Overview
Problems of English–to–Czech MT

- Large Target–Side Vocabulary
- Higher Out–of–Vocabulary Rate
- Low Reachability of Human Translations

Possible Solutions Examined
- Two–Step Translation
- Coarser Optimisation Metric
  - SemPOS (Kos and Bojar, 2009)

Two–Step Translation
Overcome Target–Side Data Sparseness

Moses runs twice in a row:
1. Translate English to lemmatized Czech augmented to preserve important semantic properties known from the source phrase.
2. Generate fully inflected Czech.

Simple vs. Two–Step Translation

<table>
<thead>
<tr>
<th>Data Size</th>
<th>Parallel Mono Bilingual BLEU</th>
<th>Simple BLEU</th>
<th>SemPOS BLEU</th>
<th>Simple SemPOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>10.28±0.40</td>
<td>30.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>12.50±0.44</td>
<td>31.40</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Minor gain in Small–Small, minor loss in Large–Large setting. Interesting mixed result in Small–Large:
- Indicates that large LM can improve BLEU score without addressing the cross-lingual data sparseness (tackled by Two–Step model and appreciated by SemPOS).
- Note that large monolingual data were used also as the LM in the first step.

Manual Annotation
- 150 sentences manually annotated by two annotators (Small–Large setting).
- Each of them mildly prefers Two–Step model.
- Equal result (23) when limited to sentences where they agree.

Rich Morphology

Large Vocabulary
- Although the training data contains roughly the same number of Czech and English tokens, Czech vocabulary has many more entries.
- Czech vocabulary has approximately double the size compared to Czech lemmas.

Higher Out–of–Vocabulary Rates
- Percentage of reference translations reachable by exhaustive search (Schwartz, 2008).
- Exhaustive search influenced by:
  - Distortion limit (default: 6).
  - Number of translation options (default: 50).
- It is much harder to reach reference translation in Czech than in English.

Lower Reachability of Reference Translations
- Percentage of reference translations reachable by exhaustive search (Schwartz, 2008).
- Exhaustive search influenced by:
  - Distortion limit (default: 6).
  - Number of translation options (default: 50).
- It is much harder to reach reference translation in Czech than in English.

Optimisation towards Coarser Metric – SemPOS

SemPOS:
- Operates on lemmas of content words.
- Ignores word order.
- Reference lemmas matched only if semantic parts-of-speech (Hajić et al., 2004) agree.
- Czech and English supported so far.

Our WMT10 System Configuration

English–to–Czech
- Standard GIZA++ word alignment based on both source and target lemmas.
- Two alternative decoding paths; forms always truecased: form+tag → form
- The first path is more specific and helps to preserve core syntactic elements in the sentence.
- Without the tag, ambiguous English words could often all translate as e.g. nouns, leading to no verb in the Czech sentence. The default path serves as a back-off.
- Significance filtering of the phrase tables (Johnson et al. 2007) implemented for Moses by Chris Dyer; default settings of filter value a=ε and the cut-off 30.
- Lexicalized reordering (oe–bi–fe) based on forms.
- Two separate 5-grm Czech LMs of truecased forms each of which interpolates models trained on the following datasets: the interpolation weights were set automatically using SRILM (Stolcke, 2002) based on the target side of the development set.
- Interpolated CzEng domains: news, web, fiction. The rationale behind the selection of the domains is that we prefer prose-like texts for LM estimation (and not e.g. technical documentation) while we want as much parallel data as possible.
- Interpolated monolingual corpora: WMT09 monolingual, WMT10 monolingual, Czech National Corpus (Kocev et al., 2000) sections SYN2000+2005+2006PUB.
- Standard Moses MERT towards BLEU.

Czech–to–English
- Far fewer configurations tested, this is the final one:
  - Two alternative decoding paths; forms always truecased:
    - form → form
    - lemma → form.
  - Significance filtering as in English–to–Czech.
  - 5-grm English LM based on CzEng English side only.
  - Lexicalized reordering (oe–bi–fe) based on forms.
  - Standard Moses MERT towards BLEU.

Using GigaWord LM as compiled by Chris Callison–Burch caused a significant loss in quality, probably due to different tokenization rules.