Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification

{ JEREMY.BARNES, TONI.BADIA }@UPF.EDU PATRIK.L@WEBINTERPRET.COM

Universitat Pompeu Fabra Barcelona

Introduction

Successful aspect-based[a] sentiment analysis systems require sophisticated NLP tools and resources, such as large-coverage sentiment lexicons, accurate parsers or annotated corpora. However, many languages lack these resources and recreating them is not a trivial task. This motivates the need to look for techniques to transfer this knowledge from one language to another.

Motivation:

- Not all languages have quality machine translation tools.
- Translation has been shown to change the sentiment of texts.

Research Questions:

- Are cross-lingual distributional semantic approaches competitive with SMT for this task?
- Given that we would like to use a minimum amount of parallel data, do techniques that use less parallel data perform equally well or near?

[a]The word "aspect" here refers to a feature of an entity. If the entity in question is a hotel, common aspects would include beds, breakfast, or wifi.

Materials

- 2016 Wikipedia Dump (English, Spanish) preprocessed to remove markup and lowercased
- OpeNER corpus (English, Spanish) annotated for four levels of sentiment at aspect-level
- Europarl v7 corpus (English, Spanish)
- Small in-domain parallel corpus created from web scraping

Methods

We compare five techniques for performing cross-lingual sentiment analysis, as well as two monolingual baselines.

Google SMT

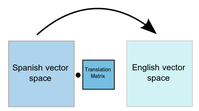
Google Translate

Constrained SMT

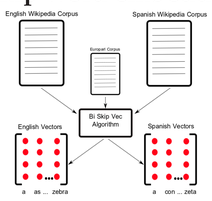
Source: On the other hand **<zone><wall/> a big advantage <wall/></zone>** of the hostel is its placement.

Translation: por otra parte **<ou1-P>una gran ventaja</ou1>** del hostal es su colocación.

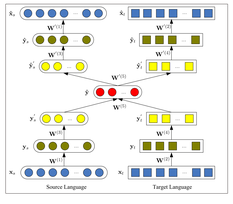
Zero-shot transfer



BiSkip word embeddings

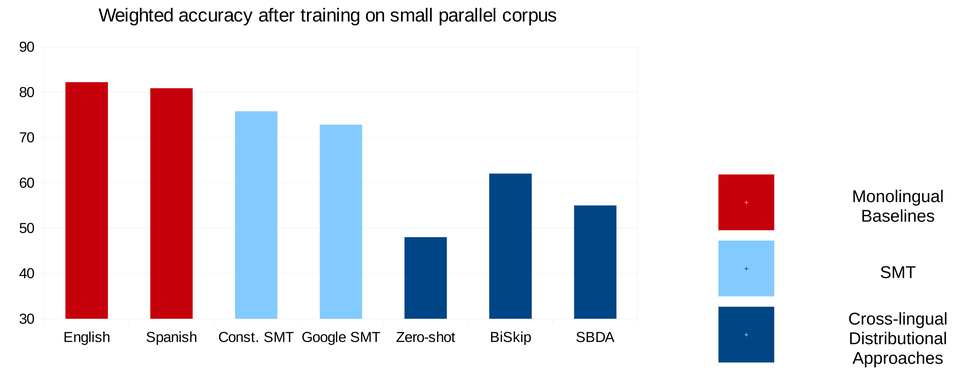


Stacked Bilingual Denoising Autoencoders (SBDA)

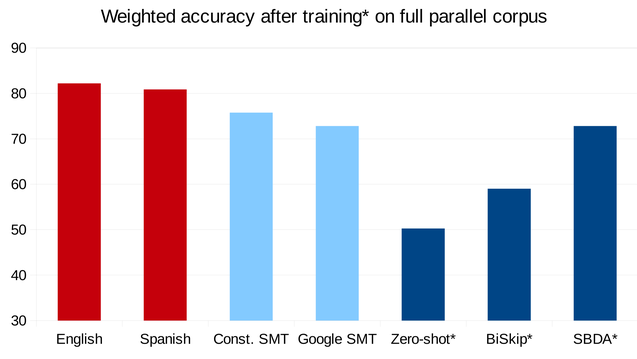


Results

The distributed representations performed similarly to machine translation given enough parallel data. However, their performance drops much faster when this data is lacking.



* Trained on 15.9 M parallel tokens



* Trained on 49 M parallel tokens

References

[1] Lambert, Patrik. 2015. Aspect-level cross-lingual sentiment classification with constrained SMT In *Proceedings of ACL 2015*

[2] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of NAACL 2015 Workshop on Vector Space Modeling for NLP*

[3] Zhou, Guangyou, Zhiyuan Zhu, Tingting He, and Xiaohua Tony Hu. 2016 Cross-lingual sentiment analysis with stacked autoencoders. In *Knowledge and Infomation Systems* 47(1):27-44.

Conclusion

- Distributed representations can be competitive with machine translation for cross-lingual sentiment analysis.
- Bilingual word embeddings show promise as they can theoretically incorporate sentiment information.
- Stacked bilingual autoencoders perform well with a large amount of parallel data, but quickly lose effectiveness.

Future Work

- Finalization of annotated corpora in truly under-resourced languages (Catalan, Basque).
- Improving bilingual word embeddings for the task of sentiment analysis.