FRI Natural language processing course 2023
First report

University of Ljubljana
Faculty of Computer and Information Science

Word Sense Disambiguation

Katarina Aleksandra Brezovar, Klara Vrabl, Meta Kokalj, and Nives Hüll

Abstract
The abstract goes here.

Keywords
Natural Language Processing, NLP, Word Sense Disambiguation, WSD, Clustering, Slovene

Advisors: Aleš Žagar

Introduction

Word sense disambiguation (WSD) is an important task in natural language processing that consists of determining the correct meaning of a word based on its context, using a predetermined list of potential meanings. This process is usually performed unconsciously by humans. WSD can be viewed as a classification problem in which the goal is to assign an occurrence of a word to its appropriate sense class based on a dictionary of possible meanings. The context in which the word occurs, including neighboring words, serves as evidence for the classification task. WSD is applicable in various domains, such as machine translation, information retrieval and hypertext navigation, content and thematic analysis, speech processing, knowledge acquisition, information extraction, etc. [1]

The goal of this project is to prepare a dataset for training WSD models using both automatic and manual methods. The project consists of four steps. To create the dataset, we will first create a list of highly polysemous words from the existing Elexis WSD dataset. Then, sentence pairs containing these words will be extracted from the ccKres corpus using clustering methods and automatic truth value assignment. The sentence pairs will be manually verified, and their truth values corrected as needed. We will also convert the existing Elexis WSD dataset to WiC format and create as many positive and negative examples as possible, which we will then join to the newly generated dataset. The resulting dataset will be used to train WSD models that can determine whether two occurrences of a word in different contexts have the same meaning or not.

Overall, this project aims to contribute to the development of WSD models that can accurately determine the correct meaning of polysemous words in natural language text, which is essential for improving the accuracy of machine translation and other NLP tasks.

Related work

Preliminary challenge in WSD [2] is the ambiguity of language itself. Despite extensive research in this field, there is still no clear understanding of the two main categories of lexically ambiguous words, homonymy and polysemy [3]. Polysemy refers to a single lexical item with more than one semantic specification, while homonymy involves multiple morphological specification with the same sound and/or spelling under different dictionary entries. The word homonym can thus be used for both homophone and homograph, which adds complexity to the issue of homonymy. Therefore, in the task of disambiguation, it is important to have a clear definition of polysemy and consistently follow the chosen approach [4].

Furthermore, there is still not a clear understanding of the difference between the two main types of linear and nonlinear polysemy. Non-linear polysemy mainly consists of metaphor and metonymy, whereas linear polysemy can be further categorized into autohyponymy, automeronymy, autosuperordination, and autoholonymy [3].

Important and frequently observed problem related to a drop in accuracy with WSD [5] is the so-called domain adaptation problem, where the system is trained on one domain but applied to a different domain. The goal of domain adaptation is to train a neural network on one dataset for which label or annotation secure good performance on another dataset from a different domain. Therefore, the challenge is to make classifiers perform well on the target dataset [6].

Current approaches rely heavily on supervised learning techniques. While they have shown promising results in WSD, they require large amounts of data for training, which can be time consuming and costly, additionally they call for annotated data. Further research is needed to develop techniques that at the same time do not rely heavily on annotated data, handle the ambiguity of language, and can be used across domains. As has been previously established, many approaches in different areas of NLP, models that are only trained on a particular domain, usually perform poorly on text from a different domain. To achieve this, semi-supervised and unsupervised approaches that can leverage large amounts of unannotated text would have to be developed [7].

WSD models for Slovene

While there have been many WSD models developed for English and other languages, for Slovene there is still a lack of different WSD models. One of the well-known ones is the one developed by RSDO [8]. Besides this one, there is also the Slovenian version of the parallel-sense annotated corpus ELEXIS-WSD [9]. We decided to use this dataset to help us find highly polysemous words. Elexis is a manually curated and annotated dataset consisting of five annotation layers for 10 European languages, including Slovene. This dataset features five annotation layers, including WSD, which is used to identify highly polysemous words. The Slovene dataset was processed using a highly accurate tool called CLASSLA tagger. Two different POS tagsets were used, which could cause confusion for the taggers. To solve this problem, the detailed tagging guidelines UD-POS [10] for Slovene were consulted. Another problem with this process was the distinction between different categories, such as DET vs PRON and CCONJ vs ADV. In order to obtain more content words, named entity components that are not proper nouns were assigned their appropriate part of speech. Finally, some corrections were made to the lemmatisation, such as manually correcting the lemmatisation of prepositions. Despite these challenges, the tokenisation process was error-free.

WiC dataset

We will also be using the WiC dataset [11] that is based on three lexical resources: WordNet, VerbNet, and Wiktionary. Word-in-Context is a binary classification task that aims to determine whether a word used in two different contexts corresponds to the same meaning or not. The dataset consists of examples with a target word and two sentences containing the target word. Each example is either positive or negative, depending on whether the two sentences have the same meaning of the target word. The dataset was compiled by obtaining all possible positive and negative examples from various sources. The test and development sets were created with the intention of obtaining a diverse and balanced set. Some of the examples were also reserved for testing and development data set, respectively. The remaining examples were used for initial training.

Our working process

We have implemented an unsupervised method for our word disambiguation task. This method does not rely on external sources of knowledge, sense inventories or machine readable dictionaries [**?**]. However, it does require a dataset with labeled word senses and an annotated corpus, both of which we do have.

We first created a `.txt` file with the 250 most polysemous lemmas found in Elexis-WSD. Then we narrowed down the list to 30 words per group member for the purpose of assigning senses to polysemous words. Our candidate words represent the 90 most polysemous words in the Elexis-WSD corpus.

Our next step was to create a `.csv` file that contained all sentences found in ccKres that included polysemous words from each group member's lists. To do that, we used `.xml` files in `ccKres_LEMATIZIRAN`. The files are structured so that the polysemous lemma is in the first column, and the sentence containing it is in the second column. We created one `.csv` file for each group member.

Then we had to vectorize each sentence in each group member `.csv` file, which was done using the `SentenceTransformer` model [**?**]. We embedded the whole sentence and then used PCA to remove columns with less correlation. Then we used the k-means clustering algorithm with a value of k set to 10, which means that we have be dividing the data into 10 distinct clusters, with each cluster corresponding to a potential sense of a polysemous word. We found the centroid of each cluster so that the algorithm could represent the central tendencies of each sense, which will help us in disambiguating the word sense of a given usage by comparing it to the centroids of the potential senses and selecting the one that is closest. Lastly, centroids and sentences containing them have been printed into a `.txt` file for each of the participants.

The next step of the project is manually assessing the algorithm's work using the Elexis-WSD corpus. We will also indicate whether the polysemous word is used in its literal or figurative meaning.

Momentarily we have only tried working with the unsupervised method but would like to implement an unsupervised method such as decision trees, support vector machines, naïve bayes or maximum entropy. For the supervised learning method, a classifier is trained on manually created training data and is used to assign senses to instances of that word. During the testing phase, the classifiers use the learned information to identify the best sense for each occurrence of the word. Overall, supervised approaches tend to be more precise than other methods [**?**]. Besides the mentioned ones, two more classes of methods can be used for word sense disambiguation: knowledge-based and hybrid methods. Semi-supervised WSD methods utilize both annotated and unannotated data. It involves using a small set of labeled data, a larger set of unlabeled data, and a set of classifiers. The algorithm is then applied to both datasets, with the annotated dataset expanding while the unlabeled dataset shrinks until a certain threshold is reached. This algorithm has been shown to achieve high accuracy when applied to smaller datasets, but there are uncertainties involved in selecting parameter values such as pool size and number of iterations. Knowledge based methods avoid the need for large training datasets and exploit the knowledge contained in WordNet, Wikipedia, dictionaries etc. They have an advantage over corpus based algorithms because one does not need an annotated corpus to implement these methods. Corpus based algorithms, however, are more precise than knowledge based ones [**?**].

References

[1] Nancy Ide and Jean Véronis. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998.

[2] Eneko Agirre and Philip Edmonds. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media, 2007.

[3] Alan Cruse. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press UK, 2010.

[4] Restu Resmiyati. The distinction between polysemy and homonymy on lexical ambiguity, 2020.

[5] Yee Seng Chan and Hwee Tou Ng. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 49–56, Prague, Czech Republic, 2007. Association for Computational Linguistics.

[6] Harsh Maheshwari. Understanding domain adaptation.

[7] Krishnanjan Bhattacharjee, S ShivaKarthik, Swati Mehta, Ajai Kumar, Snehal Phatangare, Kirti Pawar, Sneha Ukarande, Disha Wankhede, and Devika Verma. Survey and gap analysis of word sense disambiguation approaches on unstructured texts. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 323–327. IEEE, 2020.

[8] Slavko Žitnik et al. Slowsd. https://github.com/RSDO-DS3/SloWSD, 2022.

[9] Martelli Federico, Navigli Roberto, Krek Simon, Kallas Jelena, Gantar Polona, Veronika Lipp, Tamás Váradi, András Győrffy, and László Simon. Designing the elexis parallel sense-annotated dataset in 10 european languages. 2021.

[10] Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. The universal dependencies treebank for slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 33–38, 2017.

[11] Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.