Data linguistica 30

Splitting rocks: Learning word sense representations from corpora and lexica

av Luis Nieto Piña

Akademisk avhandling för filosofie doktorsexamen
i språkvetenskaplig databehandling,
som enligt beslut av humanistiska fakultetsnämnden
vid Göteborgs universitet kommer att försvaras offentligt fredagen den
13 september 2019 kl. 13.15 i Lilla hörsalen, Humanisten.

GÖTEBORGS UNIVERSITET

Göteborg 2019

TITLE: Splitting rocks: Learning word sense representations from corpora and lexica
LANGUAGE: English
AUTHOR: Luis Nieto Piña

Abstract

The representation of written language semantics is a central problem of language technology and a crucial component of many natural language processing applications, from part-of-speech tagging to text summarization. These representations of linguistic units, such as words or sentences, allow computer applications that work with language to process and manipulate the meaning of text. In particular, a family of models has been successfully developed based on automatically learning semantics from large collections of text and embedding them into a vector space, where semantic or lexical similarity is a function of geometric distance. Co-occurrence information of words in context is the main source of data used to learn these representations.

Such models have typically been applied to learning representations for word forms, which have been widely applied, and proven to be highly successful, as characterizations of semantics at the word level. However, a word-level approach to meaning representation implies that the different meanings, or senses, of any polysemic word share one single representation. This might be problematic when individual word senses are of interest and explicit access to their specific representations is required. For instance, in cases such as an application that needs to deal with word senses rather than word forms, or when a digital lexicon's sense inventory has to be mapped to a set of learned semantic representations.

In this thesis, we present a number of models that try to tackle this problem by automatically learning representations for word senses instead of for words. In particular, we try to achieve this by using two separate sources of information: corpora and lexica for the Swedish language. Throughout the five publications compiled in this thesis, we demonstrate that it is possible to generate word sense representations from these sources of data individually and in conjunction, and we observe that combining them yields superior results in terms of accuracy and sense inventory coverage. Furthermore, in our evaluation of the different representational models proposed here, we showcase the applicability of word sense representations both to downstream natural language processing applications and to the development of existing linguistic resources.

KEYWORDS: language technology, natural language processing, distributional models, semantic representations, distributed representations, word senses, embeddings, word sense disambiguation, linguistic resources, corpus, lexicon, machine learning, neural networks.

DISTRIBUTION:
Department of Swedish
University of Gothenburg
Box 200
SE-405 30 Gothenburg
Sweden

Data linguistica 30
ISSN 0347-948X
ISBN 978-91-87850-75-2
GUPEA <http://hdl.handle.net/2077/60509>

PRINTED in Sweden by GU Interntryckeri Göteborg 2019