Dictionary-based Domain Adaptation of MT Systems without Retraining

Rudolf Rosa, Roman Sudarikov, Michal Novák, Martin Popel, Ondřej Bojar
{rosa,sudarikov,mnovak,popel,bojar}@ufal.mff.cuni.cz
Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (ÚFAL)

Motivation

We want to adapt an existing MT system to the target domain (IT in WMT16) using a domain-specific bilingual dictionary, but without retraining because

- **training costs (slow)**
- **running costs (if many target domains)**
- **dictionary not suitable as parallel data**
- **black-box MT (online, commercial)**
- **non-retrainable tools (tagger, URL tok...)**

Detection of entities (dictionary entries)

- **word-based trie text search**
- **heuristic scorer for matches overlap**
- **merging menu items separated by ">"**

MT systems used

- **Moses (tuned on in-domain)**
- **TectoMT (deepMT with TM interpol.)**
- **Chimera (Moses+TectoMT combination)**

Methods of forced translation

XXX placeholders

- **substitute entities with "xxxitemAxxx",...**
- **store their translation in a separate file**
- **applicable to any MT (not dropping OOV)**

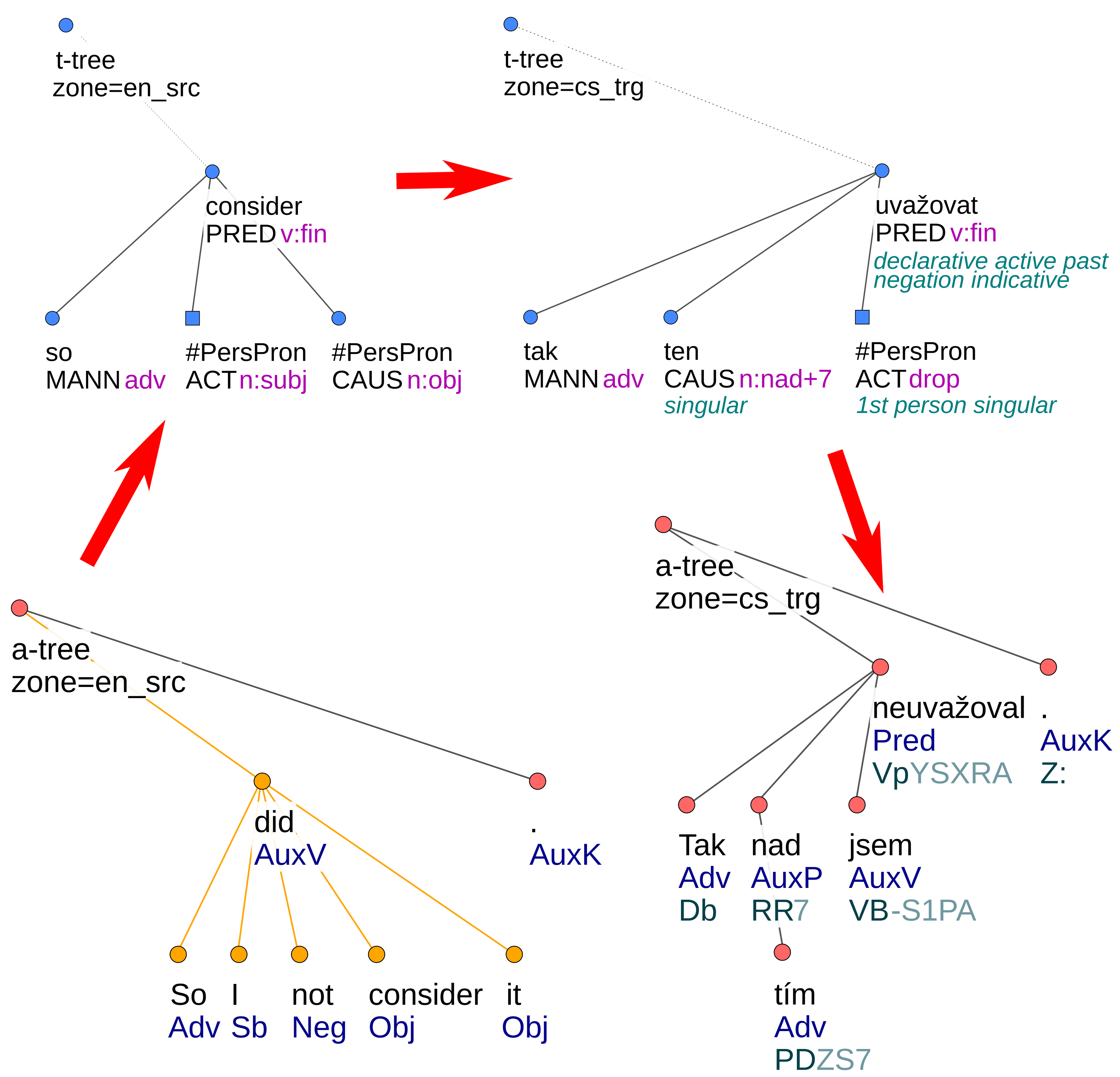
XML annotation

- **Moses with** `-xml-input=exclusive` **supports**
  select <item translation="Vypnout">Shut Down</item>.
- **better context for LM, but Moses specific**
- **incompatible with factored translation**

Wild attributes

- **TectoMT supports collapsing each entity into one node and storing its translation in a "wild" attribute.**

TectoMT pipeline



Results and conclusions

System	Annotations	→ES	→NL	→PT
Moses	(not adapted)	22.23	23.40	14.01
	XXX	23.61	24.89	15.47
	XML	**24.22**	**25.41**	**15.58**
Chimera	(not adapted)	26.01	21.82	13.11
	XXX	26.89	**23.52**	14.19
	XML	**27.40**	23.26	**14.21**

XML is better than XXX.
XXX is better than no adaptation.

System	Adapted	→CS	→ES	→NL	→PT
TectoMT	no	19.98	23.24	18.83	13.87
	wild	**21.89**	**24.31**	**19.89**	**15.51**
Moses	no	23.25	22.23	23.40	14.01
	xml	**23.71**	**24.22**	**25.41**	**15.58**
Chimera	no	**23.47**	26.01	21.82	13.11
	xml		**27.40**	**23.26**	**14.21**
	xxx	23.36			

Adaptation helps +1.3 BLEU on average.

Adaptation	→NL
(not adapted)	23.40
XML annotations	25.41
In-domain phrase table	**27.48**

The standard secondary in-domain phrase table in Moses gives the best results, but requires retraining.

System	Adapted	→CS	→ES	→NL	→PT
TectoMT	wild	3	**1–2**	3	**1**
Moses	only tune	4–5	3	4	3
Chimera	only tune	**1–2**			
	xxx	4–5	**1–2**	2	2
another		**1–2**		**1**	

WMT16 IT-task results: human ranking
Chimera combination helps (except PT).

Presented at WMT 2016, Berlin, Germany. Supported by the FP7 EU grant QTLeap (No. 610516), GAUK 1572314 and SVV 260 333. Using language resources hosted by the LINDAT/CLARIN Research Infrastructure, Project No. LM2015071.