On-The-Fly Translator Assistant (Readability and Terminology Handling)

Svetlana Sheremetyeva
National Research South Ural State University / pr.Lenina 74, 454080
Chelyabinsk, Russia
LanA Consulting ApS/ Moellekrog 4, Vejby, 3210, Copenhagen, Denmark
lanaconsult@mail.dk

Abstract

This paper describes a new methodology for developing CAT tools that assist translators of technical and scientific texts by (i) on-the-fly highlight of nominal and verbal terminology in a source language (SL) document that lifts possible syntactic ambiguity and thus essentially raises the document readability and (ii) simultaneous translation of all SL document one- and multi-component lexical units. The methodology is based on a language-independent hybrid extraction technique used for document analysis, and language-dependent shallow linguistic knowledge. It is targeted at intelligent output and computationally attractive properties. The approach is illustrated by its implementation into a CAT tool for the Russian-English language pair. Such tools can also be integrated into full MT systems.

1 Introduction

Exploding volume of professional publications demand operative international exchange of scientific and technical information and thus put in focus operativeness and quality of translation services. In spite of the great progress of MT that saves translation time, required translation quality so far cannot be achieved without human judgment (Koehn, 2009). Therefore in great demand are CAT tools designed to support and facilitate human translation.

CAT tools are developed to automate postediting and often involve controlled language. The most popular tools are translation memory (TM) tools whose function is to save the translation units in a database so that they can be re-used through special "fuzzy search" features. The efficiency of TM (as well as translation quality as such) is directly related to the problem of the comprehensiveness of multilingual lexicons. A translator who, as a rule, does not possess enough of expert knowledge in a scientific or technological domain spends about 75% of time for translating terminology, which do not guarantee the correctness of translation equivalents she/he uses. The percentage of mistakes in translating professional terminology reaches 40% (Kudashev, 2007). It is therefore essential to develop methodologies that could help human translators solve this problem, the huge resource being the Internet, if properly used. In this paper we suggest one of the possible ways to do so.

We would like to address the importance of text readability in the human translation performance. Readability relates to (though does not coincide with) the notion of translatability in MT research. Readability in human translation is associated with the level of clarity of a SL text for human understanding. Every translator knows how difficult it can be to understand professional texts, not only because of the abundance of terminology but also due to complex syntax and syntactic ambiguity. The ultimate example of a low readability text is the patent claim (Shinmori et al., 2003) that is written in the form of one nominal sentence with extremely complex “inhuman” syntactic structure that can run for a page or more. Low readability is often the case with scientific and technical papers as well.

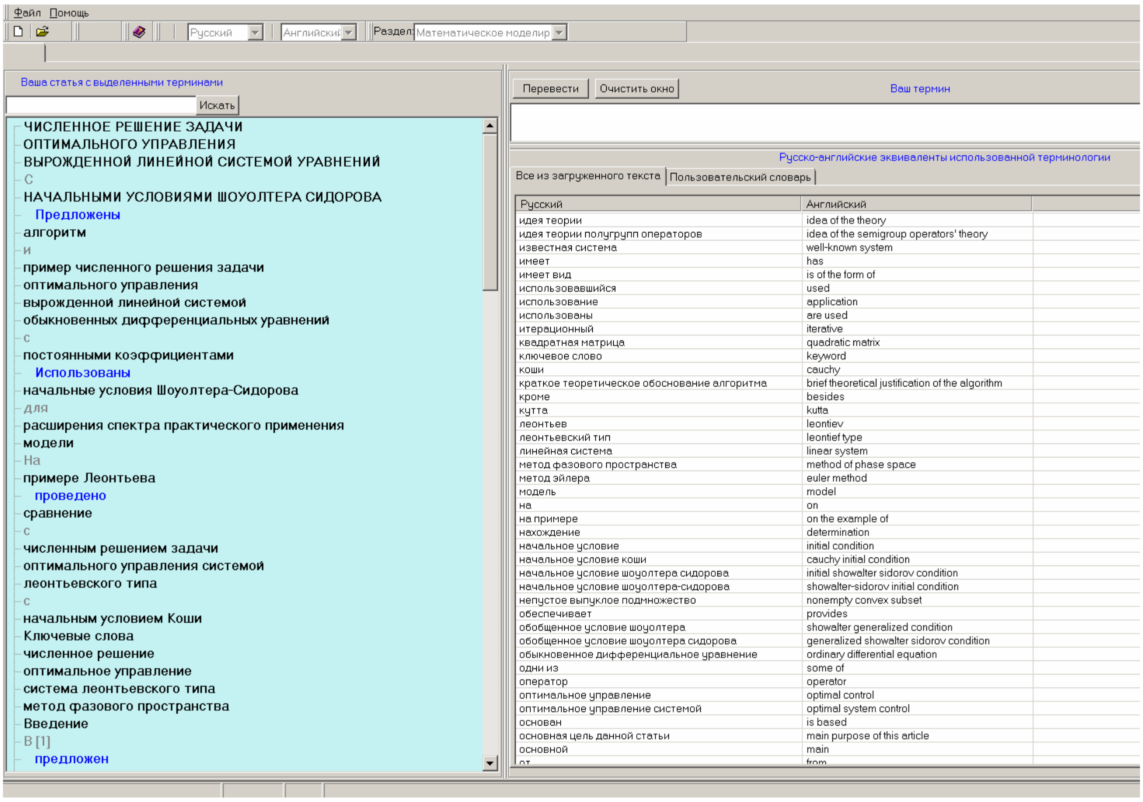
Workshop on Humans and Computer-assisted Translation, pages 22–27,
Gothenburg, Sweden, 26 April 2014. ©2014 Association for Computational Linguistics

In this paper we describe our effort to develop a portable between domains and languages CAT tool that can on-the-fly improve the readability of professional texts and provide for reliable terminology translation.

We paid special attention to multiword noun terminology, the most frequent and important terminological unit in special texts that can rarely be found in full in existing lexicons. When translated properly, multicomponent NPs do not only provide for the correct understanding of the corresponding target language (TL) term but in many cases lift syntactic ambiguity.

The tool can find a broad application, e.g., it can be useful for any non-SL speaker for a quick document digest. The settings of the tool allow the extraction of keyword translation pairs in case it is needed, e.g., for search purposes. It can also be integrated into a full MT system.

We implemented our methodology into a fully functional tool for the Russian-English language pair and conducted experiments for other domains and language pairs. In selecting Russian as a first SL we were motivated by two major considerations. Firstly, Russia has a huge pool of scientific and technical papers which are unavailable for non-Russian speakers without turning to expensive translation services. Secondly, our scientific challenge was to develop a hybrid methodology applicable to inflecting languages. Popular SMT and hybrid techniques working well on configurational and morphologically poor languages, such as English, fail on non-configurational languages with rich morphology (Sharoff, 2004). Russian is an ultimate example of such a language. It has a free word order; a typical Russian word has from 9 (for nouns) up to 50 forms (for verbs). In what follows we first present the tool and then describe the underlying methodology.


Файл
Помощь
Русский
Английский
Раздел:
Математическое моделир
Ваша статья с выделенными терминами
Искать
Перевести
Очистить окно
Ваш термин
Русско-английские эквиваленты использованной терминологии
Все из загруженного текста
Пользовательский словарь

ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧИ
ОПТИМАЛЬНОГО УПРАВЛЕНИЯ
ВЫРОЖДЕННОЙ ЛИНЕЙНОЙ СИСТЕМОЙ УРАВНЕНИЙ
С
НАЧАЛЬНЫМИ УСЛОВИЯМИ ШОУОЛТЕРА СИДОРОВА
Предложены
алгоритм
и
пример численного решения задачи
оптимального управления
вырожденной линейной системой
обыкновенных дифференциальных уравнений
с
постоянными коэффициентами
Использованы
начальные условия Шоуолтера-Сидорова
для
расширения спектра практического применения
модели
На
примере Леонтьева
проведено
сравнение
с
численным решением задачи
оптимального управления системой
леонтьевского типа
с
начальным условием Коши
Ключевые слова
численное решение
оптимальное управление
система леонтьевского типа
метод фазового пространства
Введение
В [1]
предложен

Русский	Английский
идея теории	idea of the theory
идея теории полугрупп операторов	idea of the semigroup operators' theory
известная система	well-known system
имеет	has
имеет вид	is of the form of
использовавшийся	used
использование	application
использованы	are used
итерационный	iterative
квадратная матрица	quadratic matrix
ключевое слово	keyword
коши	cauchy
краткое теоретическое обоснование алгоритма	brief theoretical justification of the algorithm
кроме	besides
кутта	kutta
леонтьев	leontiev
леонтьевский тип	leontief type
линейная система	linear system
метод фазового пространства	method of phase space
метод эйлера	euler method
модель	model
на	on
на примере	on the example of
нахождение	determination
начальное условие	initial condition
начальное условие коши	cauchy initial condition
начальное условие шоуолтера сидорова	initial showalter sidorov condition
начальное условие шоуолтера-сидорова	showalter-sidorov initial condition
непустое выпуклое подмножество	nonempty convex subset
обеспечивает	provides
обобщенное условие шоуолтера	showalter generalized condition
обобщенное условие шоуолтера сидорова	generalized showalter sidorov condition
обыкновенное дифференциальное уравнение	ordinary differential equation
одни из	some of
оператор	operator
оптимальное управление	optimal control
оптимальное управление системой	optimal system control
основан	is based
основная цель данной статьи	main purpose of this article
основной	main
от	from

Figure 1. A screenshot of the Russian-to-English CAT tool user interface at the bookmark “show all”. The left pane displays a SL interactive text of a scientific paper in mathematical modelling with explicitly marked (bold faced) nominal terminology and verbs (in blue). The left pane contains the alphabetically ordered list of all 1-4 component Russian terms with their English equivalents. On the top of the right pane there is a type-in area which permits searching for the translations of terms longer than 4 words in the tool knowledge base. The second bookmark on the top of the Ru-En equivalent area allows opening a user dictionary for the user to collect terms she/he might need in the future.

2 The Tool

The tool takes a SL text an as input and on the fly produces output at two levels:

- a marked-up interactive SL text with highlighted multi-component nominal and verbal terminology (NPs and VPs);
- a list of all single- and multi-component SL-TL units found in the input text.

Text mark-up improves input readability and helps translator quicker and better understand the syntactic structure of the input. This feature combined with on-the-fly translation of *all* 1-4 component SL text lexical units reduces translation time and effort and raises translation quality. The tool can be used as an e-dictionary where terms are searched through a type-in area in the user interface.

Translation equivalents are normalized as follows. SL NPs are outputted in nominative singular, while VPs are presented in a finite form keeping the SL voice, tense and number features. For example, in the Russian-to-English tool the Russian VP wordform "*смонтированные*"*_past participle, perfective, plural* (literally *"done")* will be outputted as "*смонтированы*"*_ finite, past, plural* = *"were mounted"*.

The tool user interface has a lot of effort-saving functionalities. A click on a unit in the marked up input text in the left pane highlights its TL equivalent in the alphabetically sorted list of translations on the right pane. It is possible to create user dictionaries accumulating terminology from different texts, saving these dictionaries and projects, etc. A screenshot of the user interface in shown in Figure 1.

3 Methodology and Development Issues

3.1 Architecture

The overall architecture of the tool is shown in Figure 2. The tool engine consists of a shallow analyzer including three fully automatic modules, - a SL hybrid NP extractor, shallow parser and imbedded machine translation module meant to translate terminology. The knowledge base contains shallow linguistic knowledge, - lexicons and rules.

The NP extractor is a hybrid stand-alone tool pipelined to the system. We built it following the methodology of NP extraction for the English language as described in (Sheremetyeva, 2009) and ported it to the Russian language.

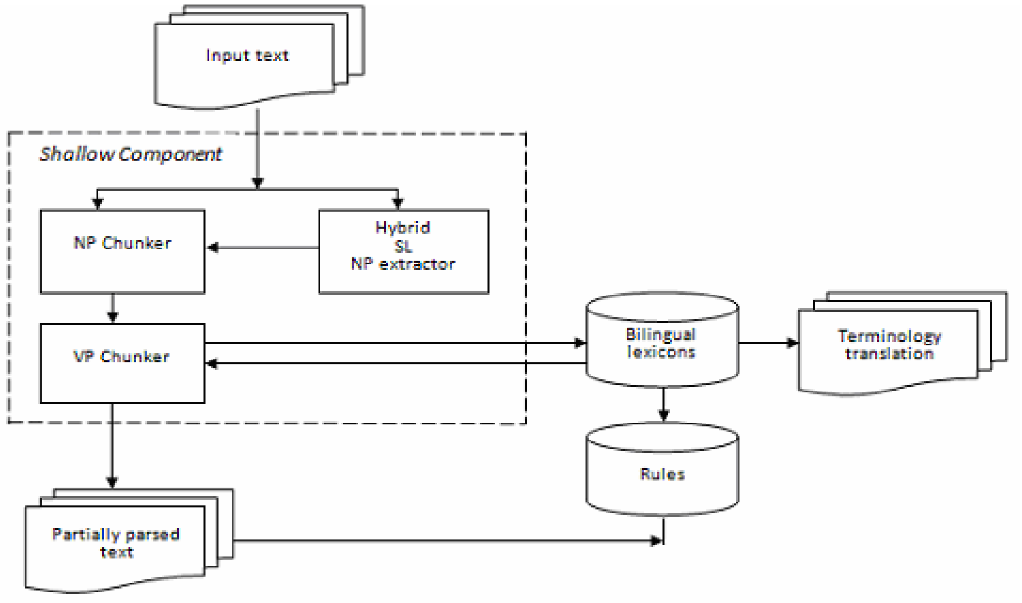


Figure 2. The architecture of the CAT tool.

The extraction methodology combines statistical techniques, heuristics and very shallow linguistic knowledge. The knowledge base consists of a number of unilingual lexicons, - sort of extended lists of stop words forbidden in particular (first, middle or last) positions in a typed lexical unit (Russian NP in our case).

NP extraction procedure starts with n-gram calculation and then removes n-grams, which cannot be NPs by successive matching components of calculated n-grams against the stop lexicons. The extraction itself thus neither requires such demanding NLP procedures, as tagging, morphological normalization, POS pattern match, 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 (http://www.translate.ru) followed by manual brush-up.

The NP 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. We excluded a lemmatizer from the original extraction algorithm and kept all extracted Russian NPs in their textual forms. The noun phrases thus extracted are of 1 to 4 components due to the limitations of the extractor that uses a 4-gram model. The extractor was also used for lexicon acquisition.

The shallow parser consists of an NP chunker, VP chunker and tagger. The first users the knowledge dynamically produced by the NP extractor (lists of all NPs of an input text in their text form). The VP chunker and tagger turn to the Russian entries of the tool bilingual lexicon. The tagger is actually a supertagger as it assigns supertags coding all morphological features, such as part-of-speech, number, gender, tense, etc.

The machine translation module translates text chunks into English using simple transfer and generation rules working over the space of supertags as found in the CAT tool bilingual lexicon.

3.2 Bilingual lexicon

To ensure correct terminology translation the bilingual lexicon of the tool should necessarily be tuned to a specific domain for which it is to be used. The lexicon is organized as a set of shallow cross-referenced monolingual entries of lexical units listed with their part-of-speech class and explicit paradigms of domain-relevant wordforms. This is the type of resource that, once build for some other purpose, can be simply fed into the system. 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 (2012 et al.,). In our work striving for correctness we combined automatic techniques with manual check and manual acquisition.

The Russian vocabulary was created in two steps. First, an initial corpus of Russian scientific papers on mathematical modelling of approximately 80 000 wordforms was acquired on Internet. We then ported the NP extractor described above to other Russian parts-of-speech and automatically extracted domain specific typed lexical units (NPs, VPs, ADJs, etc) consisting of 1 up to 4 components from the corpus. These automatically extracted lists of lexemes were further checked by human acquirers and 14 000 of them were used as a seed Russian vocabulary.

The seed vocabulary was then used to acquire longer Russian lexemes both from the initial corpus, and the Internet, which is in fact an unlimited corpus. The following methodology was applied. The seed lexical units were used as keywords in the Internet search engines. New Russian terminological units including seed terms highlighted in the two first pages of the search results were included in the lexicon. For example, for the seed (key) term *«псевдообращение»* the following multi-component terms popped-up on the Internet: *«псевдообращение сопряженной системы», «псевдообращение матриц с вырожденными весами», «псевдообращение Мура-Пенроуза»,* etc. As a result, the seed Russian vocabulary was extended to 60 000 single- and multi-component units up to seven-eight words long.

Lexical acquisition of English equivalents was done based on existing domain lexicons, parallel/comparable corpora and raw Internet resources. The last needs to be explained. In case neither existing lexicons, nor parallel/comparable corpora could provide for a reliable English equivalent, which was mostly the case with long terms, translation hypotheses were made based on different combinations of translation variants of component words. Every translation hypothesis was then checked in the Internet search engine. If an engine (we used Google) showed a translation version in the search results, the hypothesis was considered confirmed and the English equivalent was included in the tool lexicon. For example, the Russian term *«роевое представление частицы»* could not be found in any of existing lexicons, the following English equivalents of the Russian term components were found:

рой – swarm; представление - conception, expression, representation, performance, configuration; частица – bit, fraction, particle, shard, corpuscle.

If you create a translation hypothesis by using the first translation variant for every component of the Russian term you will get: *«swarm conception of a bit» or «bit swarm conception».* Used as key words in Google, the search results do not contain these words combined in a term. This translation hypothesis was rejected. Another hypothesis *«particle swarm representation»* used as key words in Google gives the English term *«Particle Swarm Optimization and Priority Representation»* from the paper on mathematical modelling by Philip Brooks, a native English speaker. «Particle swarm representation» is accepted as a correct English translation of the Russian term **«***роевое представление частицы***»**. Though tedious, this methodology allowed careful detection of the up-to-date highly reliable translation that could hardly be achieved otherwise.

3.3 Workflow

The raw SL document first goes to the automatic NP extractor, which produces a list of one- to four component noun phrases. The dynamically created NP list is then used as knowledge for the NP chunker, which by matching the extracted list against the input text chunks (brackets) noun phrases in the document. The morphological tagger completes morphological analysis of these chunks by looking them up in the NP entries of the tool lexicon. The text strings between chunked NPs is then supplied to the VP chunker that matches this input against verb wordforms, as listed in the morphological zones of verb entries. In case of a match the text string is chunked as VP and a corresponding supertag from the lexicon is assigned. The text strings which were left between NP and VP chunks are then looked up in the rest of the entries of the lexicon and tagged. The fact that in every chunking/tagging pass only the type-relevant lexicon entries are searched practically lifts the ambiguity problem in morphological analysis.

Finally, based on classified chunk borders, the document is turned into an interactive ("clickable") text with NP and VP phrases highlighted in different colours.

The output of the shallow analysis stage (fully (super) tagged lexical units) is passed to the machine translation module that following simple rules generates SL-TL lexical pairs for all the lexica of the text (See Figure 1).

4 Status and Conclusions

The viability of the methodology we have described was proved by its implementation in a Russian-English CAT tool for the domain of scientific papers on mathematical modelling. The tool is fully developed. The domain bilingual static knowledge sources have been carefully crafted based on corpora analysis and internet resources. The programming shell of the tool is language independent and provides for knowledge administration in all the tool modules to improve their performance.

The extractor of Russian nominal terminology currently performs with 98, 4 % of recall and 96, 1% precision. The shallow clunker based on the extraction results and lexicon shows even higher accuracy. This is explained, on the one hand, by the high performance of the NP extractor, and, on the other hand, by the nature of inflecting languages. Rich morphology turns out to be an advantage in our approach. Great variety of morphological forms lowers ambiguity between NP components and verb paradigms.

We could not yet find any publications describing research meant for similar output. This leaves the comparison between other methodologies/tools and ours as a future work. In general user evaluation results show a reasonably small number of failures that are being improved by brushing up the bilingual lexicon.

We intend to a) improve the quality of the tool by updating the tool knowledge based on the user feedback; b) integrate the tool into a full MT system and c) develop a search facility on the basis of the our extraction strategy.

References

Enache Ramona, Cristina Espana-Bonet, Aarne Ranta, Lluıs Marquez. 2012. A Hybrid System for Patent Translation. *Proceedings of the EAMT Conference.* Trento..Italy, May

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

Kudashev Igor S. 2007. Desining Translation *Dictionaris of Special Lexica* /I.S.Kudashev. – Helsinki University Print, – 445 p.

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

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.