Bootstrapping Unsupervised Bilingual Lexicon Induction

Bradley Hauer, Garrett Nicolai, and Greg Kondrak
University of Alberta

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 $r_s = i$
    for each target word t **do**:
      if $NED(s, t) \leq d$
      and $|r_s - r_t| \leq p$
      and $s \neq t$ **then**:
        $seed \leftarrow seed \quad \{(s,t)\}$
  return *seed*

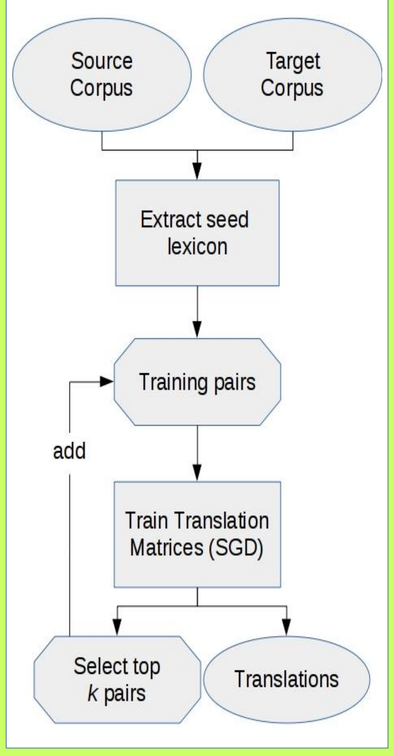
2.2 Translation Matrices

- 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_i, v_i) and SGD to train a matrix T such that $\mathbf{T}u_i = \mathbf{v_i}$
- Also train reverse translation matrix: $T'\mathbf{v_i} = \mathbf{u_i}$
- Translate source word w / vector **u**:

$$score(\boldsymbol{u}, \boldsymbol{v}) = \frac{sim(T \cdot \boldsymbol{u}, \boldsymbol{v}) + sim(T' \cdot \boldsymbol{v}, \boldsymbol{u})}{2}$$

2.2 Bootstrapping

- The translation function induced by the seed lexicon has low accuracy, but it gets some words correct.
- Key idea: add high-scoring (i.e. high-confidence) pairs to the seed lexicon.
- Training data expands to cover more of the source and target vocabularies.
- Accuracy of translations improves.
- Able to identify more high-confidence pairs to add to the training data.
- Repeat to iteratively better translations.
- Fully unsupervised!



3. Experiments

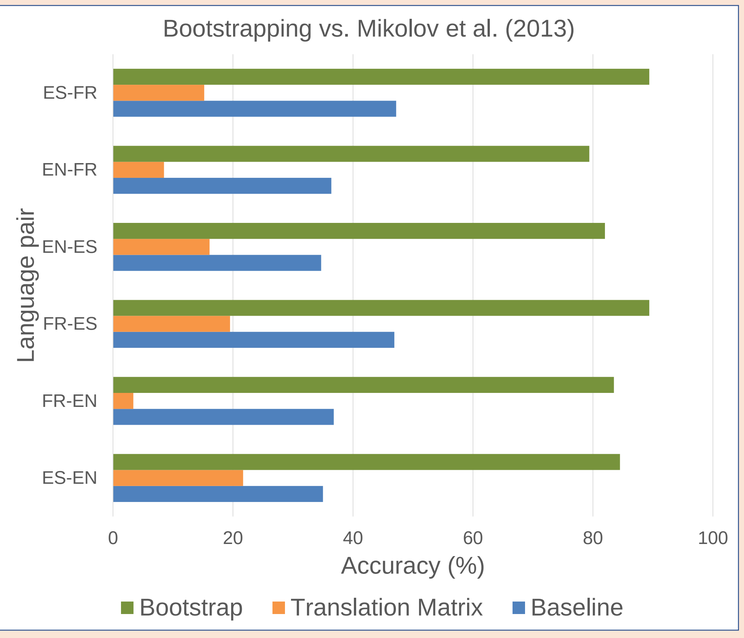
- Data: Europarl
  - Language pairs: Spanish-French (ES-FR), English-French (EN-FR), and English-Spanish (EN-ES); both directions.
  - Development on ES-FR only.
- Evaluation:
  - Following Dou and Knight (2013), use GIZA++ (Och and Ney, 2003) to align a parallel corpus, use alignment pairs to induce a gold-standard lexicon.
  - Source/target vocabularies: 2k most frequent source/target words not found in the seed lexicon.

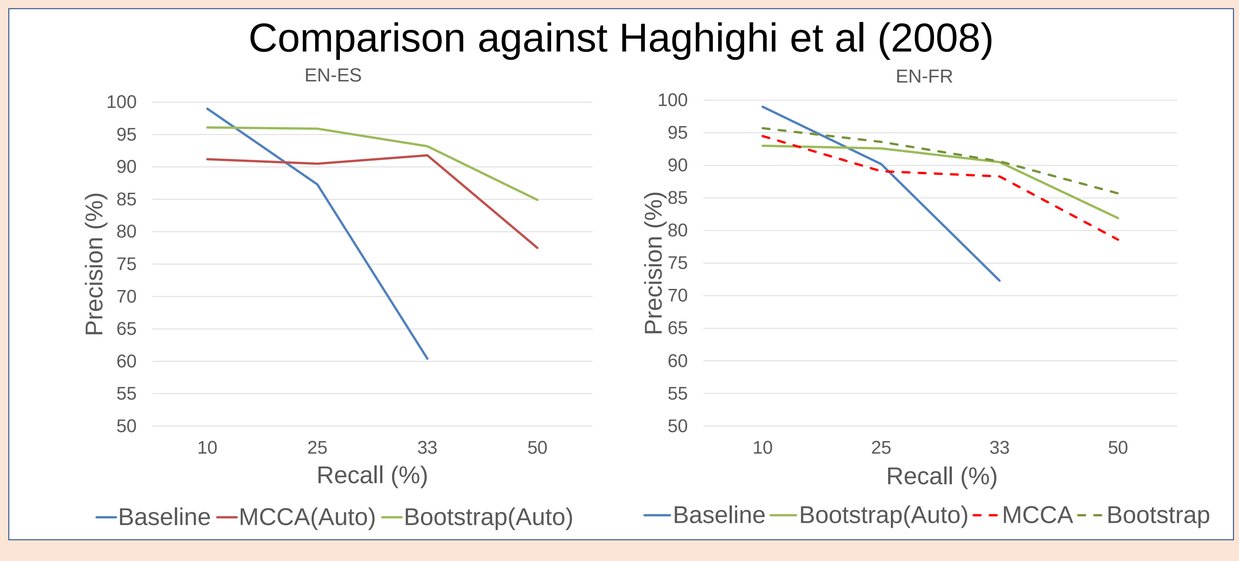
  Evaluated against:
  - Edit distance baseline.
  - Mikolov et al (2013b): one-shot unidirectional translation matrix (same seed and vectors as our bootstrapped method).
  - Reported results of Haghighi et al (2008) (MCCA)

4. Results

Lang-pair	Seed size	Accuracy (%)
ES-FR	206	87.9
EN-FR	191	80.1
EN-ES	239	83.3
FR-ES	214	93.0
FR-EN	210	79.1
ES-EN	252	88.9

Automatic seed extraction: size and accuracy





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.