On Integrating Hybrid And Rule-Based Components For Patent MT With Several Levels Of Output

Svetlana Sheremetyeva
South Ural State University/ 76 Lenin pr. Chelyabinsk 454080, Russia
LanA Consulting ApS / Moellekrog 4, Vejby, Copenhagen, Denmark
lanaconsult@mail.dk

Abstract

We present a methodology integrating hybrid and rule-based components for speeding up the development of a patent MT system. The methodology is suitable for highly inflecting languages and described on the example of translating patent claims from Russian into English. Based on different combinations of hybrid and rule-based components the system performs shallow or/and deep parsing and provides for several complementary levels of output, - (i) translation of terminology, that only involves shallow MT procedures, and (ii) full translation that is based on both shallow and deep parsing integrated either automatically, or in an interactive environment. Full translation of the patent claim is output in two formats, - a legal one sentence format and a better readable set of simple sentences. To control the quality of claim translation by better understanding the input, the system also outputs a SL claim decomposed into simple sentences.

1 Introduction

The wealth of technology contained in patents cannot be rated high enough. With ever exploding volume of patent documentation machine translation contributes a lot to strengthening the innovation process worldwide, removing language as a delimiting factor. In patent domain machine translation is a very challenging task. Only high quality patent translation could be used as a basis for important decisions on, e.g., novelty or the scope of inventor's rights. Quality requirement prompted the development of patent RBMT (Shimohata, 2005; Hong et al., 2005; Sheremetyeva, 2007; Wen and Jin, 2011) whose techniques promise correct translation but demand huge linguistic resources.

In an attempt to speed up the process of MT development and make it more robust, SMT and hybrid technologies (Ceausu et al., 2011; Eisele et al., 2008; Ehara, 2011; Espana-Bonet et al., 2011; Enache, 2012) came into patent domain. Though years of R&D in MT have resulted in great progress, the output of machine translation still cannot provide for required quality without human judgment (Koehn, 2009). In addition to traditional postediting recent work investigated the inclusion of interactive computer-human communication at each step of the translation process by, e.g., showing the user various "paths" among all translations of a sentence (Koehn, cf), or keyboard-driving the user to select the best translation (Macklovitch, 2006). One of the latest publications reports on Patent SMT from English to French where the user drives the segmentation of the input text (Pouliquen et.al, 2011). Popular SMT and hybrid techniques are problematic when dealing with inflecting languages. Statistical components of MT systems working well on configurational and morphologically poor languages, such as English, fail on non-configurational languages with rich morphology (Sharoff, 2004).

This paper reports on a novel hybrid methodology for developing an efficient patent MT system that can cope with translating patent claims. The methodology focuses on a highly inflecting SL and based on different combinations of hybrid and rule-based components provides for several complementary levels of output, - translation of terminology and full text translation in different formats. To improve the quality of full translation, the system includes an interactive module. To support quality control of

Proceedings of the 5th Workshop on Patent Translation, Nice, September 2, 2013. Yokoyama, S., ed.

claim translation the system helps the user to better understand the input by decomposing a SL claim into simple sentences thus improving its readability. Different levels of output and possible interactivity make the MT system useful for different types of users: TL-only speakers, SL-only speakers, people with some knowledge of both languages and professional translators.

The methodology is described on the model of a Russian-to-English MT system. In selecting Russian as our first inflecting SL we were motivated by two major considerations. Firstly, Russia has a huge pool of patents which are unavailable for non-Russian speakers without turning to expensive translation services. The situation is of great disadvantage for international technical knowledge assimilation, dissemination, protection of inventor's rights and patenting of new inventions. Secondly, Russian is an ultimate example of a highly inflecting language with a free word order. A typical Russian word has from 9 (for nouns) up to 50 forms (for verbs), which makes Russian a good testbed for hybrid MT covering inflecting languages.

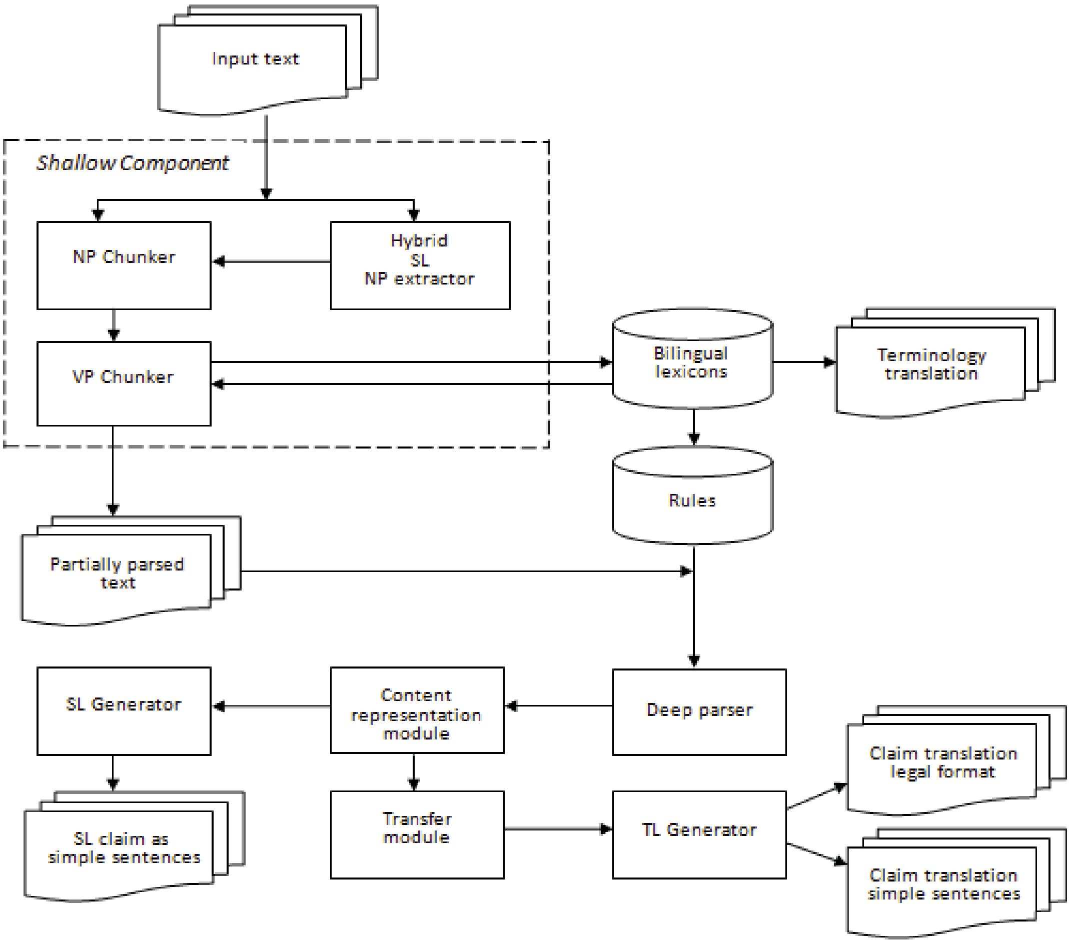


Figure 1. An overall architecture of the hybrid patent MT system with different levels of output.

2 System Overview

The system takes a Russian patent claim as input and produces translations at two major levels, - translation of claim terminology (not just any chunks), and full translation of a patent claim. Full translation of a patent claim is output in two formats, - in the form of one sentence meeting all legal requirements to the claim text, and as a better readable set of simple sentences. In addition, for the translator/posteditor to better control the quality of translation the system also improves the readability of a SL claim by decomposing it into a set of simple sentences. Partial output of the first translation level is useful for a non-Russian speaker for a quick patent digest to make a decision whether a full translation of a patent is needed. A list of translated terms is also useful for improving readability of a full claim translation[1].

This research extends our previous work on an RBMT system for translating patent claims between the low inflecting configurational English and Danish languages (Sheremetyeva, 2007). It partially reuses the program shell and some of the linguistic knowledge of its RB components. Necessary updates are made for the Russian language. The top architecture of the system follows the traditional RBMT schema, - SL analyzer – transfer -TL generation, but instead of a fully rule-based analyzer the current system includes a hybrid parser with shallow and deep components that lifts a lot of ambiguity problems and makes the whole parsing easier, less resource consuming and more robust. The core of the shallow parser is a hybrid Russian NP extractor which is a standalone tool that was integrated into the system. The full Russian parser and transfer module are designed so as to produce a final parse of a Russian patent claim in the format acceptable by the English claim generator from the earlier application. The architecture of the system is shown in Figure 1.

3 Knowledge

Patent claims must be formulated as specified by the German Patent Office and commonly accepted in Europe, the U.S., Russia and other countries. The claim must describe essential features of the invention in the obligatory form of a single extended nominal sentence with a well-specified conceptual, syntactic and stylistic/rhetorical structure.

For successful translation of patent claims two distinct types of expert knowledge are necessary: knowledge about the sublanguage of patents as legal documents and knowledge about the technical field of the invention. Both kinds of knowledge are mainly encoded in the lexicons:

(i) **a shallow bilingual (Russian/English) lexicon**, where the units are listed with their morphological features. This is the type of resource that, once build for some other purpose, can be simply fed into the system. We had a successful experience of pipelining such knowledge into an MT system in our Japanese-English project (Neumann, 2005). Acquisition of this type of knowledge for every new pair of languages is what existing SMT tools can provide either in advance or on the fly, as reported in (Enache et al., cf). We, therefore, do not dwell on acquisition of this type of resource. To demonstrate the viability of the methodology we will use our own limited semi-manually acquired set of bilingual terminological data.

(ii) **a deep (information-rich) bilingual lexicon of predicates** used in the English and Russian language patent claims; this lexicon has been specifically constructed for the current application and is meant for a multifunctional use in the modules of the system.

3.1 Deep lexicon and content representation language

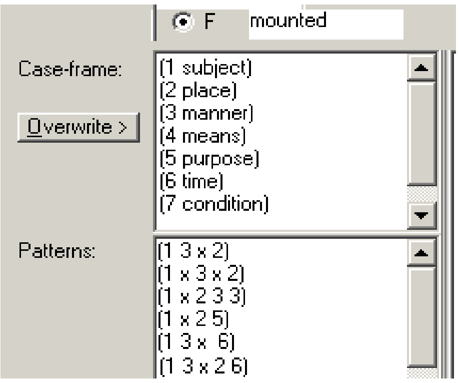
The core of the system knowledge is a deep bilingual Russian-English predicate lexicon which is organized as a set of cross-referenced set of monolingual entries and contains lexical, morphological, syntactic and semantic knowledge. Syntactic and semantic zones are as follows:

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 - code both the co-occurrences of predicates with their case-roles, and their linear order in the claim text.

Figure 2 presents a fragment of the entry of the predicate "mounted" with the case-roles and pattern zones.



● F mounted

Case-frame:
(1 subject)
(2 place)
(3 manner)
(4 means)
(5 purpose)
(6 time)
(7 condition)

Patterns:
(1 3 x 2)
(1 x 3 x 2)
(1 x 2 3 3)
(1 x 2 5)
(1 3 x 6)
(1 3 x 2 6)

Figure 2. a fragment of the entry of the predicate "mounted".

[1] It is well known that due to an extremely complex syntactic structure of the patent claim that can run for a page or so, the problem of patent readability is an issue even in a SL (Shinmori et al.,2003), let alone in translation.

The pattern (1 3 x 2), for example, can trig the realization of such clam fragment as

1:devices 3:rotatably x:mounted 2:on the leg.

3.2 Content representation language

The knowledge in predicate entries is used to support the claim content representation language as shown below.

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

where "label" is a unique identifier of a predicate/argument structure, "predicate-class" is a label of a semantic class, "predicate" is a string corresponding to a predicate, "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 (*place, instrument*, etc.,) and "value" is a case-role filler.

4 Hybrid parser

4.1 Shallow component

Russian NP extractor. The core of the shallow parsing component is a hybrid Russian NP extractor which is a standalone tool[2] pipelined to the system. It was built following the methodology of keyword extraction for the English language described in (Sheremetyeva 2009). The extractor does not rely on a preconstructed corpus, works on small texts, does not miss low frequency units and can reliably extract all NPs from an input text. The extraction methodology combines statistical techniques, heuristics and very shallow linguistic knowledge that includes a number of shallow lexicons (sort of extended lists of stop words) forbidden in a particular (first, middle or last) position in the typed unit (Russian NP in our case) to be extracted.

NP extraction starts with n-gram calculation and then removes n-grams which cannot be NPs by matching components of calculated n-grams against the stop lexicons. The extraction itself thus neither requires such NLP procedures, as tagging, morphological normalization, etc., nor does it rely on statistical counts (statistical counts are only used to sort out keywords). The latter makes this extraction methodology suitable for inflecting languages (Russian in our case) where frequencies of n-grams are low.

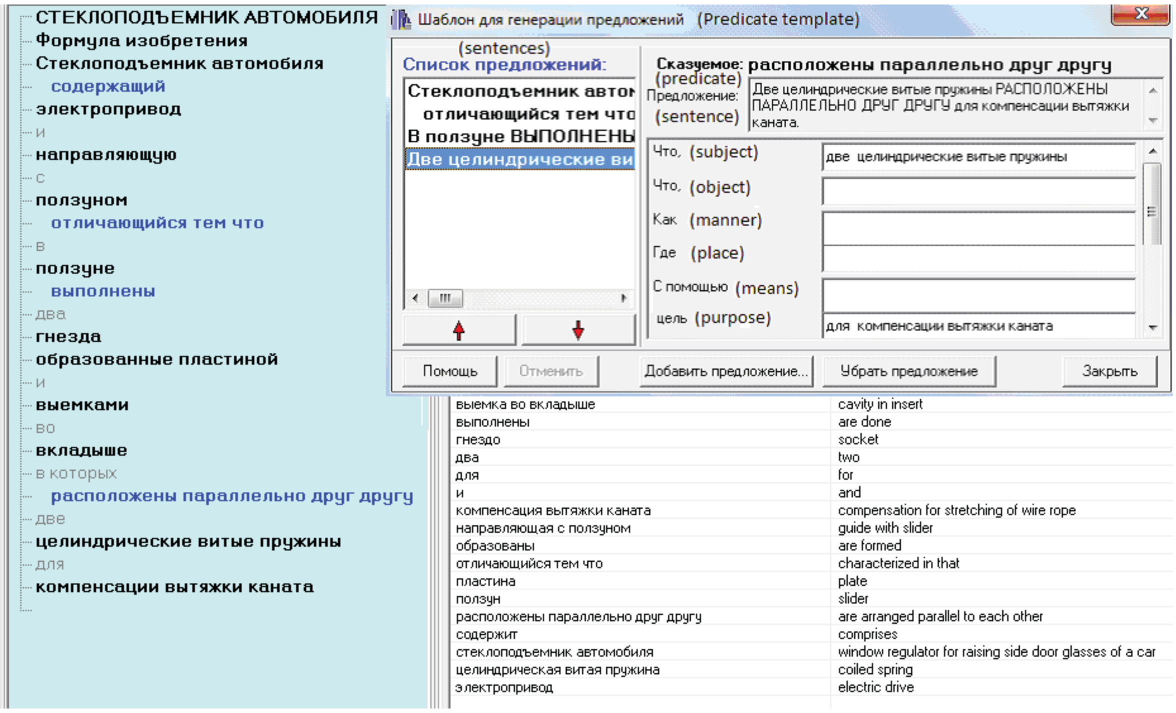
Porting the NP extractor from English to Russian consisted in substituting English stop lexicons of the tool with the Russian equivalents. We did this by translating each of the English stop lists into Russian using a free online system PROMT followed by manual brush-up. The extracted NP phrases are of 1 to 4 components due to the limitations of the extractor 4-gram model. We did not lemmatize the output of the extractor. All extracted Russian NP strings keep their text forms. This allows straightforward bracketing of these NPs in the claim text by simple matching the extracted NPs against the text. The remaining unbraketed text of the input is then matched against the morphological fields of the predicate entries in the predicate lexicon and, in case of a match, a predicate is chunked (bracketed) in the input text. This practically lifts the problem of lexical ambiguity between the forms of verbs and other parts-of-speech. Being identified as NP components and enclosed in brackets a lot of ambiguous words are simply not submitted to the VP chunker. The order of NP and VP (predicates) chunking is relevant. Noun phrases are chunked first as they are the most frequent types of phrases and thus leave less "residue text" for VP identification reducing the ambiguity.

The output of the shallow parsing components is then submitted to the bilingual lexicon with the help of which the first (partial) level of translation is performed. An example of such partial translation is shown in Figure 3 (right, bottom). If the goal of the user is just a digest, the MT procedure can stop right here. Otherwise, the shallow parse is input to the deep component that have two modes, - automatic and interactive.

4.2 Deep component

Automatic mode. The deep component takes a partially parsed claim from the shallow component as input and automatically completes the parsing procedure. It uses the knowledge from the deep lexicon and rules of our application-specific grammar, - a mixture of context free lexicalized Phrase Structure Grammar and Dependency Grammar.

[2] This tool can be used for different purposes, e.g., we also used its English and Russian versions for the acquisition of Russian-English lexicon by running it on available parallel and comparable corpora



СТЕКЛОПОДЪЕМНИК АВТОМОБИЛЯ
Формула изобретения
Стеклоподъемник автомобиля
содержащий
электропривод
и
направляющую
с
ползуном
отличающийся тем что
в
ползуне
выполнены
два
гнезда
образованные пластиной
и
выемками
во
вкладыше
в которых
расположены параллельно друг другу
две
целиндрические витые пружины
для
компенсации вытяжки каната

Шаблон для генерации предложений (Predicate template)

(sentences)
Список предложений:
Стеклоподъемник автом
отличающийся тем что
В ползуне ВЫПОЛНЕНЫ
Две целиндрические ви

Сказуемое: (predicate) расположены параллельно друг другу
Предложение: (sentence) Две целиндрические витые пружины РАСПОЛОЖЕНЫ ПАРАЛЛЕЛЬНО ДРУГ ДРУГУ для компенсации вытяжки каната.

Что, (subject) две целиндрические витые пружины
Что, (object)
Как (manner)
Где (place)
С помощью (means)
цель (purpose) для компенсации вытяжки каната

выемка во вкладыше	cavity in insert
выполнены	are done
гнездо	socket
два	two
для	for
и	and
компенсация вытяжки каната	compensation for stretching of wire rope
направляющая с ползуном	guide with slider
образованы	are formed
отличающийся тем что	characterized in that
пластина	plate
ползун	slider
расположены параллельно друг другу	are arranged parallel to each other
содержит	comprises
стеклоподъемник автомобиля	window regulator for raising side door glasses of a car
целиндрическая витая пружина	coiled spring
электропривод	electric drive

Figure 3. A screenshot of the user interface in the interactive mode. In the left pane a chunked input with highlighted NPs and VPs is displayed. On the bottom of the right pane the first level of translation results are shown. On the top in the right pane an interactive predicate template is presented which pops-up after the user clicks on a corresponding highlighted predicate.

The deep parser includes newly developed components as a Russian disambiguating tagger, a Russian bottom-up heuristic parser with a recursive pattern matching technique to recursively chunk all types of Russian phrases (NPs[3], PPs, AdvP, etc.). It preserves the inner structure of the longer chunks and marks the head of every noun phrase with its "singular/plural" feature.

At its last run the parser determines the semantic dependency relations between the identified chunks and predicates, - and assigns to the chunks their semantic status as particular case-roles of the governing predicate (see Sheremetyeva, 2003 for details).

Shallow "pre-parsing" significantly reduces all kinds of ambiguity at all stages of processing and decreases the number of rules. The final parse is then supplied into the Transfer and SL generation module to get a full translation of the claim. The final parse of the claim text displayed in the left pane of Figure 3 is shown in Figure 4 (left pane).

Interactive mode. In the interactive mode of the deep parsing component the system guides the user through the paces of "understanding" the structure of the SL claim by decomposing a complex input text into predicate phrases (simple sentences) describing individual features of the invention "disguised" in the complex telescopic claim structure. The system supports user elicitation decisions with instructions, and highlighted SL noun terms and predicates (see Figure 3, left pane). Once the user clicks on a highlighted predicate, a corresponding elicitation template is displayed in a separate pop-up window (Figure 3, top). The template is based on the knowledge in the case-role zone in the lexicon entry of the selected predicate. The user then fills the slots of the template with text elements by simply clicking on them in the interactively marked (chunked) input document. The slot fillers can be edited by supplying chunks of the text into the slots of predicate templates the user determines the dependency relations between the predicates and other chunks and defines the semantic status of these chunks as case-roles of the governing predicate. The output of this interaction proce-

[3] The deep parser combines NPs chucked by the shallow component into longer nominal phrases.

dure is a set of predicate/argument structures with partially parsed case-role fillers, which are further input into the deep parsing component. The deep parser automatically completes case-role fillers tagging and recursive chunking and outputs a set of predicate/argument structures as shown in Figure 4 (left pane). This content representation is then submitted into two system modules, – the Russian generator that outputs a Russian claim in a more readable format of simple sentences, and to the Russian-English transfer module.

4.3 Transfer module

The transfer module is fully automatic. It takes the deep parser output, - a SL set of predicate templates as input and outputs a set of TL predicate templates whose slots are filled with translated TL phrases (case-role fillers). The transfer procedure is a combination of interlingual and lexical-syntactic transfer. The interlingual transfer is based on the knowledge about predicate case-roles in the deep lexicon. It finds structural TL equivalents for every SL predicate/argument structure. The TL predicate gloss is substituted with its TL equivalent. Then the SL fillers of the case-roles are translated (See Figure 4). 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. Translation of case-role fillers can be outsourced to a foreign MT system and then put back into a predicate-argument structure. As was mentioned above this is where SMT techniques can be particularly useful.

4.4 Generation module

The claim text generation stage takes an English-oriented text representation (Figure 4, right pane) as input, and submits it to an automatic text planner which outputs a hierarchical structure of predicate templates.

The planning stage is guided both by constraints on the patent claim sublanguage and the general constraints on style. The former determines the global ordering of the claim text while the latter deals with local text coherence.

The realization stage of the generator linearizes the hierarchy of TL predicate templates and takes care of the ellipsis, conjoined structures, punctuation and morphological forms. The generic part and novelty part of the claim are generated separately. The two completely ready parts of the claim text are bound by the intermediate expression "characterized in that", the generic and novelty parts being put correspondingly before and after this expression. The output is an English text of the claim in a legal format (see Figure 5, top). Parallel to this a better readable translation of the same claim in the form of simple sentences is also generated (Figure 5, bottom).

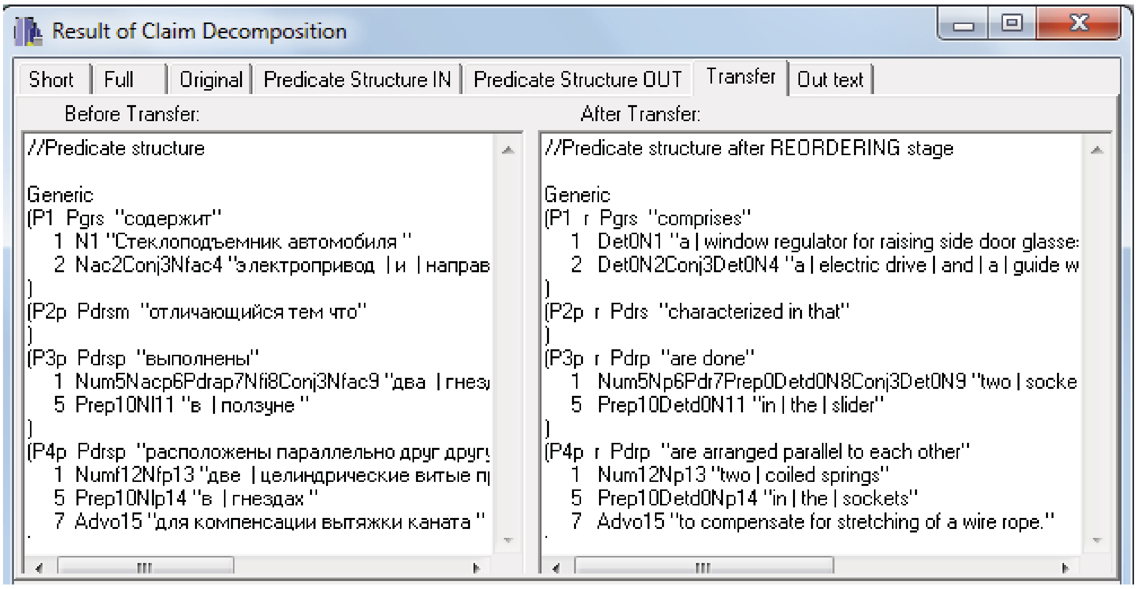
5 Status and Discussion

The methodology we have described in this paper has been implemented in a Russian-English hybrid MT system for patent claims. The system is in the late stages of development as of June 2013. The static knowledge sources have been compiled for the domain of patents about vehicles. The programming shell of the system is completed and provides for knowledge administration in all modules of the system to improve their performance. The extractor of Russian nominal terminology currently preforms with 98,4 % of recall and 96,1% precision.

The shallow clunker based on the extraction results and predicate knowledge shows even higher accuracy. This is explained, on the one hand, by the high performance of the Russian extractor, and, on the other hand, by the nature of highly inflecting languages. Rich morphology turns out to be an advantage in our approach. Great variety of morphological forms significantly lowers ambiguity between NP components and verb paradigms. We have tested shallow chunking on patents in English and, though the efficiency of English and Russian NP extractors is practically the same, chunking of the English NPs in claim texts is rather problematic due to much higher ambiguity of wordforms in English. Our conclusion is that shallow chunking based on unlemmatized extraction results better suites inflecting languages.

The interactive semantic and syntactic analysis module for the Russian language and the English generator are fully developed using the technology of earlier applications. The Russian-to-English transfer module responsible for lexical transfer and case-role translation is workable.

In the deep parsing component the morphological analysis of Russian and syntactic chunking are operational and well tested. The case-role dependency detection in the automatic mode is being currently tested and updated. We have not yet made a large-scale evaluation of our deep analysis module. This leaves the comparison between other parsers and our approach as a future work. In general preliminary MT results show a reasonably small number of failures that are being improved by brushing up the shallow knowledge and by larger involvement of predicate knowledge. This proves the viability of the suggested MT methodology. We intend to a) improve the quality of the automatic mode of our MT system by updating system knowledge based on extensive testing; b) develop a patent search and extraction facility on the basis of the patent sublanguage and our parsing strategy.



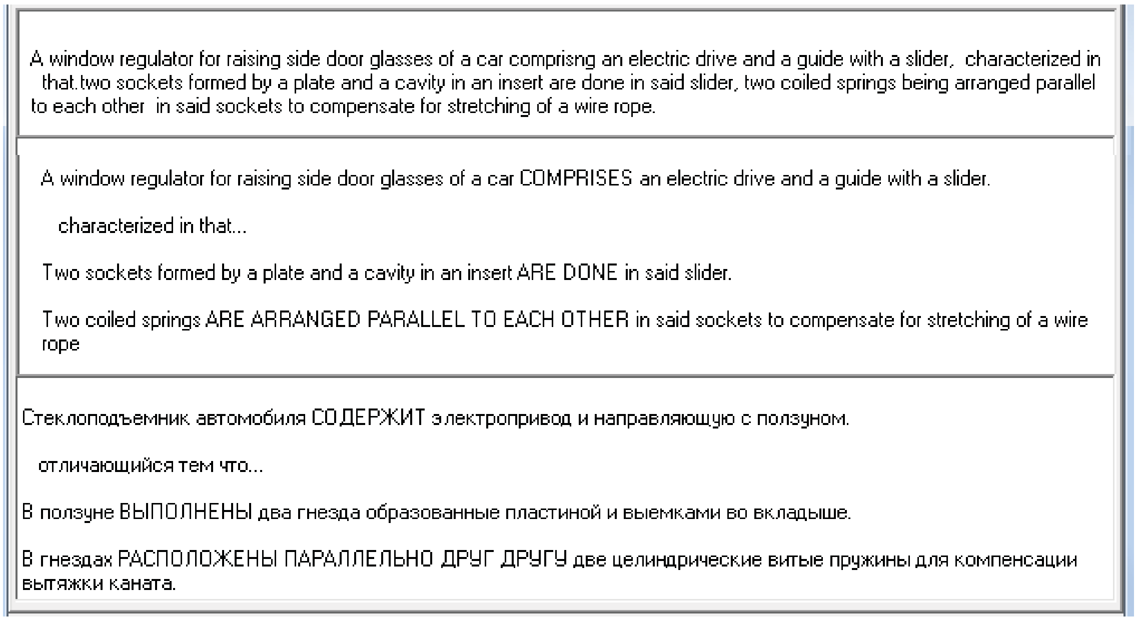
```
//Predicate structure

Generic
(P1  Pgrs  "содержит"
   1  N1 "Стеклоподъемник автомобиля "
   2  Nac2Conj3Nfac4 "электропривод  | и  | направ
)
(P2p  Pdrsm  "отличающийся тем что"
)
(P3p  Pdrsp  "выполнены"
   1  Num5Nacp6Pdrap7Nfi8Conj3Nfac9 "два  | гнез,
   5  Prep10NI11 "в  | ползуне "
)
(P4p  Pdrsp  "расположены параллельно друг друг
   1  Numf12Nfp13 "две  | целиндрические витые п|
   5  Prep10NIp14 "в  | гнездах "
   7  Advo15 "для компенсации вытяжки каната "
.
```

```
//Predicate structure after REORDERING stage

Generic
(P1  r  Pgrs  "comprises"
   1   Det0N1 "a | window regulator for raising side door glasse:
   2   Det0N2Conj3Det0N4 "a | electric drive | and | a | guide w
)
(P2p  r  Pdrs  "characterized in that"
)
(P3p  r  Pdrp  "are done"
   1   Num5Np6Pdr7Prep0Detd0N8Conj3Det0N9 "two | socke
   5   Prep10Detd0N11 "in | the | slider"
)
(P4p  r  Pdrp  "are arranged parallel to each other"
   1   Num12Np13 "two | coiled springs"
   5   Prep10Detd0Np14 "in | the | sockets"
   7   Advo15 "to compensate for stretching of a wire rope."
.
```

Figure 4. A screenshot of the developer's interface. On the left pane shown is the final parse of the example claim shown in Figure 3 (left pane). The output of the transfer module is shown on the right.



A window regulator for raising side door glasses of a car comprisng an electric drive and a guide with a slider, characterized in that.two sockets formed by a plate and a cavity in an insert are done in said slider, two coiled springs being arranged parallel to each other in said sockets to compensate for stretching of a wire rope.

A window regulator for raising side door glasses of a car COMPRISES an electric drive and a guide with a slider.

characterized in that...

Two sockets formed by a plate and a cavity in an insert ARE DONE in said slider.

Two coiled springs ARE ARRANGED PARALLEL TO EACH OTHER in said sockets to compensate for stretching of a wire rope

Стеклоподъемник автомобиля СОДЕРЖИТ электропривод и направляющую с ползуном.

отличающийся тем что...

В ползуне ВЫПОЛНЕНЫ два гнезда образованные пластиной и выемками во вкладыше.

В гнездах РАСПОЛОЖЕНЫ ПАРАЛЛЕЛЬНО ДРУГ ДРУГУ две целиндрические витые пружины для компенсации вытяжки каната.

Figure 5. Examples of MT output. On the top a full claim translation into English in the legal format is shown. In the middle the "better readable" claim translation in the form of simple sentences is shown. In the bottom a decomposed Russian input claim is given.

References

Ceausu, Alexandru, John Tinsley, Jian Zhang, and Andy Way. 2011. *Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project*. In Proceedings of EAMT 2011:

Ehara Terumasa, 2011 Machine translation system for patent documents combining rule-based translation and statistical postediting applied to the Patent MT Task. *Proceedings of NTCIR-9 Workshop Meeting,* 2011, Tokyo, Japan

Eisele, A., C. Federmann, H. Saint-Amand, M. Jellinghaus, T. Herrmann, and Y. Chen. 2008. Using moses to integrate multiple rule-based machine translation engines into a hybrid system. *In Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08.

Enache Ramona, Cristina Espa˜na-Bonet Aarne Ranta Llu´ıs M`arquez. 2012. A Hybrid System for Patent Translation. *Proceedings of the EAMT Conference.* Trento..Italy, May

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

Koehn Philipp. 2009. A process study of computer-aided translation, *Philipp Koehn*, Machine Translation Journal, 2009, volume 23, number 4, pages 241-263

Macklovitch, Elliott. 2006. TransType2: The last word. *In proceedings of LREC06*, May 2006,Genoa, Italy

Neumann Ch. 2005. A Human-Aided Machine Translation System for Japanese-English Patent Translation. *Proceedings of the Workshop on Patent Translation MT Summit*, Phuket, Thailand,.

Pouliquen Bruno, Christophe Mazenc Aldo Iorio. 2011. Tapta: A user-driven translation system for patent documents based on domain-aware Statistical Machine. *Proceedings of the EAMT Conference*. Leuven, Belgium, May.

Sharoff, Serge . 2004. What is at stake: a case study of Russian expressions starting with a preposition. *Proceedings of the ACL Workshop on Multiword Expressions: Integrating Processing*, July.

Sheremetyeva Svetlana. 2003. Natural Language Analysis of Patent Claims. *Proceedings of the workshop "Patent Corpus Processing" in conjunction with ACL 2003,* Sapporo. Japan, July.

Sheremetyeva Svetlana. 2007. On Portability of Resources for a Quick Ramp up of Multilingual MT of Patent Claims. *Workshop on Patent Translation. In conjunction of Machine Translation Summit XI.Copenhagen.*Denmark. September

Sheremetyeva, Svetlana. 2009. On Extracting Multiword NP Terminology for MT. *Proceedings of the EAMT Conference.* Barcelona, Spain, May

Shimohata S. 2005. Finding Translation Candidates from Patent Corpus. *Machine Translation Summit X, Workshop on Patent Translation.*

Shinmori A., Okumura M., Marukawa Y. Iwayama M. 2003. Patent Claim Processing for Readability - Structure Analysis and Term Explanation, *Workshop on Patent Corpus Processing. conjunction with ACL 2003,* Sapporo. Japan, July.

Wen Xiong, Yaohong Jin. 2011. A new Chinese-English machine translation method based on rule for claims sentence of Chinese patent; In Proceedings of 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE).