Data Issues and Normalization

Different corpora were processed differently:
- Tieds:
  - Sentence ends with a full stop (.)
  - Euro-numeric digits (0123456789)
- Emilee:
  - Sentence ends with a dash (--)
  - Devanagari digits (०१२३४५६७८९)
- What else can't be written in more ways:
  - Characters within ( ' ” )
  - Combining diacritics ordering ‘ति’ vs. ‘ती’
  - Candidate replaced by anusvara ए◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌●

Conclusion
- Best published BLEU score for TIDES test set achieved.
- In general, the English-to-Hindi MT comparison in problematic due to different datasets used by various research groups.
- Hierarchical models (Joshua) lead to better BLEU than Moses with morphological factors.
- Manual evaluation less conclusive about the improvement.
- Lessons learned about the data:
  - Obtaining data is easier than cleaning them up.
  - Two different corpora from different sources may overlap!

Moses vs. Joshua
- English–Hindi translation requires significant amount of rendering.

Phrase-based decoding (Moses)
- Phrase-based models (Moses) explore the space of hypothesis from left to right.
- The default rendering model promotes sentence translations.
- The best available alternative is localized rendering, swap phrase given the words in there.

Hierarchical decoding (Joshua)
- Hierarchical models (Joshua) parse the input sentence and render constituent as needed.
- Grammar extracted automatically from the parallel training corpus.

Morphology in Moses
- Moses supports explicit modeling of morphology on the target side.
- An additional language model is applied on the stream of target-side tags.

We experiment with several formalizations of Hindi morphology:

<table>
<thead>
<tr>
<th>Form</th>
<th>Tag</th>
<th>Textbook</th>
<th>2 Letters</th>
<th>4 Letters</th>
<th>8 Letters</th>
<th>10 Letters</th>
<th>Ambidextrous</th>
<th>Adjacent</th>
<th>Adjacent Alt</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Hind</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Manual Quality Judgements
- Three independent samples (100 sets, 100 sets, 50 sets).
- The annotator sees the input sentence and several hypotheses.
- The reference translation is shuffled among the hypotheses.

SRC: the private sector units are thirty to forty years old.
	- English: भरी सेवा लाभ देने का मतलब है कि वैसे ही कुछ चीजों के साथ है।
	- भरी सेवा की स्थापना का मतलब है कि वैसे ही कुछ चीजों के साथ है।
	- भरी सेवा की स्थापना का मतलब है कि वैसे ही कुछ चीजों के साथ है।
- Flags used:
  - empty incomprehensible,
  - related to input, partial translation of phrases,
  - acceptable and preserving most of the meaning, possibly still with many errors.
- Different from Maneshinath et al. (2009) who claim to have improved on average:
  - from little meaning conveyed, disjoint Hindi, most phrases correct, imperfect morph.
  - to much of meaning conveyed, non-native Hindi, low minor grammatical errors.

Morphology and more data for Moses

<table>
<thead>
<tr>
<th>System</th>
<th>Hindi</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Moses vs. Joshua
- Identical training conditions:
  - Data: Tides + Daniel Pipes, no morphology
  - Moses does not differ significantly between Treading and WCIO.

Impact of Emille training data on Moses
- BLEU almost matches manual judgments this time.
- The addition of Emille significantly decreases the quality.
- Other data slowly compensate for the loss.

A later analysis revealed that Emille overlaps with Tides development dataset, leading to model overfitting.

The research was supported by the grants GAUK 7817/1S/2015/2015 (Czech Academy of Sciences in the Czech Republic), 10400210/03/5998 (Czech Ministry of Education), P107/11/0503 (Zuzana Klusáková), and GAUK 5970089 (Grant agency of the Charles University).