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.

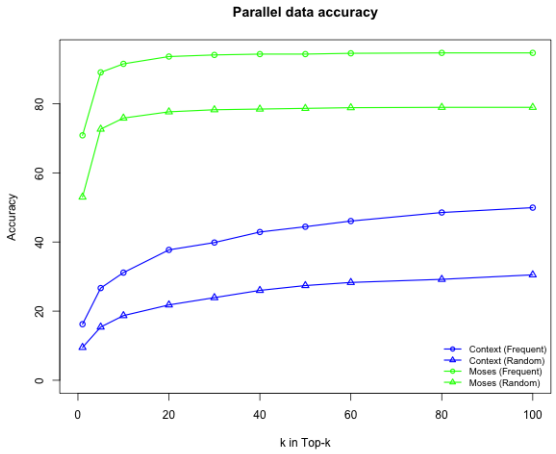

Figure 1: Bilingual lexicons from Moses lexical tables vs. using contextual cues.

3.3 Quality of Individual Cues

Performance of individual cues per language.

3.4 Combination strategies

Classification and rank aggregation.

4 Related Work

- Context: (Rapp, 1995; Rapp, 1999; Fung and Yee, 1998)
- Time: (Schafer and Yarowsky, 2002; Klementiev and Roth, 2006)
- Topics: (Mimno et al., 2009; Boyd-Graber and Blei, 2009)
- Multiple: (Schafer and Yarowsky, 2002; Koehn and Knight, 2000; Haghighi et al., 2008)
- Dependecies: (Garera et al., 2009)
- Bridge languages: (Mann and Yarowsky, 2001)

- Combination Strategies: (Koehn and Knight, 2000; Klementiev and Roth, 2006; Klementiev et al., 2008)
- Mechanical Turk: Our NAACL workshop paper.
- Other: (Monz and Dorr, 2005)

5 Conclusions and Future Work

Using cues for MT.

Acknowledgments

References

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 414–420.

Nikesh Garera, Chris Callison-Burch, and David Yarowsky. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proc. of the Annual Conference on Computational Natural Language Learning (CoNLL)*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alexandre Klementiev, Dan Roth, and Kevin Small. 2008. Unsupervised rank aggregation with distance-based models. In *Proc. of the International Conference on Machine Learning (ICML)*.

Philipp Koehn and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proc. of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Christof Monz and Bonnie J. Dorr. 2005. Iterative translation disambiguation for cross-language information retrieval. In *Proc. of International Conference on Research and Development in Information Retrieval (SIGIR)*.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 320–322.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 519–526.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proc. of the Annual Conference on Computational Natural Language Learning (CoNLL)*, pages 146–152.