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

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.
- Some English articles explicitly, as in: John contraseña.
- Possible translations:
  - John/John
  - John – Johna, Johnu, Johnovi, Johannes

Personal pronouns
- Czech is a pro-drop language.
- Subject can remain unexpressed.
- Grammatical case is not relevant.
- English articles should align to empty words.
- They often align elsewhere, but fragment statistics.
- We could remove articles from the source English.

Target agreement
- Both Czech nouns and adjectives take cases.
- *obchodní den, obchodní místo, obchodní místo etc.
  (= trading day)
- Sparse data.
  - words will be translated separately.

Verbal groups
- Variety of English analytical tenses, Czech is simple.
- Phrase-based translation system can learn that:
  - will make worse – zhorší
  - Hierarchical system can learn it even with gaps:
    - will only make matters worse – víko jen zhorší
    - If 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

SO WHAT NEXT?

SO WHAT NEXT?

ENGLISH WITHOUT ARTICLES OUT HERE


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

| Subject | System | % of times
|---------|--------|-------------
| 1       | maakt  | 99.34       |
| 2       | en-cs  | 99.57       |
| 3       | cased  | 99.33       |
| 4       | pc-trans | 99.26    |
| 5       | cited  | 99.14       |
| 6       | ued-in | 99.14       |
| 7       | sp-join | 99.16      |
| 8       | 30-14  | 99.16       |
| 9       | 5-10   | 99.16       |
| 10      | 11-15  | 99.16       |

WMT 2010, July 15 – 16, Uppsala

The research has been supported by the grant MSM0021620838.