briefly on CPA/PDEV and Czech Translational Equivalence

Jan Popelka

Basic idea: search for relations between PDEV patterns and their translational equivalents

term project in the scope of Lexical Analysis course

brought to the topic and led by Dr. Holub

quite a current topic

Patrick Hanks working on CPA/PDEV

automatic pattern recognition addressed in Lenka's thesis
Hypothesis

Knowledge of PDEV pattern for a given English verb occurrence possibly makes the choice of a Czech translational equivalent easier, i.e. decreases the number of possible translations, thus could help in the course of machine translation.

Goals

either falsify the hypothesis or discover some characteristic relations between patterns and translational equivalents
Data required:

English-Czech parallel sentences, each English verb occurrence of interest annotated with PDEV pattern and matched with corresponding Czech equivalent.

Data sources used:

PEDT & PCEDT corpora
- WSJ articles and their Czech translations
- manual praguian PDT-like annotation up to t-layer
Verb Selection (1) - Criteria

the task

10-20 verbs (out of 615 finalised PDEV verbs)
500-1000 occurrences total

verb selection criteria

pattern rich verbs, higher pattern perplexity
rich set of possible Czech translations
enough occurrences in PEDT (at least 50)

occurrence = t-node with the given lemma and lexically corresponding a-node tagged as verb
only 19 verbs matching PEDT occurrence criterium alone, criteria reconsidered

   at least 25 occurrences
   PDEV characteristics not taken into account

33 verbs to choose from

inaccurate statistics extracted by means of TectoMT and automatic t-alignment

first observations

   non-verbal translations, both verbal aspects, ”synonymity”,
   (obvious alignment errors)
verb selection (3) - results

abandon, acknowledge, admit, anticipate, argue, call, claim, deny, execute, fire, handle, launch, lead, say, signal, tell, treat, urge

18 verbs manually chosen on the basis of their translational richness (by agreement of JP & MH)
1075 occurrences total
100 random occurrences of frequent verbs
all occurrences of less frequent verbs
manually annotated czech t-tree requirement lifted in order to slightly increase occurrence count
Pattern annotation

credits to Patrick Hanks and Silvie Cinková

pattern exploitation classified either as unprecise match or figurative use

exclusion of occurrences erroneously tagged as verbs, with undecideable pattern

Inter-annotator agreement for patterns

dataset annotated twice independently, far from 100% agreement; both datasets however leading to the same conclusions
| abandon | ???|*brát v potaz | V|v-w202f13|T-wsj0118-001-p1s72a6 |
| abandon | dokončit|dokončit_:W | V|v-w598f1|T-wsj1146-001-p1s82a5 |
| abandon | odmítnout|odmítnout_:W | V|v-w2785f1|T-wsj2130-001-p1s2a20 |
| abandon | odvrátit_se|odvrátit_:W | V|v-w2975f1|T-wsj0456-001-p1s25a6 |
| abandon | odvrhnout|odvrhnout_:W | V|v-w10418f2|T-wsj0114-001-p1s14a24 |
| abandon | opustit|opustit_:W | V|v-w3161f1|T-wsj0118-001-p1s13a22 |
| abandon | opuštěný|opuštěný_^(5stit) | A||T-wsj1685-001-p1s5a5 |
| abandon | opuštěný|opuštěný_^(5stit) | A||T-wsj2136-001-p1s7a21 |
| abandon | ukončení|ukončení_^(3it) | N|v-w7115f1|T-wsj2427-001-p1s2a20 |
| abandon | ukončit|ukončit_:W | V|v-w7116f1|T-wsj0146-001-p1s7a11 |
| abandon | upustit|upustit_:W | V|v-w10600f2|T-wsj0101-001-p1s15a17 |
| abandon | vzdát_se|vzdát | V|v-w8641f1|T-wsj0456-001-p1s10a19 |
| abandon | vzdát_se|vzdát | V|v-w8641f1|T-wsj1474-001-p1s26a1 |
| abandon | vzdát|vzdát | V|v-w8640f1|T-wsj1410-001-p1s25a24 |
Translational equivalent annotation

two complex attributes, including:

- Czech translation as both m-lemma and t-lemma
- basic part of speech, valency frame (where applicable)

not always trivial, most frequent peculiarities:

- no equivalent can be found in the translated sentence
- many-to-many and one-to-many t-node relations
- preposition as significant or the only part of translation
- verb as part of an idiomatic expression
- opposite use of negation or passive voice (equivalents?)
(not) tell (the truth) - lhát

tell – s pokyny

lead – v čele s / jít v čele

say - podle

abandon - (ne)brát v potaz

treat (harshly) – (přísný) trest

fire – dát výpověď

treat – zažít přístup

signal – dát na srozuměnou

all told – celkově vzato

all told – se vším všudy

say – se slovy

lead – mít největší

fire (back) – opětovat (palbu)

Notes:

TrEd filelist avaible

Verbal Vallex Frames
Mapping (en-cz) ... running project of Jana Šindlerová and Ondřej Bojar
peculiarity handling, approx. 5% of occurrences

lemmatised string of surface word forms if needed
(not a single node, crucial use of a preposition)

technical string for all non-identifiable translations

pattern exploitation handling

non-exact pattern match merged with regular use

same sense assumed for all figurative uses
of a pattern, forming a distinct new pattern
classification criteria needed

which translations to be considered distinct, which equal, given the annotated attributes

three procedures proposed

"degrouping" ... valency frame used where available, full translation attribute string elsewhere

"as is" ... always using full translation attribute only

"grouping" ... verbs differing only in aspect grouped manually

no other grouping (synonyms, deverbative...)
grouping by verbal aspect

<table>
<thead>
<tr>
<th>Verbal Aspect</th>
<th>Verb</th>
<th>Corpus ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>požádat</td>
<td>v-w4227f1</td>
</tr>
<tr>
<td>T</td>
<td>požadovat</td>
<td>v-w4230f1</td>
</tr>
<tr>
<td>W</td>
<td>prezentovat</td>
<td>v-w4277f2</td>
</tr>
<tr>
<td>W</td>
<td>proběhnout</td>
<td>v-w4286f1</td>
</tr>
<tr>
<td>W</td>
<td>prohlásit</td>
<td>v-w4354f1</td>
</tr>
<tr>
<td>T</td>
<td>prohlašovat</td>
<td>v-w4357f1</td>
</tr>
<tr>
<td>T</td>
<td>prohlašovat</td>
<td>v-w4357f2</td>
</tr>
</tbody>
</table>
## Data Analysis (1) - Example

### Data example

| verb      | N  | c(T) | H(T)  | G(T)  | H(T|P) | G(T|P) | c(P) | H(P)  | G(P)  | H(P|T) | G(P|T) | MI   | 2 MI* | MI/H(T) |
|-----------|----|------|-------|-------|-------|-------|------|-------|-------|-------|-------|------|-------|---------|
| call      | 97 | 31   | 4.41  | 21.25 | 2.16  | 4.47  | 12   | 2.43  | 5.4   | 0.18  | 1.13  | 2.25 | 4.76  | 0.51    |
| admit     | 48 | 15   | 3.48  | 11.14 | 1.69  | 3.22  | 9    | 2.52  | 5.73  | 0.73  | 1.66  | 1.79 | 3.46  | 0.51    |
| lead      | 72 | 22   | 3.06  | 8.32  | 1.63  | 3.1   | 11   | 2.58  | 5.99  | 1.16  | 2.23  | 1.43 | 2.69  | 0.47    |
| abandon   | 34 | 20   | 4.04  | 16.4  | 2.64  | 6.22  | 7    | 1.96  | 3.88  | 0.56  | 1.47  | 1.4  | 2.63  | 0.35    |
| deny      | 62 | 13   | 2.71  | 6.54  | 1.49  | 2.8   | 8    | 2.59  | 6.01  | 1.36  | 2.57  | 1.22 | 2.33  | 0.45    |
| fire      | 26 | 11   | 2.92  | 7.57  | 1.78  | 3.43  | 5    | 1.22  | 2.33  | 0.08  | 1.06  | 1.15 | 2.21  | 0.39    |
| claim     | 71 | 24   | 2.64  | 6.25  | 1.55  | 2.94  | 6    | 1.57  | 2.97  | 0.48  | 1.39  | 1.09 | 2.13  | 0.41    |
| handle    | 56 | 32   | 4.5   | 22.61 | 3.49  | 11.26 | 4    | 1.16  | 2.24  | 0.16  | 1.12  | 1.01 | 2.01  | 0.22    |
| treat     | 31 | 20   | 3.84  | 14.32 | 2.86  | 7.26  | 2    | 0.98  | 1.97  | 0     | 1     | 0.98 | 1.97  | 0.26    |
| signal    | 37 | 13   | 2.8   | 6.98  | 1.86  | 3.62  | 4    | 1.47  | 2.77  | 0.52  | 1.44  | 0.95 | 1.93  | 0.34    |
| execute   | 32 | 20   | 4.11  | 17.31 | 3.23  | 9.38  | 3    | 0.95  | 1.93  | 0.06  | 1.04  | 0.88 | 1.85  | 0.22    |
| tell      | 97 | 24   | 3.08  | 8.44  | 2.25  | 4.75  | 10   | 1.33  | 2.51  | 0.5   | 1.41  | 0.83 | 1.78  | 0.27    |
| launch    | 65 | 26   | 4.13  | 17.48 | 3.51  | 11.4  | 3    | 1.09  | 2.13  | 0.47  | 1.39  | 0.62 | 1.53  | 0.15    |
| urge      | 42 | 17   | 3.3   | 9.86  | 2.72  | 6.59  | 3    | 0.95  | 1.93  | 0.37  | 1.29  | 0.58 | 1.5   | 0.18    |
| anticipate| 41 | 12   | 2.72  | 6.58  | 2.2   | 4.59  | 3    | 1.13  | 2.19  | 0.61  | 1.53  | 0.52 | 1.43  | 0.19    |
| argue     | 92 | 18   | 2.04  | 4.11  | 1.6   | 3.04  | 4    | 0.49  | 1.41  | 0.06  | 1.04  | 0.44 | 1.35  | 0.21    |
| acknowledge| 34 | 9    | 2.82  | 7.05  | 2.57  | 5.95  | 2    | 0.67  | 1.59  | 0.43  | 1.35  | 0.25 | 1.19  | 0.09    |
| say       | 98 | 11   | 2.74  | 6.69  | 2.74  | 6.69  | 1    | 0     | 1     | 0     | 1     | 0    | 1     | 0       |
random variables: $P$.. patterns, $T$.. translations

statistics calculated for each verb separately:

$N$ ... number of occurrences used

$c(T)$, $c(P)$ ... numbers of distinct occurrences

$H(T)$, $H(P)$ ... entropies

$H(T|P)$, $H(P|T)$ ... conditional entropies

$MI(P,T)$ ... mutual information

$MI(P,T) = H(P) - H(P|T) = H(T) - H(T|P)$

$G(T)$, $G(P)$, $G(T|P)$, $G(P|T)$

... perplexities, $(2^\text{concerning entropy})$
Motivation: even if the number of distinct translations is the same, the ease of guessing the translation might not be the same:

4 occurrences, 2 translations evenly distributed:
A A B B ... \( H(T) = 1, \ G(T) = 2 \)

4 occurrences, 2 translations
A A A B ... \( H(T) < 1, \ G(T) < 2 \)

(easier to guess as A is more likely)
## Data Analysis (3) - Correlations

Correlation matrix calculated

(Pearson's correlation coefficients)

|       | c(T)    | H(T)    | H(T|P)   | c(P)    | H(P)    | H(P|T)   | MI      | MI/H(T)  |
|-------|---------|---------|---------|---------|---------|---------|---------|----------|
| c(T)  | 1.000   | 0.6661  | 0.3433  | 0.403   | 0.2566  | -0.1349 | 0.4476  | 0.2421   |
| H(T)  | 0.6661  | 1.000   | 0.6896  | 0.1349  | 0.2193  | -0.2236 | 0.458   | 0.1041   |
| H(T|P)  | 0.3433  | 0.6896  | 1.000   | -0.5161 | -0.4802 | -0.4789 | -0.3279 | -0.6348  |
| c(P)  | 0.403   | 0.1349  | -0.5161 | 1.000   | 0.8505  | 0.5199  | 0.8096  | 0.825    |
| H(P)  | 0.2566  | 0.2193  | -0.4802 | 0.8505  | 1.000   | 0.7207  | 0.8756  | 0.912    |
| H(P|T)  | -0.1349 | -0.2236 | -0.4789 | 0.5199  | 0.7207  | 1.000   | 0.2961  | 0.4901   |
| MI    | 0.4476  | 0.458   | -0.3279 | 0.8096  | 0.8756  | 0.2961  | 1.000   | 0.9151   |
| MI/H(T)| 0.2421 | 0.1041  | -0.6348 | 0.825   | 0.912   | 0.4901  | 0.9151  | 1.000    |
Grouping and Mutual Information

during incremental translation grouping mutual information either decreases or doesn't change
degrouping thus makes more sense

minimalistic example: (no change, decrease)

\[
\begin{array}{cc}
\{TA, TB\} \rightarrow TX & a & \{TC, TD\} \rightarrow TY \\
TA & P1 & TX & P1 \\
TB & P1 & TX & P1 \\
TC & P2 & TY & P2 \\
TD & P2 & TY & P2 \\
\end{array}
\]

\[
\begin{array}{cc}
{TA, TC} \rightarrow TX & a & {TB, TD} \rightarrow TY \\
TA & P1 & TX & P1 \\
TB & P1 & TY & P1 \\
TC & P2 & TX & P2 \\
TD & P2 & TY & P2 \\
\end{array}
\]

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\text{před} & H(P) & H(P|T) & H(T) & H(T|P) & MI \\
\hline
\text{po} & 1 & 0 & 2 & 1 & 1 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\text{před} & H(P) & H(P|T) & H(T) & H(T|P) & MI \\
\hline
\text{po} & 1 & 0 & 1 & 0 & 1 \\
\hline
\end{array}
\]
Empirical Findings (1)

pattern perplexity and mutual information strongly correlated

\[ H(P) \text{ and } MI(P,T) \]
\[ H(P) \text{ and } MI(P,T)/H(T) \text{, too} \]

reasoning: the more bits of information are contained in the knowledge of an actual pattern (i.e. the richer and more evenly distributed the pattern description of a given verb is), the more this knowledge decreases the uncertainty about possible translations
Empirical Findings (2)

\[ \text{MI}(P, T) \approx 2 \text{ for the "best" verbs} \]

actual pattern knowledge reduces the set of possible translations to one quarter of its size (in terms of perplexity)

\[ \frac{\text{MI}(P, T)}{\text{H}(T)} \approx 0.5 \text{ for the "best" verbs} \]

actual pattern knowledge reduces the translation uncertainty to half its value, thus quadratically reducing in size the set of possible translations

both this quantities strongly correlated with pattern perplexity, as shown before
Empirical Findings (3)

degrouping on valency frame strengthens slightly the important correlations

reasoning: Vallex dictionary of Czech verbs shares some common aspects with PDEV, pattern knowledge seems not only to help narrow the selection of Czech verb lemmas for the translational equivalent but also their valency frames
Conclusions

Failed to falsify our hypothesis, i.e. CPA/PDEV could possibly improve MT.

Greater payoff can definitely be expected from verbs with rich pattern description in PDEV, as the mutual information was found highly correlated with pattern entropy and perplexity.

Hence, verbs having a rich set of patterns should be of primary concern for eventual pilot experiments.
Observation Confidence

Are our observations skewed because of the low number of occurrences of certain verbs?

experiment: plot correlation for increasing subsets of the 18 verbs sorted descendingly by number of occurrences used