Hierarchical Phrase-Based MT at the Charles University for the WMT 2010 Shared Task

Daniel Zeman
Charles University in Prague, Institute of Formal and Applied Linguistics
Univerzita Karlova v Praze, Ústav formální a aplikované lingvistiky
Malostranské náměstí 25, CZ-11800 Praha
`zeman@ufal.mff.cuni.cz`

Direction	BLEU	WMT10	Cased	TER
en-cs	0.0905			
en-de	0.1114	0.1150	0.043	0.805
cs-en	0.1471	0.1420	0.105	0.773
de-en	0.1617	0.1610	0.114	0.760
en-es	0.1966	0.2110	0.165	0.687
en-fr	0.2001	0.1570	0.122	0.780
fr-en	0.2020	0.1890	0.137	0.723
es-en	0.2025	0.2170	0.161	0.675

Table 1: Baseline. Train: News Commentary (incl. trigram LM); align: 4char stems; dev: NewsTest 2008.
BLEU: test: NewsTest 2009. Lowercased BLEU score.
WMT10: test: NewsTest 2010. Official lowercased BLEU.
Cased: test: NewsTest 2010. Official cased BLEU.
TER: test: NewsTest 2010. Official cased BLEU.

primary submission

Setup	BLEU	WMT10	Cased	TER
Baseline	0.0905			
Small	0.1012	0.123	0.117	0.779
Large	0.1300	0.134	0.126	0.749

Table 2: English-to-Czech. Small and Large setups truecased, align on lemmatized Czech. Small trained on Czeng/News. Large trained on Czeng/all+EMEA. Mono: 210 Mword, 6gram.

Table 3: Small Setup system in the manual evaluation. % of times it was ranked equally or better than another system.

Rank	System	% win
0.	reference	95.85
1.	dcu-combo	75.17
7.	cu-bojar	65.57
8.	uedin	62.23
9.	pc-trans	62.06
10.	cu-tecto	60.13
11.	eurotrans	54.04
12.	cu-zeman	50.10
13. – 17.	5 other systems	45.04 – 33.01

primary submission

SO WHAT NEXT?

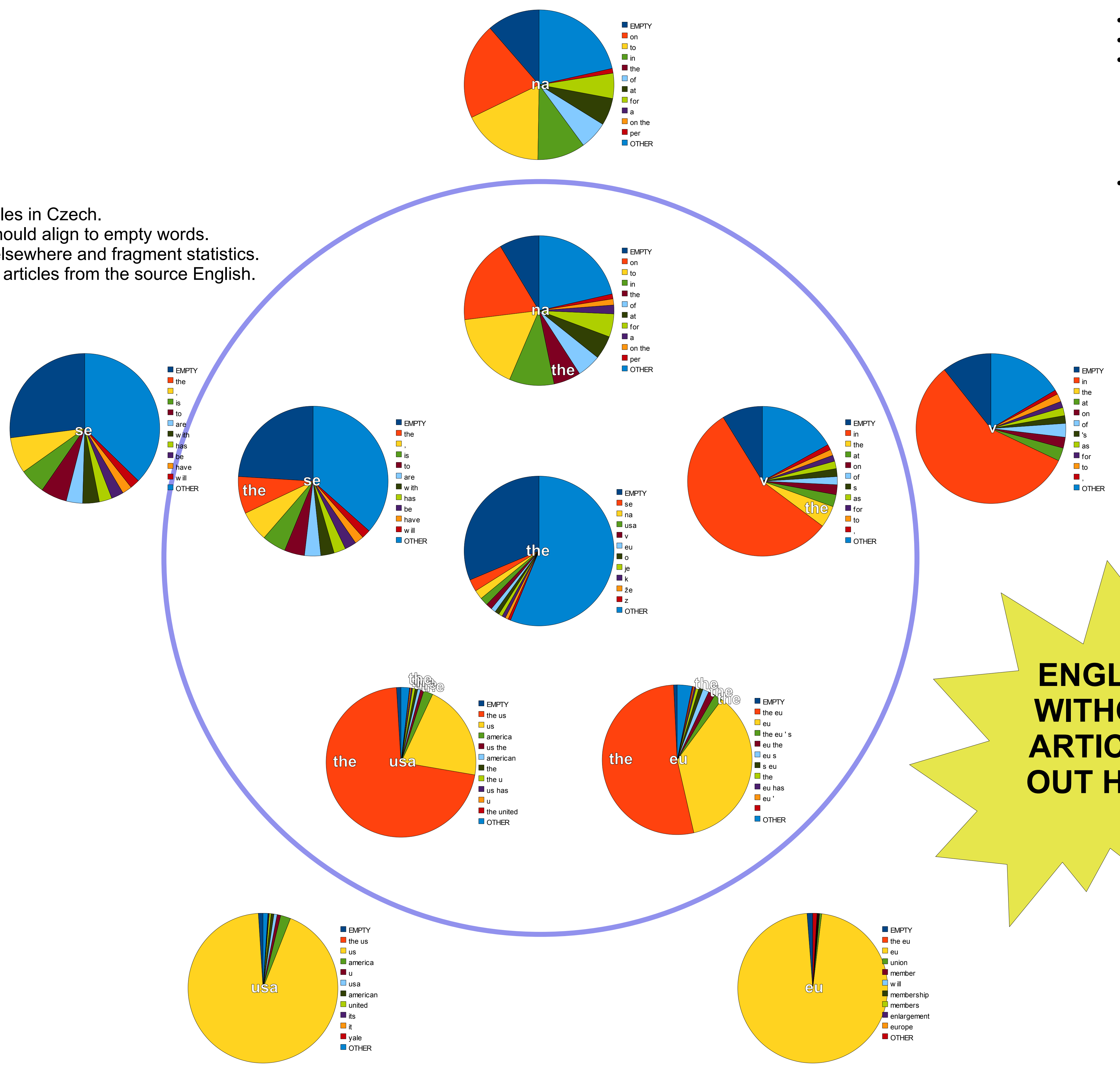
Preprocessing of the source-side text can make it look more similar to the target language in terms of word order etc. Learning may become easier then.

Subjects

- 7 grammatical cases in Czech.
- Subject should be in **nominative.**
- English subjects are marked by word order.
- Parse English, mark subjects explicitly such as in: *John/SB loves Mary.*
- Possible translations:
  - *John/SB → John*
  - *John → Johna, Johnu, Johnovi, Johne, Johnem*

Articles

- There are no articles in Czech.
- English articles should align to empty words.
- They often align elsewhere and fragment statistics.
- We could remove articles from the source English.



Personal pronouns

- Czech is a pro-drop language.
  - Subject can remain unexpressed.
  - Verb form indicates person and number.
    - As if there was a personal pronoun:
    - *(já) dělám = I do*
    - *(ty) děláš = you do*
    - *(on/ona/ono) dělá = he/she/it does*
    - *(my) děláme = we do*
    - *(vy) děláte = you do*
    - *(oni/ony/ona) dělají = they do*
- English pronouns align to Czech verbs.
- Join subjects expressed by personal pronouns with their verbs.
- Will it help?

Target agreement

- Both Czech nouns and adjectives take cases:
- *obchodní den, obchodního dne, obchodnímu dni* etc.
  (= trading day)
- Sparse data
  ⇒ phrase not seen in all cases
  ⇒ words will be translated separately.
- Ungrammatical **obchodním dne* could occur.
- Can we simulate factored translation of Moses?
- Morphological analysis of input:
  obchodního dne → obchodní GEN den GEN
- Morphological synthesis on output…

Verbal groups

- Variety of English analytical tenses, Czech is simpler.
- Phrase-based translation system can learn that:
  will make worse → zhorší
- Hierarchical system can learn it even with gaps:
  will only make matters worse → věci jen zhorší
- But it makes data sparser.
- Can we help?
- Preprocessing: whole verbal group at one place:
  - *will only make matters worse →*
    only will make worse matters
  - Question:
    - Word order of VP, *only* and *matters*?

- Another possible step:
  simplify tenses that do not exist in Czech.
  - *have since been arrested →*
    were arrested since

ENGLISH WITHOUT ARTICLES OUT HERE

Clauses

- Some English constructions do not exist in Czech.
- Restore subordinating conjunctions so they can be generated in Czech.

- Missing *if*
  - *should civil conflict erupt → if civil conflict should erupt*
- Missing *that*
  - *I like the car he bought → I like the car that he bought*

The research has been supported by the grant MSM0021620838.