A Human-Aided Machine Translation System for Japanese-English Patent Translation

Christoph NEUMANN
Cross Language KK
2-20-9 Nishiwaseda, Shinjuku
Tokyo, Japan, 169-0051
neumann@crosslanguage.co.jp

Abstract

The approach presented here enables Japanese users with no knowledge of English or legal English to generate patent claims in English from a Japanese-only interface. It exploits the highly determined structure of patent claims and merges Natural Language Generation (NLG) and Machine Translation (MT) techniques and resources as realized in the AutoPat and PC-Transfer applications. Due to its tuned MT engine, the approach can be seen as a human-aided machine-translation (HAMT) system circumventing major obstacles in full-scale Japanese-English MT. The approach is fully implemented on a large scale and will be commercially released in autumn 2005.

1 Background

The availability of patents for inventions in different languages is highly important for international trade and industry. Of up to 800,000 patent applications submitted worldwide every year, about 1/3 (250,000) are submitted to the Japanese Patent Office alone. However, only 60,000 of those patents are also submitted outside Japan (Ichikawa, 2001).

A major obstacle for submissions of Japanese patents abroad is, adding to multiple application fees, the high costs for translation and law experts handling the specific patent style. Regardless of whether the patent has already been filed in Japan or is filed abroad first, submission abroad typically involves the cooperation of at least three parties: the inventor, who has thorough knowledge about his invention, but typically only a passive knowledge of foreign languages; an international patent lawyer, who knows the legal requirements for patent submissions abroad, and a translator with the experience of translating technical texts. The threat of a *communication gap* between those parties makes numerous and constant consultations necessary, adding to time and costs for the patent submission.

To reduce those costs, however, few inventors or lawyers in Japan trust *MT* systems to be a viable alternative or support by replacing the translator. The language of patent claims is characterized by overwhelming sentence depth, length, and the abundance of technical terms – major causes for errors in parsing Japanese (Shinmori et al., 2003). Even patent-specific Japanese-English (J-E) MT systems tend to perform worse with patent translation than general-purpose systems with most other text styles.[1]

Our system largely bypasses those problems. The inventor or patent lawyer enters knowledge about the invention via a structured form and in Japanese. The system then translates the structured knowledge first into English, and finally generates a legally conform patent claim from the English information. The usual MT problems are minimized by limiting and tailoring MT to the translation of short phrases and words.

2 Related Research

Most of the computational linguistic research in the patent domain is devoted to information retrieval (Fujii and Ichikawa, 2002, Chen et al., 2003). Only few researchers focus on the morpho-syntactical specificities of patent claims (Shinmori et al., 2002). Shinmori et al. (2003) are also among the first to develop a parser explicitly dealing with the complex structure of Japanese patent claims.

The only patent-specific research in MT has been done for Russian to English by Sheremetyeva and Nirenburg (1999).

The research presented here is essentially based on two approaches: One is using controlled language input to enhance MT quality, e.g., by integrating grammar checkers (Nyberg et al., 2003) or limiting MT to lexicon mapping (Bjoern and Neumann, 1997).

The other technique integrated here is the elicitation of knowledge about the invention to allow for automatic generation of patent claims (Sheremetyeva and Nirenburg 1996, Sheremetyeva 1999, Sheremetyeva 2003).

[1] Currently, there are two specific patent MT systems for J-E: Toshiba's *The Hon'yaku* add-on package *Tokkyo-hon'yaku option V4.0*, and Cross Language's *Pat-Transer V7* , both extensions of conventional MT systems.

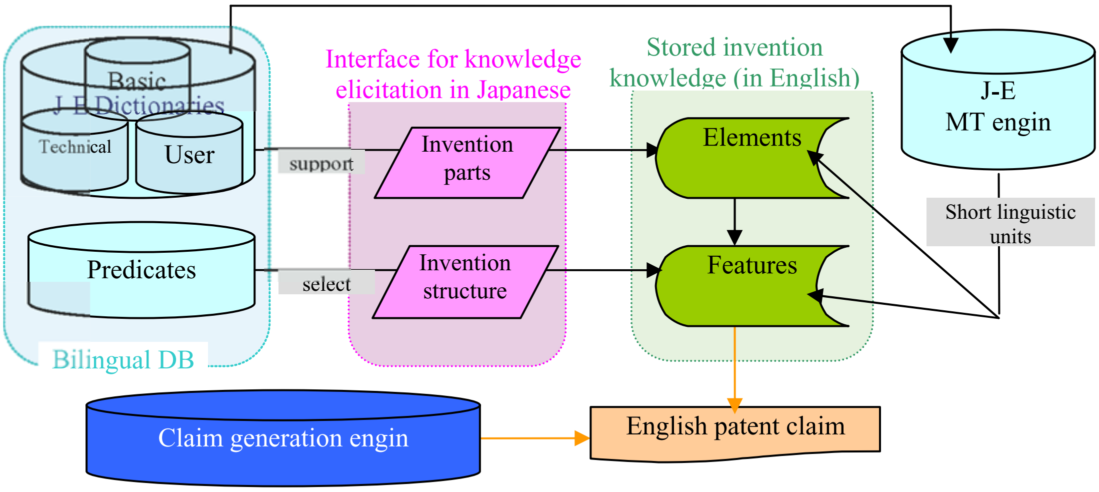


Figure 1. Schematic system architecture. The J-E MT engine is not involved in the actual generation of the English claim.

3 Overview of System

The architecture of this system is given in figure 1. The system elicitates knowledge about the invention via an interface, to which the user supplies words or smaller phrases in Japanese and structural information in form of a tree. A tailored MT engine translates those words and phrases into English. The claim generation engine builds the claim from the language-independent structure information and the set of English words and phrases.

This modularity allowed us to speed up the development process by reusing components from parent applications, namely the generation engine and the knowledge elicitation concept of *AutoPat*[2], a software for English users (Sheremetyeva 2003), and the MT engine and the dictionary of *PC-Transfer*, a J-E MT software [3].

3.1 Claim Generation Engine

One core component is an extended version of the English claim generation engine of AutoPat. AutoPat composes a patent claim ready for submission to a patent office by limiting input to the lexical and phrasal level. The specifics of the adaptation of the AutoPat engine for this system are described in Sheremetyeva (2005).

3.2 Bilingual Database

The bilingual database consists of a huge dictionary and of a lexicon of predicates. Both can be expanded by user-defined entries.

Dictionary. The dictionary is split into base and domain-specific technical dictionaries.

The *base dictionary* is a rewritten version of PC-Transfer's huge technical dictionary (2 million entries). The original entries contain much context information to allow for disambiguation in conventional MT. The tiny translation units here do not provide much context, but also do not require much disambiguation efforts. Context information was thus replaced by frequency ordering, allowing the user to choose manually less frequent meanings, if desired.

The *technical dictionaries* contain around 50,000 words each for three domains currently implemented (Machines, Electronics, IT) and were extracted semi-automatically from a large bilingual corpus of 40,000 patent texts using the technique developed by Takahashi et al. (2003). All entries here are nouns; verbs or adjectives extracted from the corpus were instead treated as candidates for predicates; see below (Other parts-of-speech are negligible in this technical language).

The **bilingual lexicon of predicate** covers verbs, adjectives and predicative nouns with a valency structure. Most of the currently 5,600 pre-defined predicates were extracted semi-automatically from a large patent corpus. Each predicate entry has a set of specific features relevant for the patent applications including:

The *Japanese predicate word* indicates the meaning of the predicate to the user, while the *English predicate word* serves as the key to the claim generation engine. The entry contains no inflection information at all.

[2] *AutoPat,* LanA Consulting, Kobenhavn, Denmark

[3] *PC-Transer Honyaku-Studio,* Cross Language KK, Tokyo, Japan.

Valency: a set of maximally 12 case-roles. The semantic status of every case-role is defined as "agent", "place", "mode", etc. thus specifying a case frame for every predicate. The set of possible case-roles is different for each predicate and is passed to the generation engine a neutral interlingua.

Syntactic features: sets of most probable postpositions and particles for case-roles to be filled with prepositional phrases (PP). A pre-defined table maps 53 postpositions and particles injectively from Japanese to English. Thus, prepositional phrases composed this way can be translated just by mapping the particles and then mapping the element name. Alternatively, the translation of non-PP fillers like adjectival phrases is supported by the MT engine.

3.3 Interface for Knowledge Elicitation and Representation

The interface design was largely inspired by the AutoPat interface. It consists mainly of separate input panels elicitating information and reference windows displaying elicitated information for each of four invention component categories, namely elements, external elements, features and steps.

Elements and external elements represent the physical parts involved in an invention. Syntactically, they are noun phrases. The user is encouraged to split up information about an element by inputting it in six designated fields like "number" or "purpose", which facilitates MT. While in the original AutoPat engine, a similar distinction into fields served mainly to help the user structure its thoughts, here, the specific linguistic properties of each field are explicitly used to facilitate MT.

Features and steps describe the static and dynamic structure of invention, what happens within the invention or how elements are related. They are created by filling case-slots of predicates (from the bilingual DB) with elements or free text. The user is encouraged to drag and drop elements with the mouse into case-slots instead of typing them, which allows the system to establish co-references between tokens. Then, he may choose the postposition or particle appropriately expressing the case from a pull down menu.

3.4 Tuned MT

PC-Transfer's MT engine is originally an all-purpose MT engine, i.e. it was designed to translate long texts of written Japanese consisting of full sentences. For our system, the MT engine was modified to deal appropriately with the small linguistic units supplied by the interface input fields. This allows avoiding a number of prominent error-prone situations in J-E MT.

Type-tuned translation. The morpho-syntactical types of an input string can be predicted from the type of its input field, such as noun phrases, simple lexemes, adverbial phrases or subordinate clauses. This way, parsing can be blocked beforehand from assigning false structures.

NP Coordination. Coordination of noun phrases is one of the major parsing problems of Japanese patent claims (Shinmori et al. 2003). Here, coordination is automatically detected and generated for sister elements in the tree structure, or, when the user assigns more than elements to a case-slot by drag-and-dropping them there from the element lists.

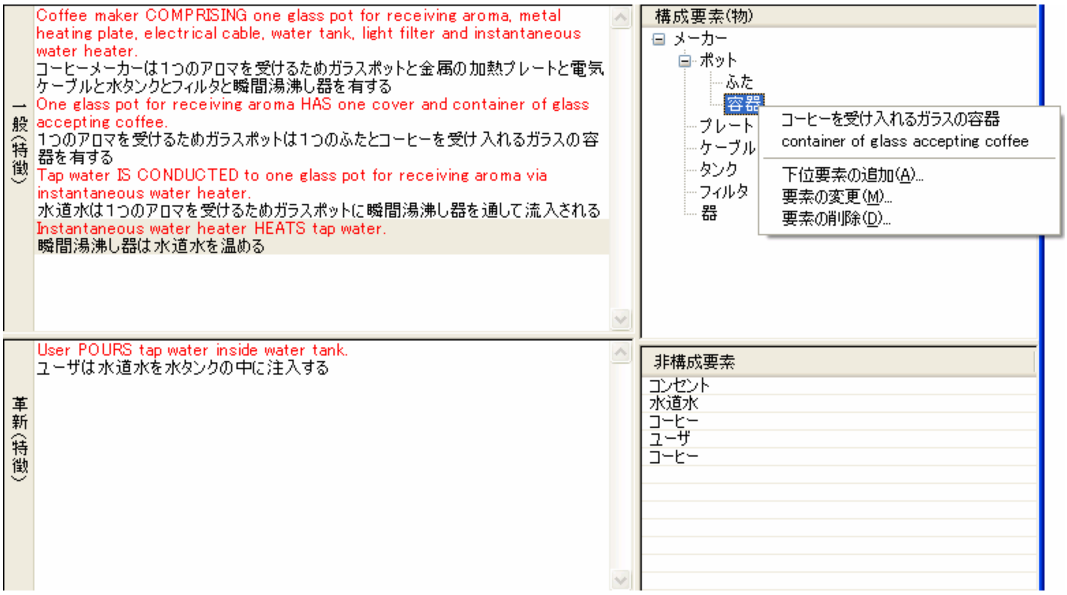
Manual referencing. Contrarily to other text forms, in patent claims, the same element must always be referred to by the same name throughout the text. Here, the user may use name variations in his Japanese draft; a special referencing window allows him to indicate co-reference information that will be passed on to the English engine.

No articles. Japanese nouns have no articles; a complex module in J-E MT engines must determine from context hints, whether the English equivalent noun should get *a, the* or zero (no article at all) (Bond 2005). This method gets even trickier with patents, as English patents include a "fourth article", the patent-specific *said.* In our system, article generation is calculated based entirely on co-reference information and position in the generated text and does not require Japanese context information. Consequently, the MT engine output is always set at "zero article".

No pronouns. Claims do not contain pronouns at all; as in Japanese, the surface realization of case-roles of the valency of a verb is not compulsory as in English, the MT engine normally fills empty case-roles in English with pronoun constructions. Consequently, the generation of pronouns is suppressed here entirely.

Attachment in noun phrases. In Japanese complex noun phrases, both attributes and dependent nouns precede the core noun. This makes it difficult to determine to which noun an attribute is attached. Here, the user indicates numbers and attributes in separate input fields so that they can easily be assigned to the core noun.

Manual lexical disambiguation. Lexical disambiguation from context information is switched off for most parts-of-speech. Compared to other text types, technical terms with clear, distinct translations are predominant in patent texts so that lexical ambiguities occur very rarely.



Coffee maker COMPRISING one glass pot for receiving aroma, metal heating plate, electrical cable, water tank, light filter and instantaneous water heater.
コーヒーメーカーは1つのアロマを受けるためガラスポットと金属の加熱プレートと電気ケーブルと水タンクとフィルタと瞬間湯沸し器を有する
One glass pot for receiving aroma HAS one cover and container of glass accepting coffee.
1つのアロマを受けるためガラスポットは1つのふたとコーヒーを受け入れるガラスの容器を有する
Tap water IS CONDUCTED to one glass pot for receiving aroma via instantaneous water heater.
水道水は1つのアロマを受けるためガラスポットに瞬間湯沸し器を通して流入される
Instantaneous water heater HEATS tap water.
瞬間湯沸し器は水道水を温める

User POURS tap water inside water tank.
ユーザは水道水を水タンクの中に注入する

Figure 2. Display of elicitated knowledge for describing a coffee machine.
On the left, templates are represented as simple sentences. On the right, the elements of the invention are represented in a hierarchical structure tree.

The user can check the English version of the invention knowledge, and can change the translation "globally" by overriding the default dictionary entry with a user dictionary entry.

Apart from these specific modifications made to the MT engine, the design of AutoPat's generation engine allows for bypassing a number of additional problems frequent in parsing of Japanese:

No complex sentences. The analysis of long sentences like patent claims with numerous propositions represented by intertwined subordinate clauses is highly error-prone (Shinmori et al. 2003) Here, the user has to create one separate feature for each proposition without having to care how those propositions interact. This is taken care of by the claim generation engine.

No relative clauses. Japanese patent claims abound of relative clauses. However, due to the absence of relative words in Japanese, the correct assignment of the case-role of antecedents within relative clauses is one of the most difficult tasks in J-E MT (Baldwin 1998). Here, the user enters propositions representing relative clauses as normal features; again, propositions are glued together only at the generation stage.

3.5 Workflow

The workflow is as follows:

Elicitation of knowledge. The system first elicits knowledge about the invention from the user by fill-in forms for invention elements and predicates. The user describes the knowledge either by looking at a Japanese claim or the invention itself.

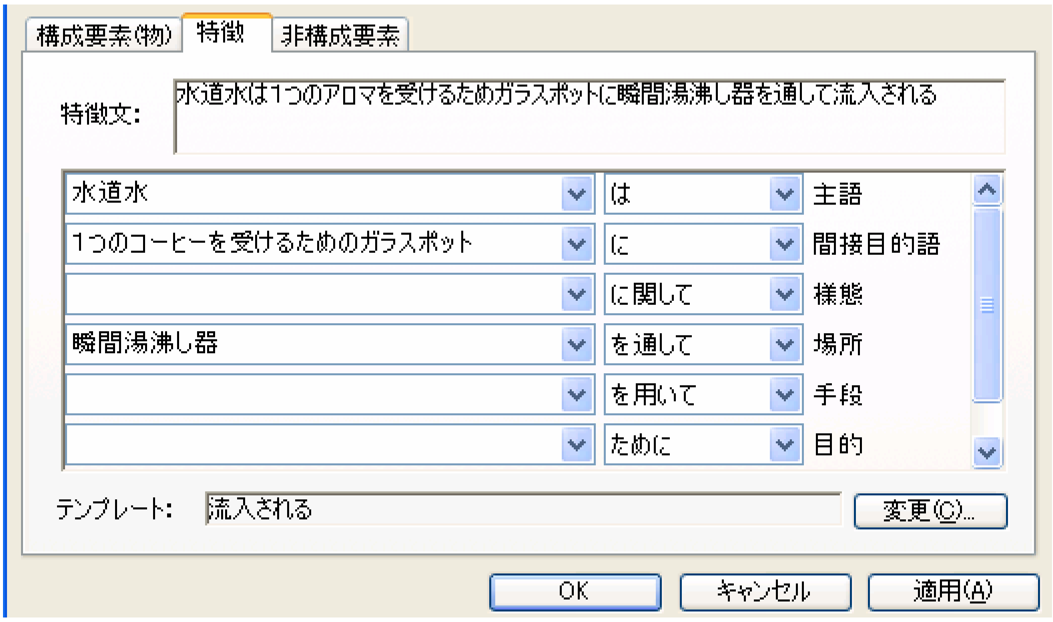
Input of elements. If the inventor of a coffee machine wants to express, for instance, that the coffee machine contains a heating plate, he chooses to add a "sub-element" to "coffee-machine". The six-fold input screen opening up right now might be filled like this (*field labels in italics;* [English in brackets])

(*core word* pureto [plate]) (*determining expression* kanetsu [heat]) (*number* ikko-no [one piece of]) (*purpose* --) (*attribute* kinzoku-no [metal-of]) (*other information* --)

The input is then passed on to the MT unit which first assigns an English translation to each field separately. Then, MT connects the English translations of all six fields to yield the English translation of the element's name. Parallely, a simple generation engine also produces the Japanese name for the element.

Display of elements (figure 2). Elements are displayed with their Japanese and English names in a hierarchical tree. External elements are displayed in a simple list. The user can change the structure of both, tree and list, by dragging elements.

Input of features. If the inventor wants, for instance, to express the fact that in the coffee machine, water is heated within the boiler unit, he first chooses the appropriate predicate *atatameru* ("to heat") from a list or looks it up in a semantic category, upon which the system provides him with a fill-in form containing the specific case-slot pattern for *atatameru* (figure 3).



特徴文: 水道水は1つのアロマを受けるためガラスポットに瞬間湯沸し器を通して流入される

水道水	は	主語
1つのコーヒーを受けるためのガラスポット	に	間接目的語
	に関して	様態
瞬間湯沸し器	を通して	場所
	を用いて	手段
	ために	目的

テンプレート: 流入される

Figure 3. Knowledge elicitation for predicate *chunyu-sareru* ("to be conducted"). Rightmost column indicates semantic cases; from middle column, user selects postpositions and particles typical for that case; invention elements are input or dragged into the fields on the left.

Now, the user writes or drags the related elements *shunkanyuwakashiki* ("boiling unit") into the agent field and *mizu* ("water") into the object field.

Content representation of features. Internally, the feature information is recorded using a simple knowledge representation language

feature:={ predicate ((case- role) (case-role)*)}
case-role::= (status value)

where *predicate* is a pointer to the predicate's lexicon entry, *status* is the interlingual name of the case-role, such as "agent", "theme", "place", "instrument" and *value* is the Japanese filler string, consisting of the Japanese particle name and the Japanese name of the element (or alternatively, free text, as for adverbial phrases).

The content representation reproduces the Japanese input ordered by the structured of the fill-in form. In the example in figure 3, Japanese input comes as (English translation below):

feature:= {*chunuyu-sareru* ((agent (*wa) suidosui*) (indirect-object (*ni*) *hitotsu-no kohi-wo ukeru-tame-no garasu-potto*) (place (*wo-toshite) shunkanyuwakashiki*) }

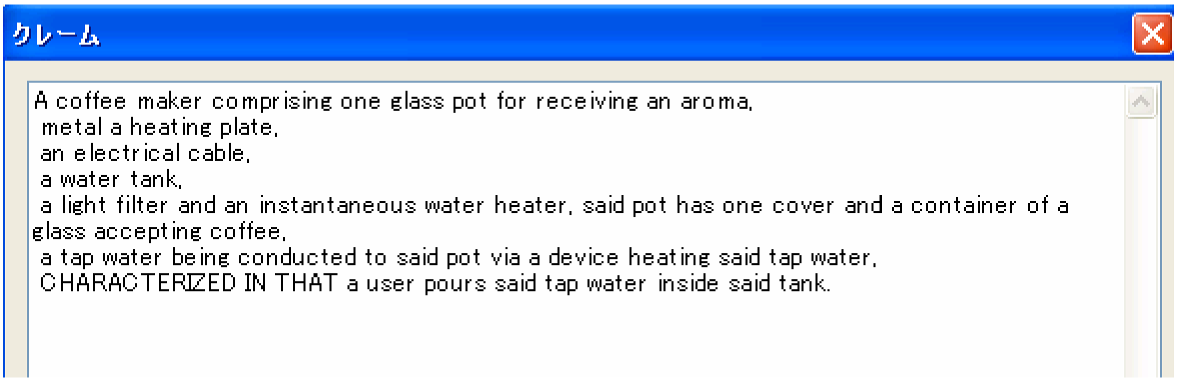
Display of Japanese feature sentence. All completed features are displayed as simple sentences in natural Japanese to the user. Those sentences are generated by the MT engine.

Generation of English feature sentences. The feature's predicate pointer is now replaced by the English keyword in a table-look up, while the fillers are either replaced by the English translations previously determined for elements, particles/postpositions or MT translates the free text input into English. The structure of the feature, including the neutral case-role names, however, stays unchanged.

feature:= {*conducted* ((agent *tap water*) (indirect-object *one glass pot for receiving coffee*) (place *instantaneous water heater*) }

This structure is now passed on to the generation engine which in return generates a simple English sentence for each feature. This sentence is displayed alongside with the Japanese equivalent (figure 2) to let the Japanese user verify that the list of features contains all intended information. Note that, while it is very difficult for most Japanese inventors to verify whether a long patent claim especially in English actually reflects the structure of their invention or not, here, invention knowledge is presented into more digestible short feature sentences in Japanese.

Generation of English claim text. From the standpoint of the Japanese interface, the main difference between English sentence generation and English claim generation is only the number of features passed to the generation engine and some co-reference information. The entire invention knowledge is represented as a set of features.



A coffee maker comprising one glass pot for receiving an aroma,
metal a heating plate,
an electrical cable,
a water tank,
a light filter and an instantaneous water heater, said pot has one cover and a container of a
glass accepting coffee,
a tap water being conducted to said pot via a device heating said tap water,
CHARACTERIZED IN THAT a user pours said tap water inside said tank.

Figure 4. Generated English claim sentence (European Patent Office format)

text::={ feature}{feature}*

For this input string, the generation engine finally returns a legal claim text in English (figure 4).

4 Evaluation

Accuracy was evaluated using a patent specific J-E MT system as the baseline. 20 single patent claims in Japanese from various technical domains were translated by the MT system. Parallelly, a trained human user created English patent claims by reading the Japanese patent claims and trying to recreate the content using this system.

The overall accuracy measure was based on a translation accuracy evaluation scheme distinguishing five accuracy levels. Our system reached 67% accuracy, while the MT system reached 50%.

Many errors may be contributed to the very early stage of the development, where many features where not yet implemented, for instance, indication of co-reference, and the MT engine was not tuned to phrase-level translations. We are optimistic that the next evaluation will have significantly improved results. Also, in the next evaluation we plan to evaluate the *speed* to create a single claim with our system as compared to a human translating from Japanese.

5 Implementation

The implementation of the Japanese user interface as well as the adaptation of both engines (patent generation and MT) to the system and their integration as well as the predicates, particles and the dictionary are mostly completed. The system is currently undergoing (summer 2005) beta-tests with professional users before it will be released commercially in autumn 2005.

Apart from the so-called apparatus claim type that is focused in this paper, the system also generates so-called method claims (used mainly in IT patents), which require a different user interface and a different design for the generation engine. However, the main features of the system remain unchanged.

6 Conclusion: Sayonara, MT (at least for J-E patents)?

This system generates English patent claims from a Japanese-only interface, merging an English patent generation engine with a Japanese-English MT software.

The typical user will be a Japanese inventor or patent lawyer with the invention and/or the Japanese patent claim in front of him. He “explains” the content of the Japanese patent claim to the system in Japanese, while the output is in English.

In this respect, the system is a Japanese-English HAMT system. J-E translation is done only on the phrase/simple sentence level, many well-known problems in J-E MT are outsourced either to the human (parsing the patent claim is replaced by the human “describing” it to the system) or to the generation engine, which handles the syntactically correct representation of many propositions packed into one complex English sentence.

This way of “avoiding” MT works well mainly due to the specific text type of claims: Most terms are technical and have only one precise translation, thus avoiding semantic ambiguity problems. The morpho-syntactical structure of claims is complex, but so predictable, that it can be represented beforehand in terms of a fill-in form. While this approach seems promising to the commercially important, but limited field of the patent claims, future research must show if it can be also be adapted to other text types with freeer, less predictable structures.

Acknowledgments

We thank the Information-technology Promotion Agency, Tokyo, Japan who partly funded this research (Grant 2004/374).

References

T. Baldwin. 1998. *The Analysis of Japanese Relative Clauses* (Master's Thesis), Tokyo Institute of Technology, Tokyo

M. Bjoern and C. Neumann. 1997. Multilingual Pre-Processed Communication on the World Wide Web. *Proceedings of the SAS Congress on Applied Computing.* San Jose

F. Bond. 2005. *Translating the Untranslatable. A Solution to the Problem of Generating English Determiners*, CSLI Publications, Stanford

L. Chen, N. Tokida, H. Adachi. 2003. A Patent Document Retrieval System Addressing both Semantic and Syntactic Properties. *Proceedings of the ACL Workshop on Patent Corpus Processing*. Sapporo, 1-6

A. Fujii and T. Ishikawa. 2002. NTCIR-3 Patent Retrieval Experiments at ULIS. *Proceedings of the Third NTRCIR Workshop*

M. Ichikawa. 2001. Towards a Global Patent System: The Japan Patent Office View. http://www.law.washington.edu/casrip/Symposium/Number7/4-Ichikawa.pdf (accessed last in july 2005)

E. Nyberg, T Mitamura, D. Svoboda, J. Ko, K. Baker, J. Micher. 2003. An Integrated system for Source language Checking, Analysis and Terminology management. *Proceedings of Machine Translation Summit IX, September.* New-Orleans

S. Sheremetyeva and S. Nirenburg. 1996. Knowledge Elicitation for Authoring Patent Claims. *Computer*, vol. 29, no. 7, 57-63

S. Sheremetyeva and S. Nirenburg. 1999. Interactive MT As Support For Non-Native Language Authoring. *Proceedings of the MT Summit VII.* Singapore

S. Sheremetyeva. 1999. A Flexible Approach To Multi-Lingual Knowledge Acquisition For NLG. 1999. *Proceedings of the 7th European Workshop on Natural Language Generation.* Toulouse

S. Sheremetyeva 2003. Towards Designing Natural Language Interfaces. *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing.* Mexico City

S. Sheremetyeva. 2005. Embedding MT for Generating Patent Claims in English from a Multilingual Interface. *Proceedings of Workshop on Patent Translation, MT Summit X.* Phuket

A. Shinmori, M. Okumura, Y. Marukawa, M. Iwayama. 2002. Rhetorical Structure Analysis of Japanese Patent Claims Using Cue Phrases. *Proceedings of the Third NTRCIR Workshop.*

A. Shinmori, M. Okumura, Y. Marukawa, M. Iwayama. 2003. Patent Claim Processing for Readability – Structure Analysis and Term Explanation. *Proceedings of the* ACL *Workshop on Patent Corpus Processing.* Sapporo*,* 56-65

H. Takahashi, T. Kawasaki, M. Makita, S. Higuchi, A. Fuji, T. Ishikawa. 2003. Nichi-ei tokkyo-koho-wo mochiita taiyaku-jisho-oyobi-honyaku-memori-no kochiku ["Building Translation Dictionaries and Translation Memories Using Japanese-US Patent Corpora"]. *IPSG-SIG Technical Report Vol. 2003 No. 057 (2003-NL-155)*, 39-46