Two-step translation with grammatical post-processing

David Mareček, Rudolf Rosa, Petra Galuščáková, Ondřej Bojar; {marecek, rosa, galuscakova, bojar}@ufal.mff.cuni.cz

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (ÚFAL)

Goal: Improve Czech grammar

- Phrase-based MT often breaks morphological agreement.
- Some of the dependencies still recovered in parses of MT output.
- => We can check and fix the agreements.

Depfix

- A rule-based grammar correction of MT to Czech.
- Applicable on top of any MT system.

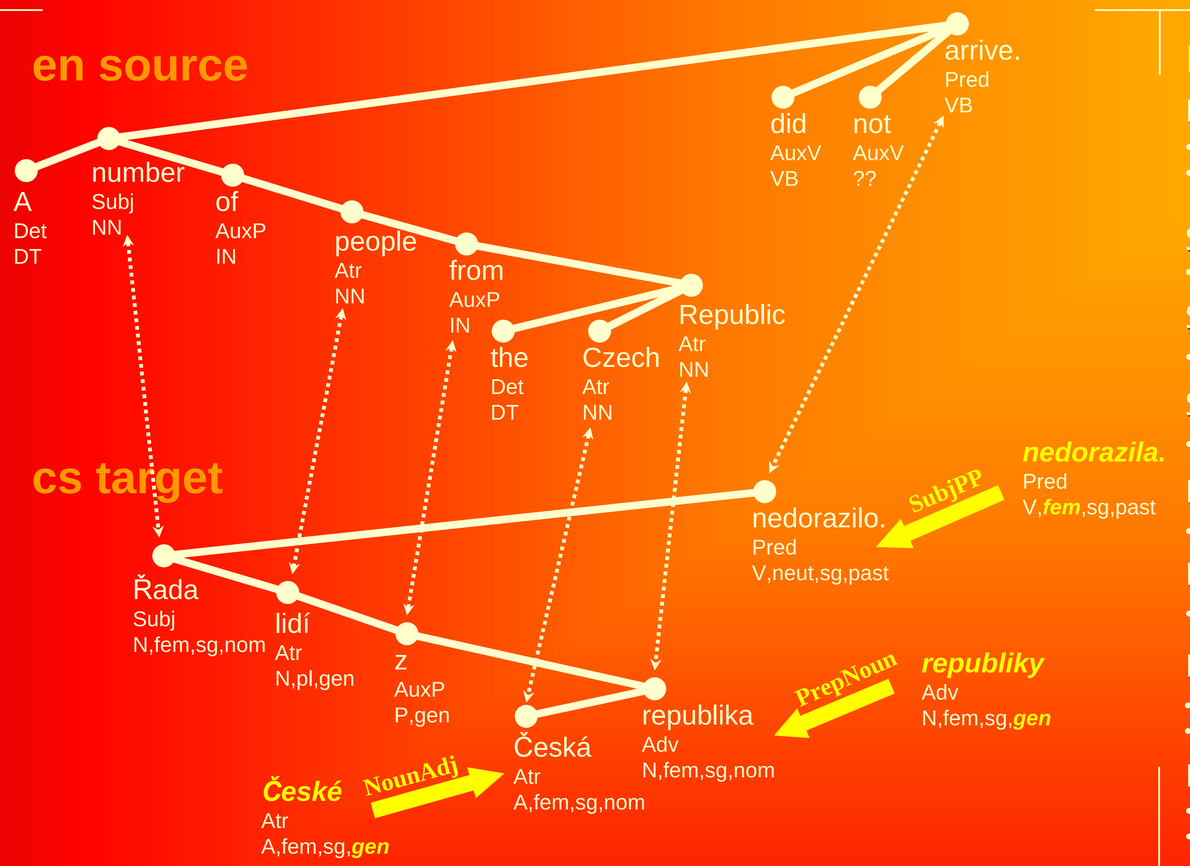
Our WMT11 submissions

cu-bojar

- Simple non-factored Moses setup.
- Truecasing based on lemmatizer output.
- Additional parallel data include Official Journal of EU.
- Large monolingual data:
  - Includes Czech National Corpus, our web collection...
  - Two LMs (5-gr and 6-gr) weighted by MERT.
  - Each LM already interpolated from domain-specific sources.

cu-marecek

- MT in three steps:
  1) Moses: English -> Lemmatized Czech
  2) Moses: Lemmatized Czech -> Czech
  3) Depfix: Rule-based grammar correction.
- Only 1-best output passed between the steps.
- The 2nd Moses trained on a large monolingual corpus.
- Lemmatized Czech ***includes*** morphological features over in English.



English source: A number of people from Czech Republic did not arrive.

Czech translated: Řada lidí z Česká republika nedorazilo. *(wrong)*

Czech after depfix: Řada lidí z České republiky nedorazila. *(correct)*

Rules

Noun number

- In Czech, plural has sometimes the same form as singular.
- Correct nouns tagged as singular to plural if the English counterpart is in plural.

Subject case

- Czech subject must be in nominative case.

Subject–Predicate agreement

- Subject and predicate must agree in morphological number.

Subject–PastParticiple agreement

- Czech subject and past participle must agree in number and gender.

Preposition–Noun agreement

- Nouns must agree with prepositions in morphological case.

Noun–Adjective agreement

- Adjectives must agree with nouns in number, gender, and case.

Reflexive particle deletion

- Czech reflexive verbs are accompanied by reflexive particles ('se', 'si').
- Particles not belonging to any verb are deleted.

Prepositions without children

- Prepositions cannot be leaves.
- Nouns are attached to them according to English source.

System	Before	After	Improvement
cmu-heaf.	16.95	17.04	0.09 ↑
cu-bojar	15.85	16.09	0.24 ↑
cu-zeman	12.33	12.55	0.22 ↑
dcu	13.36	13.59	0.23 ↑
dcu-combo	18.79	18.90	0.11 ↑
eurotrans	10.10	10.11	0.01 ↑
koc	11.74	11.91	0.17 ↑
koc-combo	16.60	16.86	0.26 ↑
onlineA	11.81	12.08	0.27 ↑
onlineB	16.57	16.79	0.22 ↑
potsdam	12.34	12.57	0.23 ↑
rwth-combo	17.54	17.79	0.25 ↑
sfu	11.43	11.83	0.40 ↑
uedin	15.91	16.19	0.28 ↑
upv-combo	17.51	17.73	0.22 ↑

WMT10

Depfix improvements in BLEU score

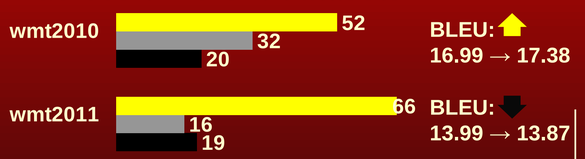
WMT11

System	Before	After	Improvement
cu-twostep	16.57	16.60	0.03 ↑
cmu-heaf.	20.24	20.32	0.08 ↑
commerc2	09.32	09.32	0.00 →
cu-bojar	16.88	16.85	-0.03 ↓
cu-popel	14.12	14.11	-0.01 ↓
cu-tamch.	16.32	16.28	-0.04 ↓
cu-zeman	14.61	14.80	0.19 ↑
jhu	17.36	17.42	0.06 ↑
online-B	20.26	20.31	0.05 ↑
udein	17.80	17.88	0.08 ↑
upv-prhlt.	20.68	20.69	0.01 ↑

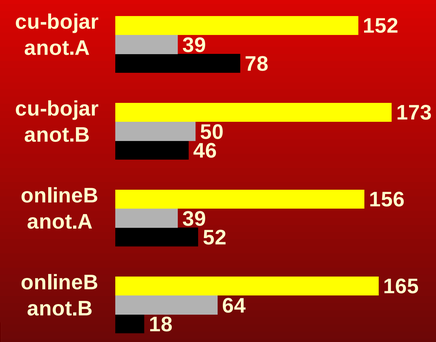
Interannotator agreement

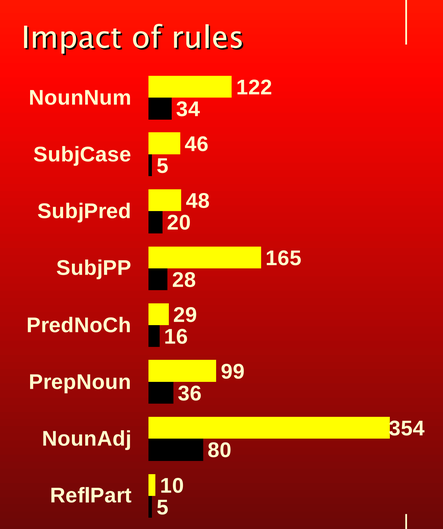
A/B	Impr.	Wors.	Indef.
Impr.	273	20	15
Wors.	12	59	7
Indef.	53	35	42

Manual and automatic evaluation across different datasets.



Manual evaluation





Full text, acknowledgement and the list of references in the proceedings of WMT 2011.