

Learning Word Representations by Embedding the WordNet Graph

Topic: Machine Learning, Natural Language Processing, Word Embeddings, Graph Embeddings

Category: Review, implementation

Contact:

- Pascal Denis (pascal.denis@inria.fr)
 - Rémi Gilleron (remi.gilleron@univ-lille.fr)
 - Nathalie Vauquier (nathalie.vauquier@inria.fr)
-

Description:

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 (NLP). This question has recently resurfaced 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 [6, 7]). Typically, these approaches directly construct word representations from large amounts of unannotated texts, and don’t make use of any linguistic resource.

Due to the ubiquity of networks in many real world applications, and the need for better graph analytics tools, another recent trend of research has been in developing graph embedding techniques [3, 5]. One specific problem is node embedding, where the goal is to encode the graph nodes as low-dimensional vectors that faithfully summarize their graph position and the topology of their local neighborhood. Several new deep learning algorithms have been proposed for node embedding (e.g., node2vec [4], deepwalk [8]), in addition to already well-established matrix factorization-based approaches like Local Linear Embedding (LLE) [9] or Laplacian Eigenmaps [1]. These approaches have been used for different types of graphs such as knowledge graphs and semantic graphs.

Objectives

The overall objective is to improve word representations with the help of existing lexical databases like WordNet (<https://wordnet.princeton.edu/>). For this, we aim to combine word embedding techniques from texts with node embedding techniques over WordNet. This internship is a preliminary step in this direction. The goal is to explore recent node embedding algorithms, in particular node2vec and deepwalk, to learn synset embeddings from the WordNet lexical database.

The tentative work-plan is as follows:

1. Review the relevant literature on word and graph embedding methods.
2. Construct similarity graphs over the 80k WordNet noun synsets, using various synset similarity algorithms [2].
3. Apply node2vec and deepwalk on these similarity graphs to derive noun synset representations.
4. Map these synset representations into word representations, and evaluate these against standard word similarity datasets.
5. If time permits, investigate a new similarity algorithm that would incorporate WordNet edge labels (e.g., hypernym vs. antonym link relations).

Experiments will be done with the help of the Mangoes toolbox (<https://gitlab.inria.fr/magnet/mangoes>). Node embedding algorithms could be integrated in the toolbox.

Skills

Basics in machine learning, graph algorithms and complexity, linear algebra. Familiarity with NLP is a plus.

<http://www.nltk.org/howto/wordnet.html>