Incorporation of a Valency Lexicon into a TectoMT Pipeline

Natalia Klyueva and Vladislav Kuboň

Institute of Formal and Applied Linguistics
Charles University in Prague
E-mail: kljueva,vk@ufal.mff.cuni.cz

Abstract

In this paper, we focus on the incorporation of a valency lexicon into TectoMT system for Czech-Russian language pair. We demonstrate valency errors in MT output and describe how the introduction of a lexicon influenced the translation results. Though there was no impact on BLEU score, the manual inspection of concrete cases showed some improvement.

1 Introduction

The work on Machine Translation systems was traditionally presented as a collaboration between linguists and computer scientists: linguists prepared data (e.g. dictionaries and transfer rules), and computer scientists implemented the baseline of the system. Linguists analyzed the output translations and on the basis of these translations suggested further improvements and then the cycle was repeated.

The interplay between linguistics and computer science as described above was true for rule-based machine translation (RBMT) systems only before the data-driven (statistical, SMT). There was no longer a need for many linguists with the knowledge of the source and the target languages: all the necessary information was acquired from data, the evaluation was done either automatically or manually by native speakers rather than by experts in linguistics.

A big advantage of RBMT systems over SMT is that the former are more controllable and predictable. The errors produced by RBMT are easy to spot and it is often obvious how to fix them (but not always easy) – e.g. by some additional rules.

In this paper, we analyze the output of MT system between Czech and Russian with respect to valency errors. We make an experiment with a rule-based MT system implemented within a framework TectoMT (for other language pairs TectoMT involves statistical methods and modules as well, our implementation of Czech-to-Russian MT system can be considered primarily rule-based) and incorporate a list of surface valency frames into the translation pipeline.

The paper will be structured as follows. In Section 2 we describe theoretical and practical aspects of TectoMT and the implementation of Czech-to-Russian MT. Section 3 presents definition of valency and an overview of a valency resource used in the experiment. In Sections 4 and 5 we describe valency errors in MT output and incorporation of the valency lexicon into a translation pipeline. Manual evaluation of the proposed method is given in Section 6.

2 TectoMT

The TectoMT system between Czech and Russian was implemented within the framework **Treex** (Popel and Žabokrtský, 2009). Treex is a modular system of NLP tools, such as tokenizers, taggers and parsers that were created to process corpora and treebanks in multiple languages. One of the main projects under Treex is the English-Czech machine translation system (Popel,

Proceedings of the 2nd Deep Machine Translation Workshop (DMTW 2016), pages 47–53,
Lisbon, Portugal, 21 October 2016.

2010). Modules of the system are easily reusable for other languages, so in order to build a Czech-Russian MT, a Czech analysis module that was already present in the system was used and the Russian synthesis module was created.

Translation scenario in TectoMT consists of a sequence of **blocks** which ensure the transformation between the four language layers: word, morphological, analytical (shallow syntactic) and tectogrammatical (deep syntactic) layers, plus a scenario for a transfer phase[1]. This division has its roots in the Functional Grammar Description theory – FGD (Sgall et al., 1986), but its implementation in Treex is slightly different from original FGD concepts.

Below, we will provide a description of layers how they are used in Treex/TectoMT.

- **The Word and Morphological Layer.** A sentence represented as a sequence of tokens. On the morphological layer, each token in the sequence is represented as a word form with a lemma and a tag assigned.

- **The Analytical Layer.** Syntactic annotation is presented in the form of a dependency tree, where each morphologically annotated token from the previous level becomes a node with an assigned *analytical function* (**afun**). Analytical function reflects the syntactic relation between a parent and a child node (e.g. Subject, predicate etc.) and is stored as an attribute of the child.

- **The Tectogrammatical Layer**. The annotation on the tectogrammatical layer (t-layer) goes deeper towards the level of meaning. Function words (prepositions, auxiliary verbs etc.) are removed from the corresponding analytical tree; they are stored as attributes of autosemantic words, leaving only content words as the nodes on the t-layer.

Initially, the baseline system was established with a minimum number of rules handling the most obvious differences between the languages, such as copula drop or negation particle handling. The BLEU score (Papineni et al., 2002) of the baseline experiment was poor – 4.44%. We attribute such a low score mainly to the automatically generated dictionary, but also some errors might be introduced by the tagger, the parser and the module of word forms generation.

We included some blocks to fix certain linguistic phenomena: list of prepositions and verbs plus their formemes, copula drop, modal verbs, fixing year construction in Russian and some other minor fixes. The improvement of specific issues in terms of BLEU was very little, but the analysis always showed some improvement in an issue that we aimed for. After applying all those blocks and enlarging the dictionary, the BLEU score reached **9.38**%. We manually explored an MT output, and here we describe one concrete type of MT errors that concerns valency.

3 Valency

The notion of valency itself is not very straightforward and can be understood differently by different researchers. Traditionally, in general linguistics, it is used to indicate that the verb requires some number of complements of a certain semantic type. Here, we will refer to valency with respect to its surface realization – morphemic endings of nouns or preposition required by a verb. So, under the notion 'valency frames' we will understand mainly surface forms of frame elements. In this experiment we focus only on verbal valency though we admit that with nominal valency we face similar challenges as with verbal.

Czech and Russian are related languages, and on the first sight there are not many distinctions in valency. In order to prove this statistically, we computed the number of different valency frames in the two languages. To extract valency frames, we exploit the MT dictionary Ruslan that was created for the experiments in MT between Czech and Russian in 1980s, (Hajic, 1987), (Oliva., 1989).

In the dictionary, verbs were assigned with their valency frames in Czech and the corresponding frames in Russian with the specification of a semantic class of all verb complements. Example

[1]For the transfer we used an automatically extracted Czech-Russian dictionary.

1 demonstrates an entry from Ruslan dictionary for the verb *vystačit* – 'to be enough', the explanatory notes are given further:

(1) VYSTAC3==R(5,PRP,?(N(D),S(I,G)),39,CHVATIT6):

- *VYSTAC3* presents a stem of the verb *vystačit* – 'be enough',
- *R* denotes a root of a tree,
- *5* is a symbol for a verb and PRP is a conjugation pattern of the Czech verb,
- *N(D),S(I,G))* is a valency frame that we will further describe in detail,
- *39* is a Russian declination pattern,
- CHVATIT6 is the Russian translation of a lexeme, coded in Latin

We transformed the entry from the original Ruslan format: lowercased the entries, transferred Ruslan encoding of letters with diacritics (coded in numbers) into common letters, converted Latin into Cyrillic letters for a Russian word. Then we selected the verbs and substituted the verb stem and the morphological information coded in special symbols with an appropriate verb ending. Out of the 2080 verbal dictionary entries from Ruslan we have analyzed 1856 unique verbs. We have divided verbs on the basis whether a verb requires the prepositional case or the non-prepositional one.

Czech and Russian non-prepositional valency slots have usually identical cases, 68 verbs (3.6%) out of all the lexicon have some discrepancy in the frame (e.g., (cz)vyhýbat se + Dat -> (ru) избегать + Gen – *to avoid*). As for the prepositional cases, 104 (5.6 %) of verbs have different surface frames containing prepositions (e.g., (cz)doufat v + Acc -> (ru)надеяться на + Acc – *to believe in*).

The main result of this transformation is a small bilingual lexicon and that is included into a TectoMT translation scenario.

4 Valency errors in MT

We found valency errors to be crucial in MT output: verb and its complements form a core of a sentence, so the mistakes in the surface form of the complements can considerably lower the quality of a sentence. The reasons for errors in valency are twofold. The most evident case is when Russian and Czech valency have some discrepancies, and the Czech structure is used in a Russian output.

In the following example, a Czech verb 'to influence' is governed by a noun in the Accusative case, and the system translated a respective noun with the Accusative case as well. However, the surface realization of the argument is different in Russian – the Russian verb requires a prepositional phrase, so the two RBMT produced an error because neither had a rule covering this discrepancy:

(2)

(src)	*ovlivnit*	*výsledky*	*voleb*	
	influence	results-Acc.Pl	elections-Gen	
	'To influence results of the elections'			
(ref)	*повлиять*	***на***	*результаты*	*выборов*
	influence	**on**	results-Acc.Pl	elections-Gen
	'To influence results of the elections'			

(tmt) **повлиять* *результаты* *выборов*
influence results-Acc.Pl elections-Gen
*'To influence results elections'

RBMT errors in valency are only partially related to some discrepancy in Czech and Russian. The system also make errors in cases when the valency structure of a verb in Czech and Russian was similar, like in the example below:

(3)

(src) *demokratičtí* *kritici* *hovoří* *o* *předpojatosti* *zákonů*
democratic critics speak about prejudice-**Gen**.Sg laws-Gen.Pl
'democratic critics speak about prejudice of law'

(tmt) *Демократические* *критики* *говорят* *о* *предубеждение* *законов*
democratic critics speak about prejudice-**Acc**.Sg laws-Gen.Pl
*'democratic critics speak about prejudice of law'

Those errors might be the result of an improper analysis of the source phrase or some error in the transfer or generation phases. The errors that originate from the differences between the frames in the two languages can be fixed by introducing the valency lexicon.

5 Exploiting valency information from Ruslan dictionary in machine translation

We have exploited the entries from the Ruslan lexicon described above within the TectoMT system to see if there is some improvement in the translation. In order to integrate the dictionary into the system, we have transformed the entries into the special format verb+formeme [2]. Formemes (Dušek et al., 2012) are morphosyntactic properties of the node which were created especially for the TectoMT, they contain surface valency information:

(4) **narazit n:na+4 => столкнуться n:с+7** – *to run into smb*

The list of formemes was incorporated into a system in the form of a block – FixValency.pm.[3] We evaluated the performance on the WMT test set (3000 sentences). We measured the BLEU score and manually checked the differences in the two outputs - before and after the new block was introduced. After implementing this block, some sentences with troublemaking verbs (verbs with different surface valency) were translated with a proper surface form. In examples below, (1TMT) is a test translation before applying the rules and (2TMT) after applying the rules.

In the following example, a Czech verb *využívat* – 'use' governs a complement in the Dative case, and in the baseline (1TMT) system, the complement received the same formeme as a default. However, in Russian the Accusative case should be used instead. This discrepancy was covered by the Ruslan entry (využívat + Dat -> использовать + Acc)[4] in the improved system (2TMT).

(5)

(SRC) *využívali* *obrovských* *amerických* *zakázek*
used-3Pl huge-**Gen** american-**Gen** contracts-**Gen**
' they made use of huge American contracts'

[2] In formemes, and according to the Czech tradition, cases are indicated as numbers, e.g. 4 is Accusative, 7 is Instrumental.

[3] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/CS2RU/FixValency.pm`.

[4] využívat n:2 => использовать n:4 in the block FixValency.pm

(1TMT) **они использовали огромных американских заказов*
*they used huge-**Gen** american-**Gen** contracts-**Gen**
‘ they made use of huge American contracts’

(2TMT) *они использовали огромные американские заказы*
they used huge-**Acc** american-**Acc** contracts-**Acc**
‘ they made use of huge American contracts’

However, there were cases when this rule worsened the translation. In Example 6, the prepositional complement was translated properly by (1TMT) because a rule for the preposition transfer from another module[5] was applied (**n:pro+4 -> n:для+2** – n:for+Acc -> n:for+Gen). In the version with the lexicon, this rule was overridden by the rule from a new FixValency.pm module ("připravit n:pro+4" => "готовить n:про+4"). The latter verb-formeme Russian equivalent is a mistake in the Ruslan lexicon.[6]

(6)

(SRC) *v kuchyni se pro hosty připravuje čaj.*
in kitchen refl **for** guests-**Acc** prepare tea
‘In the kitchen the tea for the guests is preparing’

(1TMT) *В кухне для гостей готовится чай.*
in kitchen for guests prepare-refl tea
‘In the kitchen the tea **for** the guests-**Gen** is preparing’

(2TMT) **В кухне про гости готовится чай.*
*in kitchen for guests prepare-refl tea
‘In the kitchen the tea **about** the guests-**Acc** is preparing’

In some sentences, both translations were incorrect due to various reasons. In Example 7, the light verb phrase *nabývá účinnosti(Gen) vs. вступит в силу (в + Acc)* – ‘takes effect’ is different in Czech and Russian; it should have been translated with another verb and another noun. The rule has no effect in this case, as the translation is wrong all the same.

(7)

(SRC) *zákon nabývá účinnosti 6 prosince*
law gains effect 6 December
‘The law takes effect on 6 December’

(1TMT) *закон приобретать эффективности 6 декабря*
law *gains *effect-**Gen** 6 December
‘The law gains effect on 6 December’

(2TMT) *закон *приобретать *эффективность 6 декабря*
law *gains *effect-**Acc** 6 December
‘The law takes effect on December 6’

The above examples show that using the valency resource helps in some cases and harms in some others. Also, there was no significant influence on the BLEU score: **9.40%** without and **9.37%** with the module FixValency.pm.

[5] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/CS2RU/RuleBasedFormemes.pm`

[6] As the dictionary was compiled by non-native Russian speakers, there are a few errors in the lexicon and this one illustrates how people automatically assign a surface frame from their native Czech language to the verb in Russian.

6 Manual evaluation

For such a small experiment, the BLEU score can not necessarily indicate if this valency module helped or not – we evaluated the experiment only on the reference translations with one example. We evaluated manually the cases where a valency frame was changed according to the lexicon. It was impossible to count the exact number of cases when the pattern from the lexicon was applied because mostly the valency pattern stayed the same.

We have marked a list of changes between the (1TMT) and (2TMT) outputs indicating whether the introduction of a new rule:

- lead to some improvement like in Example 5
- worsened the translation like in Example 6
- did not have any effect as both variants were incorrect – Example 7

Effect	number of differences	Percentage
improved	28	58.3 %
worsened	3	6.2%
no effect	17	35.4%
Total	48	100%

Table 1: Manual evaluation of changes after adding FixValency.pm

From the table we can see that in the majority of cases the verbal valency is improved, or it has no effect on the translation which is wrong this way or that. However, such a little fix did not bring any sufficient gain or loss when considering the automatic evaluation metric BLEU.

7 Conclusion

We described the experiment with introducing valency information into an MT system between Czech and Russian. The BLEU score showed no improvement, but the manual evaluation revealed the cases where the valency errors were fixed.

Our initial assumption that errors in valency would occur only when there is some discrepancy in Czech and Russian valency structures turned out to be false. Many words were marked as a valency error even though the Czech and Russian verbs had the same frame with the same morphological cases. This may be due to the low performance of analysis or synthesis modules of the system, when the wrong case/preposition can be used even if the valency patterns for Czech and Russian are identical. So it is crucial to improve other 'core' modules of the system to insure the proper integration of the lexicon into the translation pipeline.

8 Acknowledgment

This work has been supported by the grants FP7-ICT-2013-10-610516 (QTLeap) and LINDAT/CLARIN project No. LM2015071 of the Ministry of Education, Youth and Sports of the Czech Republic.

References

Ondřej Dušek, Zdeněk Žabokrtský, Martin Popel, Martin Majliš, Michal Novák, and David Mareček. 2012. Formemes in English-Czech Deep Syntactic MT. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 267–274, Montréal, Canada. Association for Computational Linguistics.

Jan Hajic. 1987. RUSLAN: An MT System Between Closely Related Languages. In *Proceedings of the Third Conference on European Chapter of the Association for Computational Linguistics*, EACL '87, pages 113–117, Stroudsburg, PA, USA. Association for Computational Linguistics.

Karel Oliva. 1989. A parser for czech implemented in systems q. Explizite Beschreibung der Sprache und automatische Textbearbeitung, MFF UK Prague, .

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. pages 311–318.

Martin Popel and Zdeněk Žabokrtský. 2009. Improving English-Czech Tectogrammatical MT. *The Prague Bulletin of Mathematical Linguistics*, (92):1–20.

Martin Popel. 2010. English-Czech Machine Translation Using TectoMT. In Jana Šafránková and Jiří Pavlů, editors, *WDS 2010 Proceedings of Contributed Papers*, pages 88–93, Praha, Czechia. Univerzita Karlova v Praze, Matfyzpress, Charles University.

P. Sgall, E. Hajicová, J. Panevová, and J. Mey. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Springer.