**BACKGROUND**

Cross-Lingual Word Embeddings (CLWEs)
- represents words from two (or more) languages in a shared embedding space.
- allows retrieval of translation pairs through Bilingual Lexicon Induction (BLI)

Why CLWEs/BLI? Machine Translation?
- low supervision
- no need for sentence-aligned parallel corpora
- data requirements:
  - monolingual corpora
  - small seed dictionary

However, most studies on CLWEs focus on:
- homogeneous language pairs (i.e. European Languages)
- moderate to high resource languages (>100M tokens)

**METHODS**

**STRATEGY**
- Increase the size of the existing monolingual corpus.

**CLWE METHODS**
- Find a CLWE approach applicable to a low-resource setting.

**SEED LEXICON**
- Determine the optimal seed lexicon, in terms of content and size.

**TRANSFER LEARNING**
- Leverage linguistically similar and/or high-resource languages.

**Ultimate Goal**: Improve the BLI performance of EN-HIL CLWEs.

1. Increased corpus size.
2. Well-curated seed lexicon.
3. Adapted applicable methods.
4. Transfer learning: leverage high-resource languages.

**EXPERIMENTS & RESULTS**

1. **MWEs: fastText vs SGNS**
   - fastText was able to capture spelling differences and gender, while SGNS was able to capture orthographic and morphological regularities.

2. **Seed Lexicon**
   - High coverage
   - Coverage Target Model vs.
   - EN-HIL Corpus
   - fastText vs SGNS
   - fastText vs SGNS

3. **Transfer learning with pivot languages.**
   - Success of transfer learning is dependent on the quality of the source and target monolingual word embeddings.

**CONCLUSION**

Well-curated seed lexicon for attaining improved retrieval.

** target languages **

<table>
<thead>
<tr>
<th>Target Languages</th>
<th>Corpus</th>
<th>Word Vectors</th>
<th>P@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-Es-HIL</td>
<td>1.2M tokens</td>
<td>9.7% (SGNS)</td>
<td>9.25%</td>
</tr>
<tr>
<td>EN-TL</td>
<td>Wikipedia built</td>
<td>64.6k (fastText)</td>
<td>15.90%</td>
</tr>
</tbody>
</table>