Incorporation of a Valency Lexicon into a TectoMT Pipeline

Natalia Klyueva and Vladislav Kubon

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague

kljueva,vk@ufal.mff.cuni.cz

October 21, 2016
1 **TectoMT for Czech-Russian**
   - TectoMT scenario for Czech-Russian
   - System Improvement

2 **Valency**
   - Defining valency
   - Valency in Slavic Languages
   - Ruslan dictionary - format transfer
   - Discrepancies in valency frames

3 **Experiment: implementation of the dictionary into TectoMT**
   - Ruslan frames transformed into formemes
   - Any improvement?

4 **Conclusion**
- Czech analysis module
- Czech-Russian transfer: Czech-Russian dictionary; formemes (+ surface valency frames)
- Russian synthesis: specific blocks for Russian like copula drop
System improvement

- Fixing verbal aspect.
- Enlarging the dictionary.
- The list of formemes with prepositional complements:
  - Some blocks to fix certain linguistic phenomena were added/improved: copula drop, modal verbs, fixing year construction in Russian etc.
- Surface valency frames added as formemes
## System improvement

| Experiment and improvements                          | BLEU score |
|-----------------------------------------------------|------------|
| Baseline                                            | 4.44%      |
| Fixing verb tenses and aspect                       | 5.09%      |
| Adding preposition formemes                         | 6.62%      |
| Larger dictionary                                   | 7.04%      |
| Fixes in Czech analysis (punctuation)               | 9.04%      |
| Fixes in rules                                      | 9.40%      |
| Fixes in valency                                    | 9.37\(^1\)% |

**Table:** Baseline and improvements

\(^1\)Trust BLEU score under 20?
Valency

- Valency: Capability of a word to bind arguments
- Deep valency vs. surface valency
- Surface valency frame - formeme
- noun, verbal valency
The verbs that can cause mistranslations:

CZ: účastnit se konference (lit. to participate conference.Gen)
RU: участвовать в конференции (to participate in conference.Loc)
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

Transformed into vystačit s + Ins - chvatit’ + Gen
Discrepancies in valency frames

- The same simple valency frame: Nom вyzывать + Acc -> Nom вызывать + Acc – to call + Acc
- The same prepositional complements: пусобить на + Acc -> воздействовать на + Acc to influence on + Acc
- different simple frame: (cz)vyhýbat se + Dat -> (ru) избегать + Gen – to avoid
- different prepositional frame: (cz)doufat v + Acc -> (ru) надеяться на + Acc – to believe in)
Discrepancies in cases

|        | Czech |        |        |        |        |
|--------|-------|--------|--------|--------|--------|
|        | 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      |
| Inst   | 5     | 0      | 0      | 3      | 1355   |

Different surface frames: 90 (1.47%)
Total number of surface frames: 6160 (100%)
Number of verbs with different frames: 68 (3.66%)
Total number of analyzed verbs: 1856 (100%)

Table: Co-occurrence of the same cases in Czech and Russian based on Ruslan dictionary
Discrepancies in prepositional phrases

| Czech frame | Russian frame | freq |
|-------------|---------------|------|
| na+Acc      | na+Acc        | 82   |
| do+Gen      | v+Acc         | 80   |
| z+Gen       | iz+Gen        | 76   |
| k+Dat       | k+Dat         | 58   |
| s+Ins       | s+Ins         | 57   |
| od+Gen      | ot+Gen        | 29   |
| v+Loc       | v+Loc         | 26   |
| o+Loc       | o+Loc         | 22   |
| do+Gen      | do+Gen        | 19   |
| k+Dat       | dlja+Gen      | 16   |
| na+Acc      | o+Loc         | 15   |
| na+Acc      | k+Dat         | 14   |
| před+Ins    | ot+Gen        | 12   |
| o+Acc       | na+Acc        | 10   |

**Table:** Prepositional case correspondence – Ruslan dictionary
Discrepancies: overall statistics

Valency frames correspondences

Equal prepositional frames: 90.80%
Different prepositional frames: 3.60%
Different simple frames: 5.60%
"vztahovat n:k+3" => "относить n:k+3",
"vystačit n:s+7" => "хватить n:2",
"vztáhnout n:4" => "отнести n:4",
"vznášet n:4" => "задавать n:4",
"vžít se n:do+2" => "вжиться n:v+4",
"vzdálit se n:od+2" => "удалиться n:от+2",
"vyzývat n:4" => "вызывать n:4",
"vyzvat n:4" => "вызвать n:4",
"vyznačovat n:4" => "обозначать n:4"
Any improvement?

- BLEU: 9.40% - > 9.38%
- manual evaluation of 100 sentences:

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

Table: Manual evaluation of changes after adding FixValency.pm
exploiting old language resources
challenges for morphologically rich languages
valency lexicon: no impact in BLEU, but the manual evaluation showed some improvement
Thank You!