Semi-Supervised Text Classification with Word Representations

December 9, 2015

1 Semi-Supervised Text Classification

This project aims at automatically classifying text documents with very high accuracy. These classification tasks will range from predicting the topic of a document to predicting whether the document is spam or what the sentiment reported in this document is.

The classifiers will be learned from both a small set of labeled documents and a very large set of unlabeled ones. To exploit unlabeled documents, semi-supervised techniques based on word representations will be investigates. Specifically, word clusters as well as word embeddings will be induced from unlabeled data and integrated into the text classifiers. Recent work has suggested that such word representations can significantly contribute to the performance of supervised models as described below.

2 Word Representations

2.1 Brown Word Clustering

The Brown algorithm is a hierarchical clustering algorithm that creates clusters of words that are semantically related by virtue of their having been embedded in similar contexts. It is used to address the data sparsity problem inherent in natural langugage processing (NLP). The Brown clusters have beeen shown to be useful in many NLP tasks, including named entity recognition (NER) [1] , dependency parsing [2], as well as relation extraction [3].

2.2 Deep Learning

Word representations induced by deep learning techniques have recently been proposed to further improve the performance of NLP models or, more generally, AI. Deep learning and the representations it induces have consequently been adopted by the major Internet companies such as Google and Facebook[4].

The word representations to be investigated in this project include the word embeddings resulting from word2vec, a deep learning Google project that has significantly impacted the recent developments in NLP[5]. The implementation of these word embeddings is publicly available at `http://deeplearning4j.org/word2vec.html` and will be extensively used in this project.

[1] `http://www.aclweb.org/anthology/P10-1040`

[2] `http://www.cs.columbia.edu/~mcollins/papers/koo08acl.pdf`

[3] `https://www.cs.nyu.edu/~asun/pub/ACL11-FinalVersion.pdf`

[4] See `http://static.googleusercontent.com/media/research.google.com/en//people/jeff/CIKM-keynote-Nov2014.pdf` for how Google applies deep learning techniques as well as `http://homes.cs.washington.edu/~yejin/cse599.html` and `http://yanran.li/resources/` for a comprehensive bibliography about deep learning in NLP

[5] See `http://u.cs.biu.ac.il/~yogo/nnlp.pdf` for an introduction to these deep learning techniques as well as `http://arxiv.org/pdf/1402.3722v1.pdf` for a detailed description of the mathematics behind word2vec

3 Your Task

Your task is to train semi-supervised text classifiers using different word representations and compare their performance in meaningful experiments. The word representations to be investigated include word clusters as well as word embeddings. Additional representations can be investigated.

The project will use as standard academic corpora as well industrial ones. The industrial corpora will be provided by dMetrics [6]. dMetrics.com is a venture-backed MIT spinoff whose innovations in natural language processing have received five consecutive US National Science Foundation awards. Experiments will be performed on both types of corpora and results will compared and discussed. Specifically, to successfully complete this project, you will:

- Train semi-supervised text classifiers using word representations based on Brown clustering.
- Train semi-supervised using word representations based on the word2vec embeddings.
- Evaluate the performance of the trained classifiers and discuss results.

If the results are state-of-the-art, they can be reported in a scientific paper, although this is not a requirement to successfully complete the project.

4 Required Skills

- Java programming is an absolute must.
- Working knowledge of machine learning is required.

5 What you will gain from this project

- Learn about semi-supervised machine learning, clustering and deep learning.
- Gather valuable programming experience with big data.
- Gain industrial experience by working on an industrial project.

[6] See `www.dmetrics.com`