Extracting Translation Verb Frames

Ondřej Bojar, Jan Hajič
{bojar,hajic}@ufal.mff.cuni.cz
Presented by Petr Homola

September 24, 2005

Outline

- Motivation: Why are translation verb frames needed
- Related research: Extracting verb subcategorization frames
- Our method: observe frames, select nice examples, optionally filter
- Evaluation
- Conclusion and further research
- Neglected MT divergences

Motivation

- Structural MT relies on accurate syntactic dictionaries.

  Need to support an older MT system RUSLAN (Hajič [1987]) with a Czech-to-English dictionary.

  More on the project of reusing an older MT system in Bojar et al. [2005].

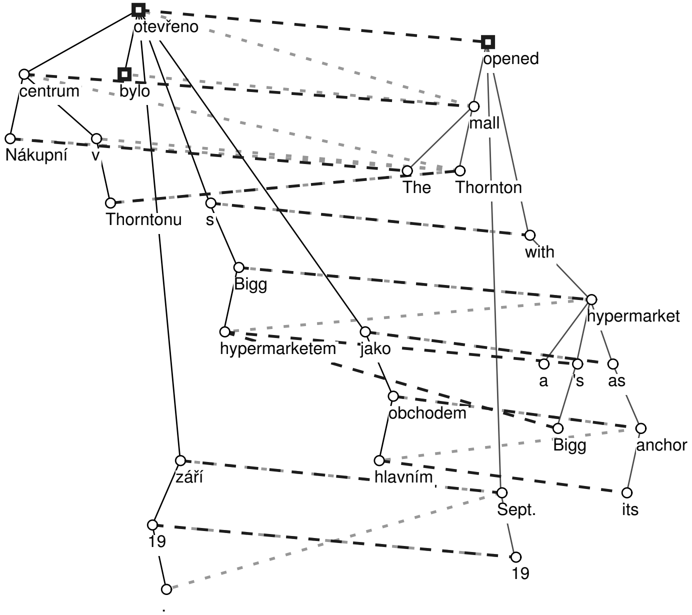
- For Czech and English, there is no MT dictionary.

  Available machine-readable dictionaries, such as Svoboda [2001] or WinGED, do not contain required syntactic information either at all or in an inaccessible form.

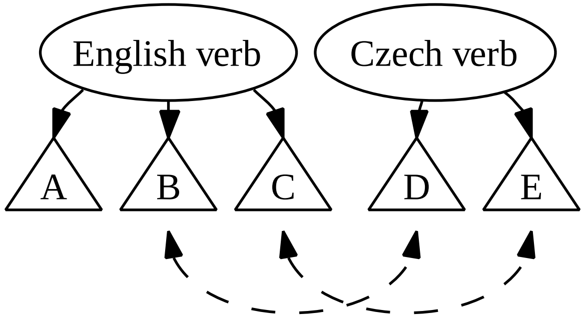
  Other researchers (Čmejrek et al. [2003]) had to use a limited dictionary with single word translations only.

The Goal

Given: Two (surface) syntactic trees + GIZA alignments in both directions
We need: Surface translation frames, such as *dělit=divide na+accusative=into*



⇒



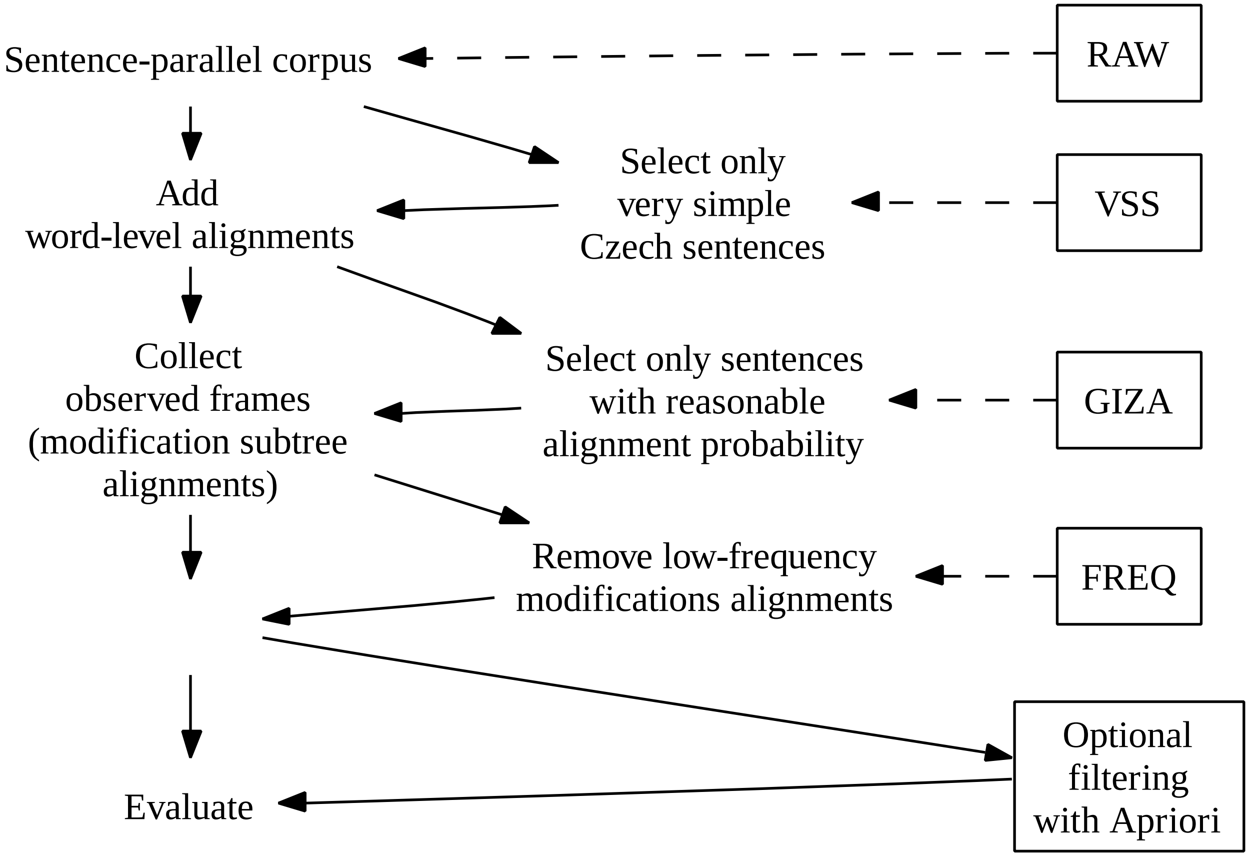
Related: Subcategorization Acquisition

A lot of work done, for a review see Korhonen [2002] or Zeman and Sarkar [2000].

But:

- They aim at telling whether a verb's modification is a COMPLEMENT or an ADJUNCT.
- We aim at telling the type (i.e. surface form) of an English modification given the type (surface form) of the corresponding Czech modification.

Our Pipeline with Various Filtering Methods



Sentence-parallel corpus
RAW
Select only very simple Czech sentences
VSS
Add word-level alignments
Select only sentences with reasonable alignment probability
GIZA
Collect observed frames (modification subtree alignments)
Remove low-frequency modifications alignments
FREQ
Optional filtering with Apriori
Evaluate

Training Data

Prague Czech-English Dependency Treebank 1.0 (PCEDT, Čmejrek et al. [2004]):

- Sentence-based translation of 21600 sentences from Wall Street Journal section of Penn Treebank 3.
- Syntactic trees for English derived from manual PTB annotation.
- (Surface) syntactic trees for Czech built automatically using an adaptation of Charniak's parser (Charniak [2000]).

. . . but the CD offers more, such as deep syntactic (tectogrammatical) trees.

Word Alignment

PCEDT is aligned only at the sentence level.

GIZA++ (Och and Ney [2003]):

- Unsupervised learning of translation models IBM1-5 and HMM.
- As a side-effect produces word alignments: one-to-many correspondence of a word in a source language sentence to words in the target language sentence.
- Employed in both directions; usually the intersection of alignments is quite reliable.
- Never evaluated alone on Czech-English (no hand-aligned data ready)

  . . . but we are adding the alignments for PCEDT test set now

Collecting Observed Translation Frames

1. For every Czech verb occurrence, follow the GIZA link to find the corresponding English verb.

2. For every modification of the Czech verb find the corresponding modification, such that the overall number of GIZA links between the whole modification subtrees in both directions is the highest.

3. The same procedure for those English modifications that have not been paired yet.

4. Mark modifications with no GIZA link as linked to NULL.

Optional Filtering of Obtained Frames

Apriori (Agrawal et al. [1993])[1]:

- Extracts common subsets from a set of sets.
- Given $\{a, b, c\}, \{a, b\}$
  and a certain threshold on the requested support of the subsets
  outputs $\{a\}, \{b\}, \{a, b\}$

The effect of Apriori on observed frames is twofold:

- Unreliable modifications (not supported by enough observations) are removed.
- Unseen frames are allowed, if they constitute subframes of known observations with support strong enough.

[1] An efficient implementation available at `http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html`

Evaluation Methods

The whole MT system is not ready yet
⇒ evaluate the method alone on manually paired surface frames.

Three different algorithms make use of the obtained translation frames:

- A: Translate slot by slot regardless the verb.

  All frames joined. Given a Czech modification form, the most probable English modification form is suggested.

- B: Translate slot by slot taking the verb into account.
- C: Translate according to the best matching frame.

  The English forms are chosen from the translation frame (of the given verb) with largest intersection of expected and observed Czech modification forms.

Simpler methods can serve as back-offs ⇒ CBA and BA.

Results

Test data: 140 sents., 400 occs. of 200 different verbs, 1005 modifications

Clean-up	Apriori	Method	F	P	R	
`giza`	Yes	A	68.4	52.9	96.7	← the best F-score
`giza`	Yes	CBA	66.4	50.5	96.7	
`giza`	Yes	BA	66.1	50.2	96.7	
`giza`	No	BA	66.1	49.8	98.0	
`raw`	No	A	58.2	41.3	98.9	← baseline
`vss`	Yes	BA	55.9	41.4	86.3	← the best of `vss`
`vss`	Yes	C	30.0	33.5	27.2	← the worst result

⇒ the simplest method trained on securely aligned frames works best

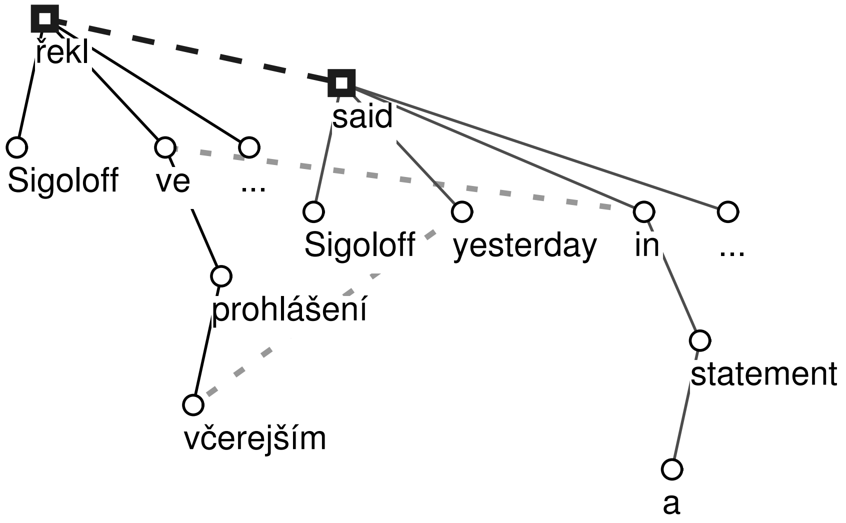
⇒ using simple Czech sentences hurts (hurts alignment ⇒ hurts extraction)

Conclusion and Further Research

- A method for extracting translation verb frames described.
- Evaluation suggest that the bottleneck lies in the quality of word alignment.
- Filtering with Apriori to remove unreliable modifications helps.
- Not really clear why the most simple evaluation method achieves best results.

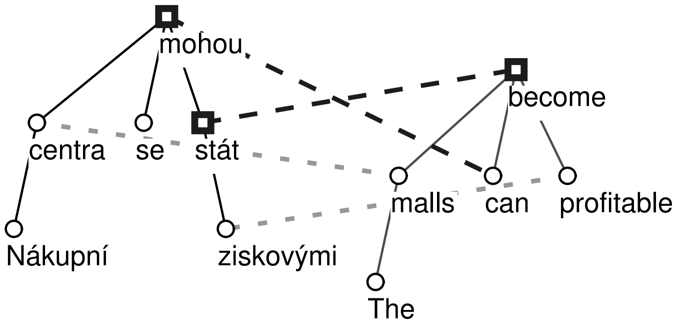
- Desirable to incorporate the obtained frames in an MT system.
- Necessary to model known syntactic divergences more adequately.

Neglected MT Divergences: Modification Shift



Sigoloff řekl ve včerejším prohlášení . . .
Sigoloff said yesterday in a statement . . .

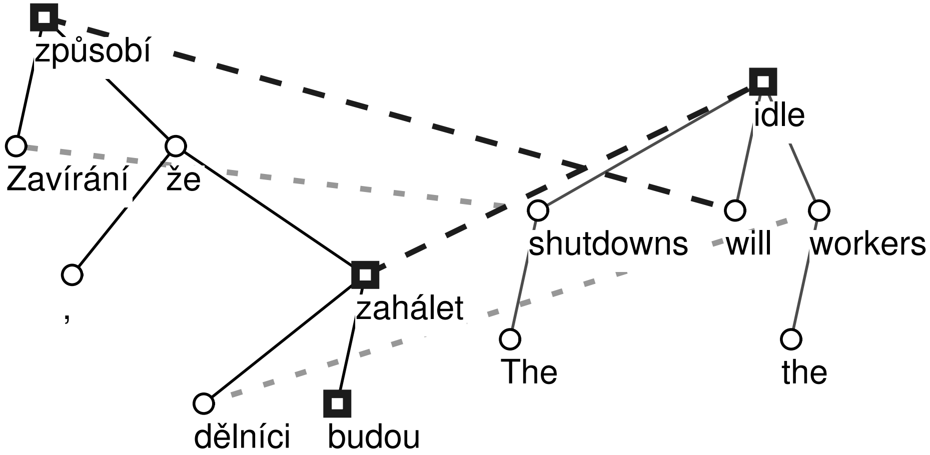
Neglected MT Divergences: Head Switching (1)



Nákupní centra se mohou stát ziskovými . . .
The malls can become profitable . . .

. . . caused by different dependency annotation guidelines.

Neglected MT Divergences: Head Switching (2)



Zavírání způsobí , že dělníci budou zahálet . . .
The shutdowns will idle the workers . . .

. . . real head switching (English verb transformed to a Czech subclause)

References

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, New York, NY, USA, 1993. ACM Press. ISBN 0-89791-592-5.

Ondřej Bojar, Petr Homola, and Vladislav Kuboň. An MT System Recycled. In *Proceedings of MT Summit X*, pages 380–387, September 2005. ISBN 974-7431-26-2.

Eugene Charniak. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL-2000*, pages 132–139, Seattle, Washington, USA, April 2000.

Martin Čmejrek, Jan Cuřín, and Jiří Havelka. Czech-English Dependency-based Machine Translation. In *EACL 2003 Proceedings of the Conference*, pages 83–90. Association for Computational Linguistics, April 2003. ISBN 1-932432-00-0.

Martin Čmejrek, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In *Proceedings of LREC 2004*, Lisbon, May 26–28 2004.

Jan Hajič. RUSLAN: an MT system between closely related languages. In *Proceedings of the third conference on European chapter of the Association for Computational Linguistics*, pages 113–117. Association for Computational Linguistics, 1987.

Anna Korhonen. Subcategorization Acquisition. Technical Report UCAM-CL-TR-530, University of Cambridge, Computer Laboratory, Cambridge, UK, February 2002.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, 2003. ISSN 0891-2017.

Milan Svoboda. GNU/FDL English-Czech Dictionary, 2001. `http://slovnik.zcu.cz/`.

Daniel Zeman and Anoop Sarkar. Learning Verb Subcategorization from Corpora: Counting Frame Subsets. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 2000. ELRA.