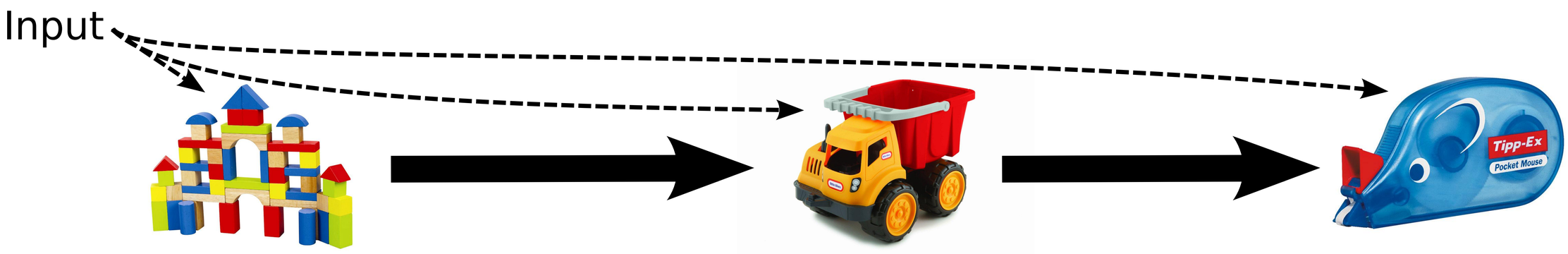
CUNI in WMT15: Chimera Strikes Again

Ondřej Bojar, Aleš Tamchyna {bojar,tamchyna}@ufal.mff.cuni.cz
Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics



TectoMT

- hybrid (rule-based/statistical) MT system
- transfer at a deep syntactic layer (t-layer)
- our combination: get an extra phrase table for Moses from TectoMT output

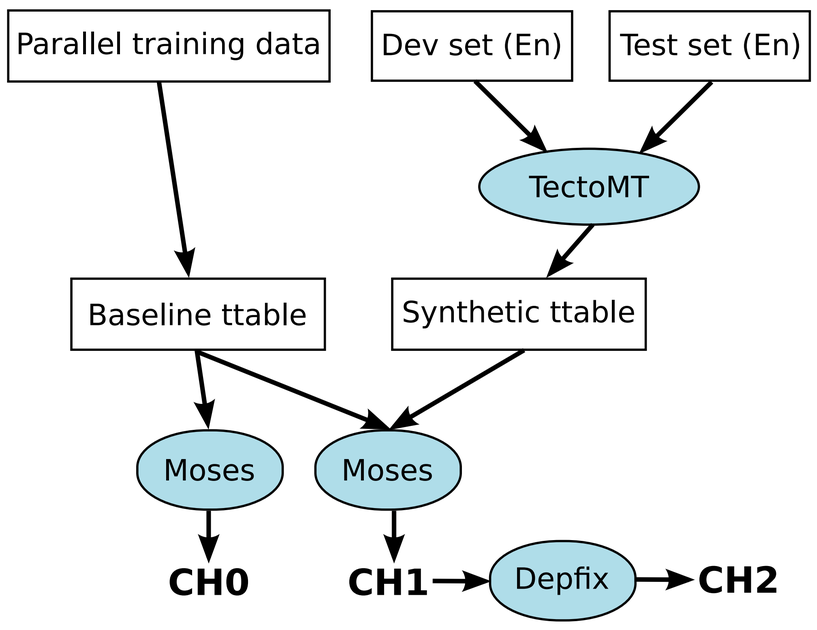
Moses

- phrase-based SMT
- large-scale data
- morphological tags as factors for a better grammatical coherence

Depfix

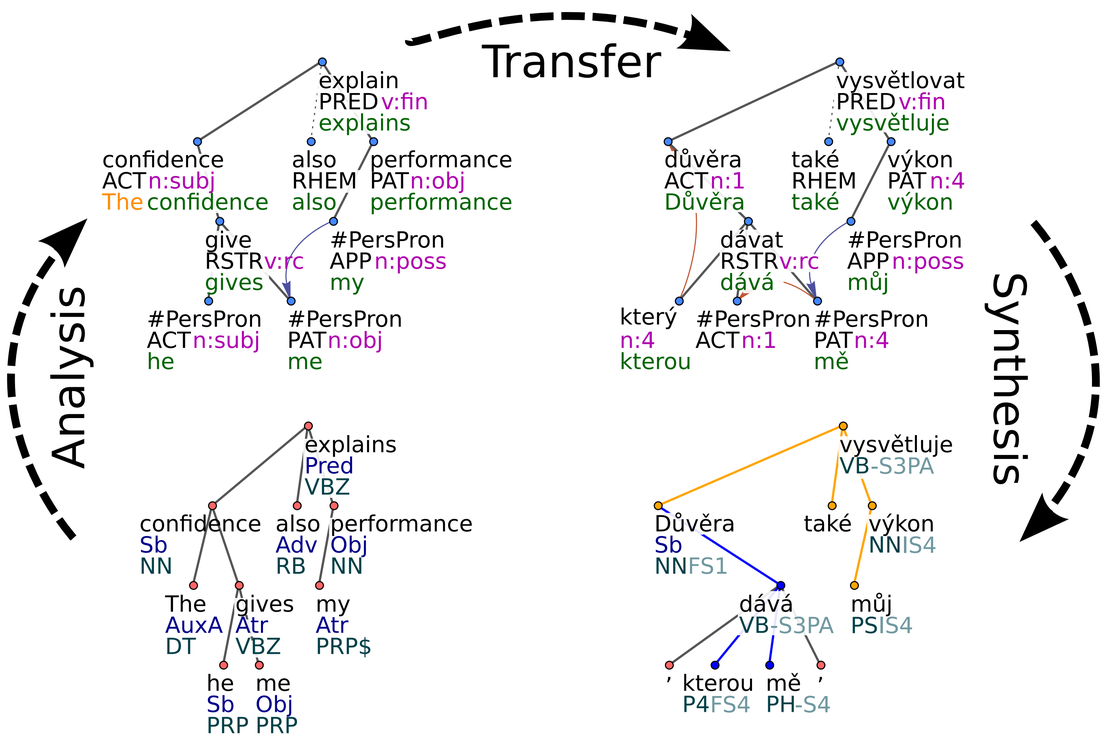
- rule-based error correction in (S)MT output
- output parse corrected based on source

Poor Man's Combination



- not restricted to tokens in 1-best outputs, can use alternative translations to better "glue" system outputs together

TectoMT Overview



- novel word forms (unseen in the parallel data)
- grammatical coherence, clause structure

Why TectoMT Helps

- TectoMT phrase table matches the test set ⇒ Moses can apply longer phrases
  - better grammatical coherence
  - search is simplified
- TectoMT provides many novel translations
- reduction of modelling errors

Constrained vs. Unconstrained

- monolingual: **44.3** vs. **392.3** million sentences
- parallel: **15.5** vs. **52.6** million sentence pairs

	Constrained	Full	Delta
CH0	21.28	22.59	1.31
CH1	23.37	24.24	0.87
Delta	2.09	1.65	

Gains from extra data

Gains from adding TectoMT

Language Models

long
- 7-gram LM on word forms
- mainly WMT monolingual data, individual years interpolated

big
- 4-gram LM on word forms
- use all available data

morph
- 10-gram LM on morphological tags

longmorph
- 15-gram LM on tags
- goal: capture sentential patterns

LMs	BLEU
long	21.32
long morph longmorph	22.00
big	22.00
long morph	22.01
long longmorph	22.14
big morph	22.21
big long	22.26
big morph longmorph	22.28
big longmorph	22.29
big long morph	22.48
big long longmorph	22.69
all	22.59

WMT Results

System	BLEU	TER	Manual
CH2	18.8	0.715	0.686
CH1	18.7	0.717	–
JHU-SMT	18.2	0.725	0.503
CH0	17.6	0.730	–
GOOGLE TRANSLATE	16.4	0.750	0.515
CU-TECTOMT	13.4	0.763	0.209

Chimera placed first among English→Czech MT systems in WMT for three years in a row.

Presented at WMT 2015, Lisbon, Portugal.

This research was supported by the grants H2020-ICT-2014-1-644402 (HimL), H2020-ICT-2014-1-644753 (KConnect), and SVV 260224. This work has been using language resources developed, stored and distributed by the LINDAT-CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).