Verbal Valency in the MT Between Related Languages

Natalia Klyueva
Inst. of Formal and Applied Linguistics
Charles University in Prague
kljueva@ufal.mff.cuni.cz

Vladislav Kuboň
Inst. of Formal and Applied Linguistics
Charles University in Prague
vk@ufal.mff.cuni.cz

Abstract

The paper analyzes the differences in verbal valency frames between two related Slavic languages, Czech and Russian, with regard to their role in a machine translation system. The valency differences are a frequent source of translation errors. The results presented in the paper show that the number of substantially different valency frames is relatively low and that a bilingual valency dictionary containing only the differing valency frames can be used in an MT system in order to achieve a high precision of the translation of verbal valency.

1 Introduction

Numerous experiments, such as Česílko (Hajič et al., 2000) and Apertium (Sánchez et al., 2007), with the machine translation (MT) between related languages support the claim that direct (word for word or phrase for phrase) methods guarantee better translation quality than complicated MT architectures. The more related the source and target languages are, the better the results provided by simple direct methods. Very closely related languages have similar morphological and syntactic properties, their lexicon usually also demonstrates a great number of similarities not only with regard to the lexical values, but also to important phenomena as e.g. the valency. For the translation of those languages it is therefore possible to ignore valency completely, because the system can rely on the similarity (or even identity) of valency frames of corresponding words and thus it is possible to translate expressions from individual valency slots directly, as e.g. in the Czech-to-Slovak MT system Česílko.

The languages which belong to the same language group, but which are not as closely related constitute a greater challenge, they require a different treatment of the verbal valency. In subsequent sections of this paper we present an examination of differences between Czech (a western Slavic language) and its Eastern Slavic counterpart, Russian.

Experiments in automatic extraction of verbal valency frames from different resources were carried out by many researchers. One of the first attempts was made in early 90's by (Rosen et al., 1992) where the process of English verb frame derivation from a learner's dictionary is described. The similar goal for extracting verb frames for both Czech and English was set in a research by (Bojar et al., 1984). Valency frames were extracted automatically from a parallel treebank PCEDT, resulting in a list of verbs and their modifications.
To the best of our knowledge such experiments were not carried out for related languages.

2 Existing resources

Manually built and handchecked dictionaries of verbal valency frames exist both for Czech and Russian. Vallex (Žabokrtský et al., 2007) is a lexicon of Czech valency frames having its roots in FGD (Functional Generative Description) theory. For Russian language, verbal valency frames can be found in the TKS (Tolkovo-Kombinatornyj Slovar - Explanatory combinatorial dictionary) – cf. (Mel'čuk, 1984). The lexicon of TKS is based on a Meaning-Text theory, it contains rich syntactic and semantic information for lexical entries of all parts of speech.
The formalisms on which Vallex and TKS are based are different in many ways, therefore it is almost impossible to map the entries from both dictionaries directly.

The first attempt to achieve a high quality MT between Czech and Russian, the transfer-based system Ruslan, was carried out in 80's (Oliva, 1992). This project left a valuable resource in a

form of a bilingual dictionary that includes various kinds of information necessary for lexical, morphological, syntactic and semantic transfer. In our current work we use only morphological and syntactic information from this dictionary.

Another system we work with is an MT system between closely related languages Česílko (Hajič et al., 2000), which uses a direct word-for-word (and tag-for-tag) translation. Initially the system translated between Czech and Slovak languages reaching rather high quality, as the two languages are very closely related. When other languages from Slavic group - Polish, Lithuanian and Russian - were included into the system, it became evident that some additional shallow syntactic rules must be used.

3 Valency

Valency frame of a verb contains syntactic and semantic information crucial for proper analysis and synthesis of a sentence. In our work we will use a notion of a valency frame at the level of shallow syntax, we will not take into consideration deep syntactic structure. So we avoid such terms as Actor, Patient, Recipient etc., and we use rather surface forms of the verbal actants - cases: Nom, Gen, Dat, Acc, Ins, Loc for which we use shortcuts n, g, d, a, i, l respectively. Our work is carried out on the two Slavic languages, Czech and Russian, and for the sake of simplicity we partly follow the representation of verb structure used in the MT system Ruslan. In addition we use the following terms added for the present experiments (Czech case is always listed first, followed by a Russian one enclosed by brackets):

Simple frame constituents:

n(n) means that Czech nominative case corresponds to the same case in Russian.

a(d) means that whereas accusative form is used in Czech, Russian uses dative case.

Frame constituents including prepositions:

s(i,s(i)) means that the Czech preposition *s* (with) requires an instrumental case in Czech and the same situation holds for Russian language.

Other constituents:

(inf(inf)) means that both languages use infinite form of an additional verb as a valency constituent.

A translation valency frame therefore consists of a set of simple and/or prepositional or other constituents for both Czech and Russian. Example:

trvat|**(n(n),na(l,na(l)))**|nastaivat' - to insist

3.1 Dictionary of Ruslan

Dictionary entries in Ruslan contain morphological, syntactic and semantic information. In the first stage of our study we do not make use of semantic features, leaving it for future experiments.

The dictionary has 10023 entries, 2080 of which are verbs. Let us now present two examples of original dictionary entries from Ruslan, one for a noun and one for a verb:

NA2PAD==H(@(*A),FI1023, IDEJA) - *idea*.

H represents a nominal declension class(hrad).

DOBE3H==R(5,TI,?(N(N)),D2,KONC2IT6SJA) - *to finish running*

R represents a verb, 5,TI - conjugation type (tisknout), (N(N)) - the valency frame of an intransitive verb with a single slot for a subject in nominative case in both languages.

,D2,KONC2IT6SJA - conjugation class + Russian lexical equivalent of the verb.

4 Classification of valency frames

Out of the 2080 verbal dictionary entries from Ruslan we have analyzed 1856 unique verbs. The reason of this difference is the fact that the original dictionary contains a number of verbal pairs with identical valency frames, usually two variants of a Czech lemma in the present and past tense. We made a classification of how the Czech valency frames correspond to the Russian ones. We have sorted verbs on the basis whether the verb requires the prepositional case or the simple one. The most important categories of verbs are those showing differences between both languages - these verbs will serve as a basis of a list of verbs with different valency frames which will be used for an improvement of our experimental MT system. The subsequent subsections describe examples for all analyzed categories of words.

4.1 Equal simple frame constituents

Cases when a verb have an actant structure without a preposition and Czech and Russian frames correspond to one another:

vyzývat|(n(n),a(a) or n(n))|vyzyvat' - to call

The most typical sequence of frame patterns is **n(n),a(a)** , which represents simple transitive verbs. 1317 (70 % of all verbs) have this structure. The fact that Czech and Russian have practically the same number of cases that are meaningful [1]

[1] Vocative case is not used in modern Russian unlike in Czech

Table 1: Case correspondences

Cs/Ru	Nom	Gen	Dat	Acc	Ins
Nom	3070	8	10	6	3
Gen	0	25	0	4	0
Dat	0	3	178	7	0
Acc	3	19	12	1388	7
Ins	5	0	0	3	1355

when speaking of verb valency makes the comparison easier and it apparently also influences the number of identical frames.

4.2 Different simple frame constituents

The first group of verbs that will form our list of verbs having different valency frames in both languages are those translation pairs in which Czech and Russian verbs govern different simple cases:
vyžadovat|(n(n),**a(g)** or n(g),i(i))|trebovat'
- to demand, Acc in Czech, Gen in Russian:
povšimnout|(n(n),refl(si),g(a))|zametit'
- to notice, Gen in Czech, Acc in Russian
rušit|(n(n),**a(d)** or n(d),i(n))|mešat'
- to disturb, Acc in Czech, Dat in Russian
hýbat|(n(n),a(a),**i(a)**)|dvigat'
- to move, Ins in Czech, Acc in Russian

Table 1 presents the statistics of simple frame patterns giving a picture of how simple cases in Czech and Russian mutually correspond. [2]

As we can see from the table, Czech and Russian non-prepositional valency slots have usually identical cases, the list of verbs exhibiting differences is very short.

4.3 Equal prepositional frame constituents

Verbs in this class have the valency slots containing prepositions. We have considered the translation frames to be equal in a case when prepositions are translated straightforwardly or typically from Czech into Russian. The problem is to set a border between typically translated prepositions and those translated differently. This issue lies outside of the scope of our study. We have used the data from (Nadykta, 2007), in which the author addresses in detail many aspects of Czech and Russian prepositions. Following are verbs and frames that constitute a typical translation of each other according to our criteria:
do(g,v(a)):ponořit|(n(n),do(g,v(a)))|pogruzit' - to sink into
z(g,iz(g)):vycházet|(n(n),z(g,iz(g)))|vychodit' - to go out from

4.4 Different prepositional frame constituents

To select verbs that have different prepositional frames we just excluded verbs with similar frame patterns described in the previous section. 104 (5.6 %) of verbs belong to this group. Below are some examples of such verbs:
záležet|(n(n),**na(l,ot(g))**)|zaviset' - to depend on
narazit|(n(n),**na(a,s(i))**)|stolknut'sja - to face
We also define some special cases which are irrelevant from computational point of view, as they will be processed as the common cases. They may still be of some interest to theoretical study of verb valency differences.
Those special cases form a rather small group of verbs that:
1. they are followed by an infinitive:
přestat|(n(n) or inf(inf) or **v(l,inf)**)|perestat' - to stop + inf
2. they govern identical prepositions that have different case:
klást|...**před(a,pered(i))** or na(a,na(a))| klast' - to put behind
3. they govern a preposition in one language, while in the other a simple case is used:
vystačit|(n(d),**s(i,g)**)|chvatit' - to be enough

5 Statistics of Valency Difference List

The main output of our work is a list of verbs that have different valency structure in Czech and Russian. Table 2 shows the statistics of those verbs with regard to our classification on simple and prepositional case frames.

Table 2: Types of valency frames incorrespondences

Type of difference	N of verbs	Percentage
Simple case	68	3.6%
Prepositional case	104	5.6%
Total	1856	100%

[2] Locative case is not included as it is governed by a preposition in both languages.

6 Evaluation

In this section we present a semi-manual evaluation of our list of verbs carried out on sentences translated by the Česílko MT System. In the process of MT evaluation we have evaluated only parts of sentences that include a verb and its arguments and we have determined whether our data might improve the translation. The test did not evaluated overall translation quality due to the observation that because of the overall imperfection of the system there are many other errors that have greater influence on the translation quality and which would bias the evaluation of our experiment. We aim primarily at an estimation to which extent the knowledge of differing valency frames ultimately might improve the translation quality by its own, not in combination with other phenomena. We are actually aiming at a kind of upper boundary of possible improvement.

The evaluation was carried out on a relatively small sample of 100 sentences translated from Czech into Russian.

As mentioned above, we have evaluated not the whole sentences, but smaller units. In accordance with (Lopatková et al., 2009), we took linguistically motivated units (segments) containing only one finite verb. This made it easier to analyze valency issues of concrete verbs. This approach was motivated by the fact that in complex sentences it might be difficult to define a verb and its arguments when a clause is divided into two or more parts by an embedded segment, and a verb is situated in another part of a sentence than its dependent arguments:

*Mnozí provozovatelé **považovali** naši shůzku, k níž došlo bezprostředně po konferenci v Anapolisu, kde se sešli představitelé všech arabských států včetně Sýrie a Izraele, **za projev** nevůle...*

(Many observers considered our meeting which took place immediatelly after the conference in Anapolis, where the deputies of all Arabic states including Syria and Israel met, to be a manifestation of ill will...)

In the evaluated phrase the verb *považovat* and its dependent prepositional construction *za projev* stands more than 20 tokens from one another, and could not be analyzed properly without breaking a sentence into several less complex segments.

The evaluation process was performed in several steps:

Table 3: Errors in verbal valency

mistakes	34	12,45 %
improvements	16	5,86 %
Total No. of verbs	273	100 %

1. Detect segments of sentences with Czech verbs with different valency structure

2. Determine whether the verbs and their arguments have been translated into Russian by the MT system in a correct way

2b. ...and whether or not adding our Valency DATA can improve the translation quality (Sometimes even this will not help because of the totally different structure)

The table 3 describes the results of the evaluation: the **mistakes** column presents a number of incorrectly translated verbal valency constructions, the **improvements** column shows the number of cases where our valency list could have helped to achieve better results.

The table shows that errors in verbal valency occur in slightly more than 10 % of all verbs. Almost half of those mistakes can be captured by our list of valency differencies that contains most frequent verbs. Here comes an example of an error in MT, that can be improved:

*pokračovat v diplomatických snahách.*LOC(cz) (continue diplomatic attempts)
**prodolzhat' v diplomatičeskich popytkach.*LOC(ru - Česílko MT)(v + loc)
*prodolzhat' diplomatičeskie popytki.*ACC(ru - improved)

The verb *pokračovat* - to continue - in Czech has as its arguments the preposition *v* + noun in locative case, the entry from our data (pokračovat (v(l,a)) prodolzhat') will make sure that a noun in accusative case will follow the verb in Russian.

7 Conclusion

In this article we have shown that the number of really different verbal valency frames between Czech and Russian is relatively low and that instead of using a complete bilingual valency dictionary it is reasonable to create only a list of differences and to translate the remaining verbs and their constituents in a default manner. We have also evaluated the expected impact our data will have on translation of verbs and their arguments. This evaluation shows that although the valency

dictionary will definitely improve the translation quality, it's influence is relatively small and it will be necessary to investigate also other phenomena in order to achieve a more substantial improvement.

Nevertheless, this experiment has also brought interesting results from the linguistic point of view. It shows that in the future it might be possible to translate both existing valency dictionaries for Czech and Russian and compare them. This might bring an enrichment of the frames contained in both dictionaries. The extension of our list of differences will then come as a side effect of this process.

Acknowledgments

This research is supported by grants MSM0021620838 and GA405/08/0681.

References

Hajič, J., Hric, J., and Kuboň, V. 2000 *Machine translation of very close languages*. In Proceedings of the 6th Applied Natural Language Processing Conference

Sánchez-Martínez, Felipe and Forcada, Mikel L. 2007 *Automatic induction of shallow-transfer rules for open-source machine translation*. In Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation 181-190

Alexandr Rosen, Eva Hajicová, Jan Hajic. 1992 *Derivation Of Underlying Valency Frames From A Learner's Dictionary.*, COLING 1992, 553-559

Lopatková, M., Klyueva, N., Homola, P. 2009 *Annotation of Sentence Structure; Capturing the Relationship among Clauses in Czech Sentences.* In Proceedings of the Third Linguistic Annotation Workshop, Law III, ACL-IJCNLP 2009 Conference 74-81

Karel Oliva. 1989 *A Parser for Czech Implemented in Systems Q.* in Explizite Beschreibung der Sprache und automatische Textbearbeitung, MFF UK Prague, 1989.

Igor Mel'čuk and Alexander Zholkovsky. 1984. *Explanatory Combinatorial Dictionary of Modern Russian.* Vienna: Wiener Slawistischer Almanach

Ondřej Bojar and Jan Hajič. 2005. *Extracting Translation Verb Frames.* Proceedings of Modern Approaches in Translation Technologies, workshop in conjunction with Recent Advances in 2-6.

Zdeněk Žabokrtský and Markéta Lopatková. 2007. *Valency Information in VALLEX 2.0: Logical Structure of the Lexicon*. Prague Bulletin of Mathematical Linguistics, (87):41-60, 2007.

Nadzeya Nadykta. 2007. *The Use of Prepositions in Czech and Russian according to Parallel Corpus Data*. Diploma, Praha