To Word Senses and Beyond: Inducing Concepts with Contextualized Language Models

Bastien Liétard and Pascal Denis and Mikaela Keller
University of Lille, Inria, CNRS, Centrale Lille, UMR 9189- CRIStAL, F-59000 Lille, France
first_name.last_name@inria.fr

Abstract

Polysemy and synonymy are two crucial inter-related facets of lexical ambiguity. While both phenomena are widely documented in lexical resources and have been studied extensively in NLP, leading to dedicated systems, they are often being considered independently in practical problems. While many tasks dealing with polysemy (e.g. Word Sense Disambiguation or Induction) highlight the role of word's senses, the study of synonymy is rooted in the study of concepts, i.e. meanings shared across the lexicon. In this paper, we introduce Concept Induction, the unsupervised task of learning a soft clustering among words that defines a set of concepts directly from data. This task generalizes Word Sense Induction. We propose a bi-level approach to Concept Induction that leverages both a local lemma-centric view and a global cross-lexicon view to induce concepts. We evaluate the obtained clustering on SemCor's annotated data and obtain good performance (BCubed F1 above 0.60). We find that the local and the global levels are mutually beneficial to induce concepts and also senses in our setting. Finally, we create static embeddings representing our induced concepts and use them on the Word-in-Context task, obtaining competitive performance with the State-of-the-Art.

1 Introduction

A crucial challenge in understanding natural language comes from the fact that the mapping between word forms and lexical meanings is many-to-many, due to polysemy (i.e., the multiplicity of meanings for a given form)[1] and synonymy (i.e., the multiplicity of forms for expressing a given meaning). Both polysemy and synonymy have been thoroughly studied in NLP, but mostly as independent problems, giving rise to dedicated systems. Thus, Word Sense Disambiguation (WSD) aims at correctly mapping word occurrences to one of its senses [**?**], while Word Sense Induction (WSI), its unsupervised counterpart, aims at clustering word occurrences into latent senses directly from data [**?**, **?**]. More recently, researchers have proposed the task of Word-in-Context (WiC), which consists in classifying pairs of word occurrences depending on whether they realize the same sense or not [**?**]. All these works take a word centric view, which aims at identifying or characterizing the different senses of a given word, where these senses are bound to a word. Another line of work, which takes a broader lexicon-wide perspective, is concerned with identifying synonyms, which are equivalence classes over different words that point to the same concept [**?**, **?**], where concepts are semantic entities that are not bound to a word. In WordNet [**?**, **?**], concepts are called synsets, defined as sets of synonyms. However, outside of lexical resources, synonymy and polysemy are usually considered as independent problems in the NLP literature. Yet, these two views are complementary. In lexicology, they correspond to two perspectives on the word-meaning mapping: semasiology and onomasiology. The former is the word-to-meanings view, where one can observe polysemy by looking at the different meanings a given word has. The latter is the meaning-to-words view, in which one can study synonymy by looking at the inventory of words that speakers use to express the same meaning.

In this paper, we propose a new task, called Concept Induction, that directly aims at learning concepts in an unsupervised manner from raw text. More precisely, this task aims at learning a soft clustering over a target lexicon (i.e., a set of words), in such a way that each cluster corresponds to a (latent) concept. Thus, this task both addresses polysemy (since polysemous words should appear in multiple clusters) and synonymy (since synonymous words should appear in the same cluster(s)). Inducing concepts can be interesting for many external applications, like building lexical resources for low-resources languages [**?**], and can bring a different perspective in computational studies of meaning, moving the usual word-centric focus to a more meaning-centric state.

Our approach to Concept Induction relies on word occurrences for a target lexicon, represented as word embeddings derived from a Contextualized Language Model (in this case,

[1] In this paper, we take polysemy in its most comprehensive definition, also including homonymy.

BERT Large [**?**]), which are then grouped, using hard clustering algorithms, into concept denoting clusters. While these concept clusters could in principle be obtained directly from word occurrences, we propose a bi-level methodology that leverages both a local, lemma-centric clustering (i.e., operating on only specific word occurrences), and a global, cross-lexicon clustering (i.e., operating on all words occurrences). From this perspective, our approach generalizes, and in fact builds upon classical Word Sense Induction, in that word senses are learned jointly alongside with concepts. We hypothesize that an approach taking both complementary resolutions in account will lead to improved Concept Induction and Word Sense Induction, i.e. that the two objectives can be mutually beneficial.

To validate our approach, we carried out experiments on the SemCor dataset, which provides a set of concepts (taking the form of WordNet synsets) related to word occurrences. We found that our bi-level clustering approach accurately learn concepts, achieving F1 scores above 0.60 on the task of Concept Induction compared to WordNet's synsets, outperforming competing approaches that use only local and global views. This demonstrates the benefits of our bi-level approach, and its ability to leverage both local and global views when inducing concepts. Interestingly, we show that the benefits go both ways: our proposed approach outperforms lemma-centric approaches when evaluated for WSI. Finally, we show that concept-aware static embeddings derived from our approach are also competitive with state-of-the-art approaches efficient on the Word-in-Context task, while using less training data. Through the new task of concept induction, we also contribute in a new way to the ongoing debate regarding the ability to align vector representations extracted from Contextualized Language Models to the semantic representations posited by (psycho-)linguists. In this vein, we conduct a qualitative evaluation of obtained clusters to ensure they indeed reflect concepts and gather synonyms. The source code we used for experiments is available at `https://github.com/blietard/concept-induction`.