FRI Natural language processing course 2022
Project report

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, because we know from experience 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]. The dataset was constructed from example usages in various lexical resources. It was compiled with constraints of not having more than three instances for some target word and not having repeated contextual sentences across instances.

Ideas for implementation

TODO

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.