Bilingual Lexicon Induction from Monolingual Sources for Low Resource Languages

Abstract

Statistical machine translation relies on the availability of substantial amounts of human translated texts. Such bilingual resources are available for relatively few language pairs, which presents obstacles to applying current statistical translation models to low-resource languages. In this work, we induce bilingual dictionaries from more plentiful monolingual corpora using a diverse set of cues, including: cross-lingual vector space models, the frequencies of words over time, orthographic similarity, etc. We report the efficacy of these monolingual cues and contrast their performance for a language pair where plentiful bilingual resources are available. We further evaluate the accuracy of bilingual dictionaries induced between English and XX, YY, ZZ. Rather than evaluate only on the 1000 most frequent nouns, as previous work on lexicon induction has done, we further evaluate on a random sample of lower frequency words. We introduce a novel, space-efficient extension to the locality sensitive hashing (LSH) scheme that exploits cross-lingual, phrasal distributional statistics.

1 Introduction

Statistical methods for machine translation continue to push the state of the art in automatic translation. However, they crucially rely on the availability of large numbers of translations aligned across two languages. Generation of these parallel corpora require the efforts of bilingual speakers and are extremely expensive to produce in sufficient quantities to induce a high quality statistical translation system. As a result, these methods can not be successfully applied to the majority of word's languages and especially those less frequently taught. In terms of the community's evaluation of progress, most shared tasks involve european languages for which generous quantities of multilingual parliament proceedings are available.

At the same time we now have unprecedented access to vast and continually expanding monolingual resources. Moreover, they often contain additional metadata which can provide non-sentential cues for inducing bilingual resources; suggesting we might substantially reduce and eventually eliminate the requirement for explicitly aligned bilingual translations. Recent examples include exploiting temporal information to induce Named Entity lexicons ((**?**; **?**)), and topic information to generate translations ((**?**)). Moreover, resources such as these are likely to be extremely useful for numerous multilingual NLP tasks ().

Examples

2 Inducing Bilingual Lexicons

Our goal is to generate bilingual resources for pairs of languages for which we have sufficient monolingual data. These languages currently include .

List languages

This article consitutes the first report on this effort: since more data is continuously becoming available, we expect to add more languages to this list in the future.

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.

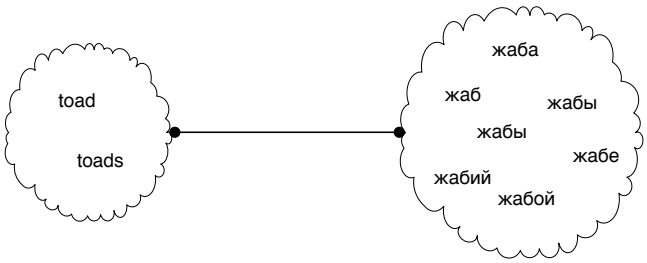

Figure 1: Inducing lexicons using contextual similarity. An example set of word forms for *toad* for English and Russian.

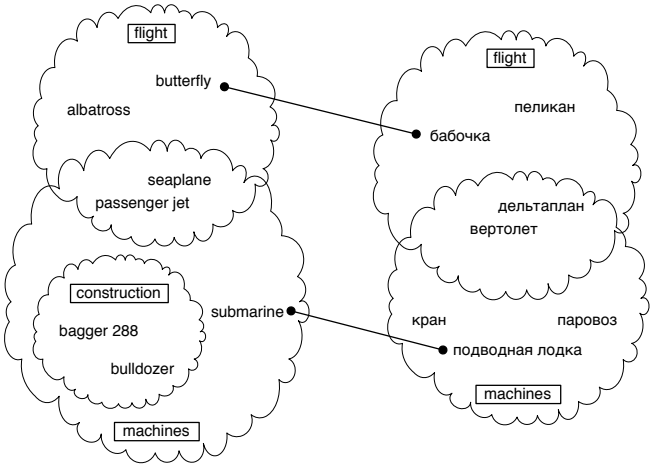

Figure 2: Hierarchical clustering: associations can be inferred between sets of semantically related words. Intersections between sets e.g. "flying machines") can also substantially limit the sets of potential translation candidates.

- rank aggregation

3 Contextual Similarity

Grouping semantically related words through distributional statistics (e.g. (Pereira et al., 1993)) is a way to reduce both the search space and associate translations. We can find associations between clusters instead of individual lexemes or phrases, by intersecting them we can substantially reduce the space of possible translations (Fig. 2). .

Verbiage is very close to CCB's proposal

4 Experimental Evaluation

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

4.1 Data and Other Resources

Describe the data:

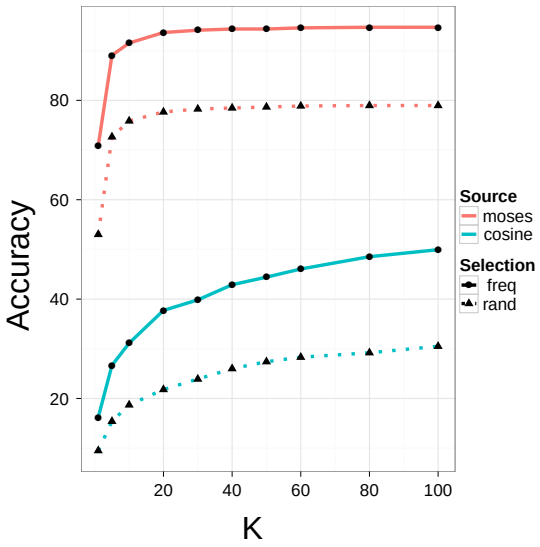

Figure 3: Moses as compared to cosine.

- Wiki
- News

Describe the resources:

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

4.2 Quality of Available Resources

Parallel data experiments:

- Moses lexical tables vs. monolingual cues (e.g., Fig. 4).
- Use Moses lexical tables as seed dictionaries.

4.3 Quality of Individual Cues

Performance of individual cues per language.

4.4 Combination strategies

Classification and rank aggregation.

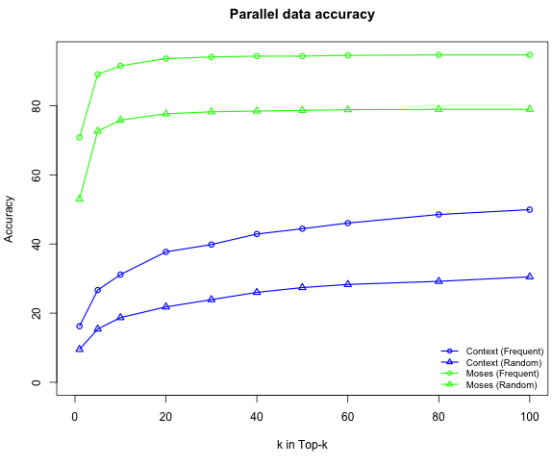

Figure 4: Bilingual lexicons from Moses lexical tables vs. using contextual cues.

5 Related Work

- Context: (**?**; 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)

6 Conclusions and Future Work

Using cues for MT.

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)*.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of english words. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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.