HABLex: Human Annotated Bilingual Lexicons for Experiments in Machine Translation
Brian Thompson,* Rebecca Knowles,* Xuan Zhang,* Huda Khayrallah, Kevin Duh, Philipp Koehn
Johns Hopkins University

Overview

What is in the HABLex dataset?
- Human-generated *alignments* of words and phrases.
- Development and test set.

When to use the HABLex dataset?
Benchmarking methods for *bilingual lexicon integration* into neural machine translation.

Why is bilingual lexicon integration desirable?
- high-tech vocabulary
- low resource
- user requirement
- improve rare word translation

What are the challenges of bilingual lexicon integration?
- Arbitrary dictionaries have problems: e.g. overlap entries, ineffective
- Hard to evaluate only based on BLEU.
  In need of bilingual lexicons **tailored to dev and test set**.

**Domain:** Patent
**Corpus:** World Intellectual Property Organization (WIPO) COPPA-V2
**Language Pairs:** Russian -> English, Korean -> English, Chinese -> English

Two-step process:
1. Identifying *rare words* on the source side of the test and development sets.
2. Human annotators correcting or validating automatic alignments of the identified words.

* These authors contributed equally to this work.

<table>
<thead>
<tr>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entries</td>
<td>Sentences</td>
</tr>
<tr>
<td>Ru</td>
<td>9040</td>
</tr>
<tr>
<td>Ko</td>
<td>5593</td>
</tr>
<tr>
<td>Zh</td>
<td>1773</td>
</tr>
</tbody>
</table>

BL (Baseline system): First train on general domain data, then fine-tune on Patent data.

Recall: Percentage of the time the system output contains the correct lexicon translation.