Learning Distributed Word Representations and Applications in Biomedical Natural Language Processing

Jiaping Zheng\textsuperscript{a}, Hong Yu\textsuperscript{b}

\textsuperscript{a} University of Massachusetts Amherst
\textsuperscript{b} University of Massachusetts Medical School

Abstract

A common challenge for biomedical natural language processing (BioNLP) is data sparsity. Distributed word representation approaches have been developed recently that represent words by learning from a large unlabeled data, such approaches have shown to overcome the data sparsity and improve NLP tasks. In this tutorial, we will introduce various word representation approaches, including the traditional Brown Clusters and the more recent distributed word representation (word embeddings), and their potential to improve other NLP tasks. Although we primarily discuss BioNLP applications, the methodologies are applicable to most of learning-based approaches (e.g., image processing and machine translation).

Keywords:
Word Representation; Natural Language Processing.

General Topics

Introduction

A common challenge for biomedical natural language processing (BioNLP) is data sparsity. Distributed word representation approaches have been developed recently that represent words by learning from a large unlabeled data set. Such approaches have shown to overcome the data sparsity and improve NLP tasks [1], [2]. In this tutorial, we will introduce various word representation approaches, including the traditional Brown Clusters [3] and the more recent distributed word representation (word embeddings [2], [4]), and their potential to improve other NLP tasks. Although we primarily discuss BioNLP applications, the methodologies are applicable to most of learning-based approaches (e.g., image processing and machine translation).

Cluster Based Approaches

There are several approaches to learn a word representation from unlabeled data. The first approach is word clustering. In this approach, words are grouped into clusters. Words that are similar are assigned to the same cluster. One prominent example of this approach is the Brown Clusters. Words are clustered hierarchically to maximize the mutual information of bi-grams. An example of a learned cluster is shown in Figure 1 (taken from [5]). It has been shown to improve parsing [6], [7] and Named Entity recognition [8], [9].
Tutorial Structure

We will first introduce the common NLP tasks, such as Named Entity Recognition, and the widely adopted methods to solve the problem. We will then focus on various word representation approaches. We will conclude with discussions of the applications of these word representations.

Educational Goals

Word representations have been demonstrated to improve performance in a variety of tasks in the general NLP domain. The biomedical domain could also benefit from these methods. Our goal is to promote the adoption of these methods and advance the state of the art performance in the biomedical NLP tasks, especially given that there is an enormous amount of unlabeled biomedical data from repositories such as PubMed.

Although some of the methods discussed in this tutorial require large amount of data and computing resources, it is straightforward to incorporate word representations in existing NLP methods to tackle data sparseness problems. There have also been efforts to distribute learned representations to enable other researchers to take advantage of these word features [2], [13], [14].

Expected Attendees

This tutorial is most relevant to informaticians and researchers working with textual data.

References


Address for correspondence

Jiaping Zheng
jzheng@cs.umass.edu
University of Massachusetts Amherst, USA