Bilingual Lexicon Induction for Less Commonly Used Languages

Abstract

1 Introduction

2 Inducing Bilingual Lexicons

Cues and similarity metrics:

- Context (including contextual NEs), using dependency contexts for the resources rich side (i.e. English).
- Time
- Topics (i.e. wiki categories)
- Edit distance

Combination strategies:

- cue scores as classification features: use seed dictionaries for supervised data.
- rank aggregation

Evaluation: tokens for inducing translations. Evaluation metrics: why precision at top-k?

3 Experimental Evaluation

3.1 Data and Other Resources

Describe the data:

- Wiki
- News

Describe the resources:

- Dictionaries: generally, noisy
- Parallel data: some languages may have small amounts of parallel data

3.2 Quality of Available Resources

Parallel data experiments:

- Moses lexical tables vs. monolingual cues (e.g., Fig. 1).
- Use Moses lexical tables as seed dictionaries.

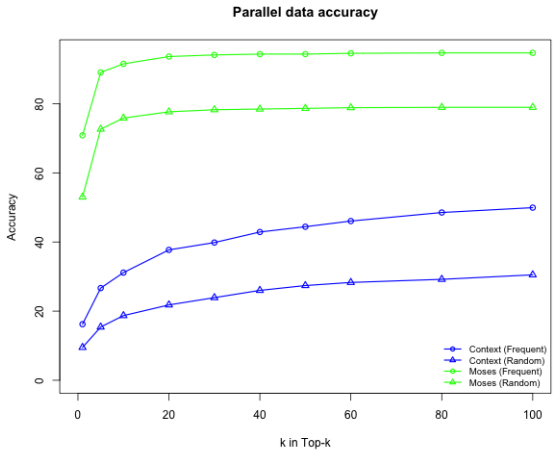

Figure 1: Bilingual lexicons from Moses lexical tables vs. using contextual cues.

3.3 Quality of Individual Cues

Performance of individual cues per language.

3.4 Combination strategies

Classification and rank aggregation.

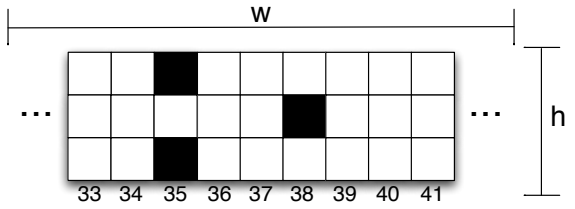

Figure 2: A Spectral Bloom Counter of width w and height $h = 3$. Locations 35 and 38 respectively contain the bit sequences 101 and 010.

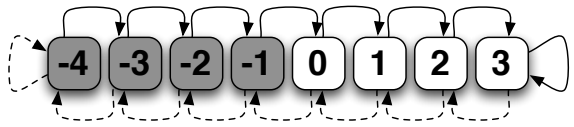

Figure 3: A SumSign Markov Chain using $b = 3$ bits, and thus $2^b = 8$ states. Dark and light states correspond to whether a given bit of an LSH signature will be set to 0 or 1, respectively. States are numerically labelled in order to reflect the similarity to a low bit order integer data type, that never overflows.

Space Efficient Online LSH

4 Related Work

- Context: (**?**; **?**; **?**)
- Time: (**?**; **?**)
- Topics: (**?**; **?**)
- Multiple: (**?**; **?**; **?**)
- Dependecies: (**?**)
- Bridge languages: (**?**)
- Combination Strategies: (**?**; **?**; **?**)
- Mechanical Turk: Our NAACL workshop paper.
- Other: (**?**)

5 Conclusions and Future Work

Using cues for MT.

Acknowledgments