modularity of APTrans and the specificity of its components a foreign MT system can easily be integrated into the system architecture. This is a complementary way of speeding up the MT development.

We are planning to continue our research in both directions, - developing in-house machine translation resources and experimenting with foreign MT systems to integrate into APTrans those of them that show good results in their performance.

Bibliographical References

Bart. M., Mellebeek, K. Owczarzak, J.Van Genabith & A.Way. (2006). Multi-Engine Machine Translation by Recursive Sentence Decomposition. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, August 2006.

Neumann Ch. (2005). A Human-Aided Machine Translation System for Japanese-English Patent Translation. Proceedings of the Workshop on Patent Translation in Conjunction with MT Summit, Phuket, Thailand, September 16

Paul M. (2001): Translation Knowledge Recycling for Related Languages. *Proceedings of MT Summit VIII18-22* September. Santiago de Compostela, Galicia, Spain.

Pinkham J., M.Corston-Oliver, M. Smets & M.Pettenaro. (2001). Rapid assembly of a large-scale French-English MT system. *Proceedings of MT Summit VIII18-22*. September. Santiago de Compostela, Galicia, Spain.

Takeda K. (1994). Portable Knowledge Sources for Machine Translation. Proceedings of COLING 1994, 15th International Conference on Computational Linguistics, August 5-9. Kyoto, Japan.

Sheremetyteva, S. and S. Nirenburg. (1999). Interactive MT As Support For Non-Native Language Authoring. *Proceedings of the MT Summit VII. September 13-17, 1999, Singapore*.

Sheremetyteva S. (2003a). Towards Designing Natural Language Interfaces. Proceedings of the 4th International Conference "Computational Linguistics and Intelligent Text Processing" Mexico City, Mexico, February 16-22.

Sheremetyeva S. (2003b). Natural Language Analysis of Patent Claims. *Proceedings of the workshop "Patent Corpus Processing" in conjunction with 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003*), Sapporo. Japan, July 7-12.