Representation Learning for Natural Language Processing

10 September 2015

Team and contact

- Équipe Magnet, INRIA/Cristal : http://team.inria.fr/magnet.
- Pascal Denis (mailto:pascal.denis@inria.fr), Mikaela Keller (mailto:mikaelle.keller@inria.fr)

Keywords

Machine Learning (ML), Natural Language Processing (NLP)

Context

The success of Machine Learning algorithms for regression and classification depends in large part on the choice of the feature representations that are used to characterize the input of the problem. This aspect is particularly crucial for problems that live in very high-dimensional spaces. In turn, much human effort is spent on hand-crafting "good" features: these are usually knowledge-based and engineered by domain experts in time consuming trial and error iteration cycles.

This phase of feature engineering is very common in statistical Natural Language Processing (NLP), in where the goal is to capture through features linguistic informations that are relevant to a specific task. A natural question is how one can automate this phase of learning useful features from raw data.

A very popular approach for NLP and text mining problems has been to learn so-called word embeddings: these are real lower-dimensional vectors that are constructed by contrasting word contexts over large quantities of texts. Different algorithms have been proposed for learning these word representations, ranging from neural networks to clustering approaches to purely statistical approaches like Canonical Correlation Analysis.

While these lower-dimensional representations appear to capture some important latent syntactic and semantic information, they tend to be more useful for some tasks (e.g., Part-of-Speech tagging, dependency parsing) than for other (e.g., sentiment analysis).

This raises the question of learning word embeddings in a way that is task-specific: it is indeed crucial to learn representations that correlate well with the underlying unknown labels.

A second open research direction consists in deriving representations for larger text objects, like pairs of words, but also sentences or paragraphs. This suggests two directions, either finding ways to compose word vectors for these larger units, or directly learning low-dimensional representations for them.

Objectives

The first objective of this project is to review the main approaches for learning word embeddings, and to re-implement some of them and construct word embeddings for a language other than English (e.g., French). A second, more programmatic objective would be to start tackling one of the research direction outlined above, such as learning task-specific word embeddings and evaluate this on an NLP problem (e.g., sentiment analysis or discourse relation classification).

References

Collobert, Ronan, et al. Natural language processing (almost) from scratch. The Journal of Machine Learning Research 12 (2011): 2493-2537.

Paramveer S. Dhillon, Dean P. Foster, and Lyle H. Ungar. Multi-view learning of word embeddings via CCA. Advances in Neural Information Processing Systems. 2011.

Igor Labutov and Hod Lipson. Re-embedding words. The 51st Annual Meeting of the Association for Computational Linguistics, 2013.

R. Lebret and R. Collobert. Word Embeddings through Hellinger PCA. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 482-490, Association for Computational Linguistics, 2014.

Pranava S Madhyastha, Xavier Carreras Pérez, and Ariadna Quattoni. Learning task-specific bilexical embeddings. In Proceedings of COLING, 2014.

Mikolov, Tomas, et al. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. 2013.

Levy, Omer, and Yoav Goldberg. Dependency based word embeddings. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014.

Levy, Omer, and Yoav Goldberg. Neural Word Embedding as Implicit Matrix Factorization. Advances in Neural Information Processing Systems. 2014.

Turian, Joseph, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010.

Skills

Algorithms and complexity, basics in Machine Learning, linear algebra and probability. Some familiarity with Natural Language Processing or quantitative linguistics would be a plus, but is not required.