Spectral Graph-Based Methods for Learning Cross-Lingual Word Embeddings

Aurélien Bellet, Pascal Denis, Mikaela Keller

September 28, 2017

Team and contact

- Équipe Magnet, INRIA/CRIStAL: http://team.inria.fr/magnet
- Aurélien Bellet (aurelien.bellet@inria.fr), Pascal Denis (pascal.denis@inria.fr), Mikaela Keller (mikaela.keller@inria.fr)

Keywords

Machine Learning, Natural Language Processing, Word Embeddings, Graph-based Learning.

Context

How to adequately represent words as vectors is a long-standing and crucial problem in the fields of text mining and Natural Language Processing. This question has recently re-surfaced due to the recent surge of research in "deep" neural networks, and the development of algorithms for learning distributed word representations or "word embeddings" (the best known of which is probably `word2vec` [4]). The main appeal of these low-dimensional word representations is two-fold: they can be derived directly from raw text data in an unsupervised or weakly-supervised manner, and their latent dimensions condense interesting distributional information about the words, thus allowing for better generalization while also mitigating the presence of rare and unseen terms. To date, most algorithms for learning word embeddings use some variants of neural networks [4, 5] or standard spectral methods like Principal Component Analysis (PCA) [3] or Canonical Correlation Analysis (CCA) [2]. The quality of word embeddings is typically evaluated either by directly testing for syntactic or semantic relationships between words (including word similarity or syntactic/semantic analogy tasks) [4] or by using word embeddings as features in downstream NLP tasks (e.g., text classification, Part-of-Speech or Named Entity tagging) [9].

Objectives

The goal of this internship is to learn (cross-lingual) word embeddings using graph-based nonlinear manifold learning methods such as Isomap [8], Local Linear Embedding (LLE) [6] and Laplacian Eigenmaps [1]. Given a weighted graph where the weight between two nodes encodes a notion of similarity, these methods aim to learn vectorial embeddings of nodes such that the similarity structure is preserved.

The input graph structure provides a unified framework to learn word embeddings for a single language, and for several languages jointly (by projecting words from different languages into a shared embedding space). In our context, the nodes would be the words and the edge weights would represent the context in which the words appear in the training corpus (e.g., word co-occurrence within some window as in the skip-gram model [4]). In the cross-lingual setting, one may consider

the union of several graphs: the co-occurrence graph mentioned above for each language, and an "alignment graph" derived from alignment frequencies between languages as considered in previous work on cross-lingual embeddings (see [10, 7]).

The tentative work-plan is as follows:

1. Review the relevant literature on word embeddings and graph-based spectral methods.
2. Propose methods to learn mono-lingual and cross-lingual word embeddings through graph-based spectral methods applied to appropriately-designed weighted graphs, and evaluate the performance of the resulting embeddings.
3. If time permits, investigate some further questions, such as (i) design graphs representing richer contexts, and (ii) learn embeddings for higher-order objects (multi-word expressions, sentences, etc).

Skills

Basics in machine learning, graph algorithms and complexity, linear algebra and programming (Python preferred). Familiarity with natural language processing is a plus.

References

[1] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396, 2003.

[2] P. S. Dhillon, D. P. Foster, and L. H. Ungar. Multi-view learning of word embeddings via cca. In *NIPS*, pages 199–207, 2011.

[3] R. Lebret and R. Collobert. Word Embeddings through Hellinger PCA. In *EACL*, 2014.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. Technical report, arXiv:1301.3781, 2013.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, 2013.

[6] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[7] A. Søgaard, Y. Goldberg, and O. Levy. A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments. In *EACL*, 2017.

[8] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[9] J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In *ACL*, pages 384–394, 2010.

[10] S. Upadhyay, M. Faruqui, C. Dyer, and D. Roth. Cross-lingual Models of Word Embeddings: An Empirical Comparison. In *ACL*, 2016.