University *of Ljubljana*
Faculty *of Computer and Information Science*

Cross-lingual sense disambiguation

Erazem Pušnik, Rok Miklavčič, and Aljaž Šmaljcelj

Abstract

TODO

Keywords
Natural language proccessing, word context, disambiguation

Advisors: Slavko Žitnik

Introduction

In human language, one word can often be used in multiple ways. It is important to understand word usage to properly develop natural language processing applications. While the problem is easy for humans to solve, since we know from experience of how the world works, computers aren't able to naturally do so.

Word Sense Disambiguation is an important method of dealing with a word context and it involves the use of syntax, semantics and word meanings in context. It can be used in multiple applications, like machine translation, speech processing, information retrieval, text processing and more. The goal of WSD is to ground the meaning of words in certain contexts into concepts as defined in some dictionary or lexical repository. WSD requres two main information sources: context and knowledge. Sense-tagged corpora provide knowledge, leading to data-driven or corpus-based WSD. Use of lexicons or encyclopaedia lead to knowledge-driven WSD. Context is established from neighbouring words and the domain of discourse.

While simple approaches like naive Bayes classifier or Corpus Lesk algorithm, more complex neural network approaches became popular in the 2010s. As an exmaple, Wiriyathammabhum et al. (2012) applied Deep Belief Networks (DBN) and Yuan et al. (2016) proposed a semi-supervised LSTM model with label propagation.

Related work

The English variant of this task is described by Wang et. al. [1] with a Word-in-Context (Wic) task. This task is defined as a binary classification of sentence pairs. Given two sentences who both contain the specified (target) word, the task determines whether the target word is used in same context (sense) in both sentences.

Dataset used in the mentioned task is also named Wic (ang. *Word-in-Context database*) [2]. It was constructed from example usages in various lexical resources and compiled with constraints of not having more than three instances for some target word and not having repeated contextual sentences across instances.

In the article [3] the corpus is constructed using stop word removal which can reduce the size of the corpus up to 25% and improve the accuracy of the model. The article provides insights to two different methods of word sense disambiguation - shallow approach where the model only considers the surrounding words and deep approach where the model is knowledge-based and supervised.

In case we don't have a labeled dataset available, we could perform unsupervised learning. One of the methods is Word sense induction, specifically context clustering. The method works under the assumption that words are contextually similar if they appear in similar windows of context [4].

Ideas for implementation

As a part of our task we will attempt to construct a new corpus that will contain pairs of sentences which contain one particular word. Corpus will also contain information whether this word is used in same context in both sentences or not.

We will obtain sentences from another corpus Gigafida[1]. To construct pairs of sentences we will predefine selected words in Slovene that have ambigious meanings and then search Gigafida to obtain multiple sentences

[1] http://eng.slovenscina.eu/korpusi/gigafida

that contain these selected words. As the selected words can have different forms, we will search by lemmas.

Regarding context extraction we will use word sense induction, specifically context clustering. It is a unsupervised method where for every instance of a target word in corpus we will compute context vector. We will then perform clustering of all the vectors into predefined clusters. For every newly given sentence with a target word we will compute its context vector and check for closest cluster. Two instances of same word in a pair of sentences will be determined as being used in similar context if both context vectors will be placed in the same cluster. However, the process of creating a corpus will only be semi automatic as we will need to manually check the corpus for any instances when the results were not correct.

We will also analyse the resulting corpus by comparing its accuracy with human's that is said to be around 80% [2] for English language. For Slovene human threshold we will first attempt to determine same context between ourselves and take the average as the human accuracy for Slovene.

References

[1] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

[2] Mohammad Taher Pilehvar and José Camacho-Collados. Wic: 10, 000 example pairs for evaluating context-sensitive representations. *CoRR*, abs/1808.09121, 2018.

[3] Madeeh Nayer El-Gedawy. Using fuzzifiers to solve word sense ambiguation in arabic language. *International Journal of Computer Applications*, 2013.

[4] Mohammad Nasiruddin. A state of the art of word sense induction: A way towards word sense disambiguation for under-resourced languages. *arXiv preprint arXiv:1310.1425*, 2013.