Word Translation Without Parallel Data (2018)

Poster by: A. Puigdemont Monllor, J. Ciesko

Faculty of Informatics, Masaryk University

Objectives

- Develop an unsupervised method for cross-lingual word embedding alignment using Generative Adversarial Networks (GANs).
- Minimize cross-lingual alignment errors with Procrustes analysis to improve semantic similarity.
- Enhance retrieval accuracy with Cross-domain Similarity Local Scaling (CSLS).

Introduction

Cross-lingual word embeddings provide a powerful way to model semantic similarity between words across different languages, supporting applications in machine translation, information retrieval, and transfer learning. Traditional methods require large parallel corpora, which are expensive and unavailable for many language pairs. This model enables cross-lingual alignment without labeled data by mapping monolingual embeddings in different languages through adversarial learning and refining alignment with the Procrustes approach.

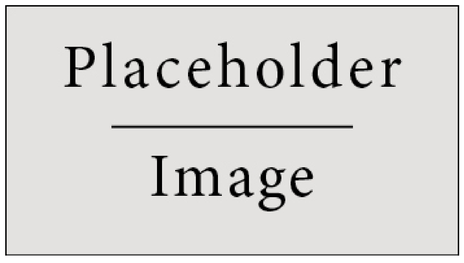

Figure 1:Model Architecture Overview

Methods

- **GAN Training**: The model uses a generator G that maps source embeddings X_s to the target space, and a discriminator D that distinguishes between transformed source embeddings and target embeddings X_t. The GAN objective is to minimize:

$$\min_G \max_D \mathbb{E}_{X_t \sim P_t} \log D(X_t) + \mathbb{E}_{X_s \sim P_s} \log(1 - D(G(X_s))).$$

- **Procrustes Analysis**: After initial GAN training, Procrustes alignment refines the mapping by minimizing the mean squared error between aligned embeddings:

$$R = \arg\min_{R \in \mathbb{O}_d} ||RG(X_s) - X_t||_F,$$

where R is an orthogonal matrix and $|| \cdot ||_F$ denotes the Frobenius norm.

CSLS Scoring

Cross-domain Similarity Local Scaling (CSLS) rescales similarity to reduce the "hubness" problem common in high-dimensional spaces. CSLS adjusts cosine similarity by penalizing densely clustered embeddings:

$$\text{CSLS}(x, y) = 2\cos(x, y) - r_x - r_y,$$

where r_x and r_y are the average similarity of x and y to their k-nearest neighbors, providing a more robust similarity metric.

Results

Empirical results indicate that the GAN-Procrustes-CSLS combination significantly outperforms baseline methods on bilingual dictionary induction and unsupervised machine translation tasks. CSLS notably improves retrieval by 20-30% on test datasets, validating its role in enhancing semantic alignment.

Conclusion

The integration of adversarial GANs, Procrustes alignment, and CSLS facilitates effective cross-lingual embedding mappings without parallel data. This robust alignment supports practical tasks in machine translation and language processing for resource-limited languages.

References

Mikolov, T., et al. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *NeurIPS*.
Goodfellow, I., et al. (2014). Generative Adversarial Nets. *NeurIPS*.

Contact Information

- Web: http://www.university.edu/smithlab
- Email: john@smith.com

PLACEHOLDER LOGO PLACEHOLDER LOGO