1. Introduction and Motivation

Unsupervised Bilingual Lexicon Induction

Consider two related languages, source and target.
- Given a word in the source language, find a word in the target language with the same meaning.
- Unsupervised: resources are limited to two corpora, one in each language, of the same genre (to ensure sufficient overlap in vocabularies), but no alignment or parallelism.
- Our method extracts a small initial seed and bootstraps to produce high-quality translations.

2. Methods

2.1 Seed Lexicon Extraction

- We assume source and target languages are related.
- Related languages typically have cognates: words with a shared linguistic origin.
- Cognates often have similar spelling, frequency, and meaning.
- We can use similarity to find cognates and build a seed lexicon.
- Examine pairs of high-frequency words: let \( r_w \) be the frequency rank of word \( w \) in its corpus.
- We tune frequency and similarity thresholds on development data.

```
function EXTRACT_SEED(m, p, d):
    seed ← Ø
    for \( i \) from 1 to \( m \) do:
        \( s \) ← source word such that \( s_i = r_w \)
        for each target word \( t \) do:
            if \( \text{NED}(s, t) \leq d \) and \( \|s - t\| \leq p \) then:
                seed ← seed \cup \{s, t\}
    return seed
```

2.2 Translation Matrices

- For each language, source and target, word2vec (Mikolov et al., 2013a) creates a vector space; every word is a vector in the space of its language.
- Key idea: learn a linear transformation between the source and target vector spaces.
- Use the seed lexicon pairs \( (u, v) \) and SGD to train a matrix \( T \) such that \( T u = v \).
- Also train reverse translation matrix: \( T v = u \).
- Translate source word \( w \) / vector \( u \):

\[
\text{score}(w, v) = \frac{\text{sim}(T \cdot u, v) + \text{sim}(T^{-1} \cdot v, u)}{2}
\]

3. Experiments

3.1 Data: Europarl
- Language pairs: Spanish-French (ES-FR), English-French (EN-FR), and English-Spanish (EN-ES); both directions.
- Development on ES-FR only.

3.2 Evaluation:
- Following Dou and Knight (2013), use GIZA++ (Och and Ney, 2003) to align a parallel corpus, use alignment pairs to induce a gold-standard lexicon.
- Source/target vocabularies: 2k most frequent source/target words not found in the seed lexicon.

4. Results

Comparison against Haghighi et al (2008)

```
<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Seed size</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES-FR</td>
<td>206</td>
<td>87.9</td>
</tr>
<tr>
<td>EN-FR</td>
<td>191</td>
<td>80.1</td>
</tr>
<tr>
<td>EN-ES</td>
<td>239</td>
<td>83.3</td>
</tr>
<tr>
<td>FR-ES</td>
<td>214</td>
<td>93.0</td>
</tr>
<tr>
<td>FR-EN</td>
<td>210</td>
<td>79.1</td>
</tr>
<tr>
<td>ES-EN</td>
<td>252</td>
<td>88.9</td>
</tr>
</tbody>
</table>
```

4 Conclusion

- Novel method combines lexical and frequency information to extract a seed lexicon from non-parallel corpora.
- Combined with a word-embedding-based bootstrapping method, we have created a fully unsupervised bilingual lexicon induction algorithm which outperforms prior work.
- Innovative bi-directional scoring improves results and gives a more robust algorithm.
- Can be applied to low-resource languages – a large text corpus in each language is the only requirement.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada, Alberta Innovates – Technology Futures, and Alberta Advanced Education.