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

Cross-lingual sense disambiguation

Zala Erič, Miha Debenjak, and Denis Derenda Cizel

Abstract

This is the second report for the Natural Language Processing course. Our ambition is to prepare a dataset that will be used to compare meaning of words in different sentences. So far, the report contains an overview of the field and a basic idea.

Keywords

NLP, sense disambiguation, word-in-context ...

Advisors: Slavko Žitnik

Introduction

In human language, the same word can have different meanings depending on its intended use in a sentence. In the context of natural language processing, word sense discrimination can be described as the ability to determine the meaning of a word used in a given sentence or context. We know word syntactic ambiguity, which can be solved by part-of-speech (POS) taggers, and semantic ambiguity (word sense disambiguation), which is part of our project, and it's harder to resolve than syntactic ambiguity. Word Sense Disambiguation (WSD) has to decide what the sense of a word is based on the word in context and a fixed inventory of potential word senses. This principle can be used in various applications: informational retrieval, knowledge acquisition and information extraction, machine translation, question answering, lexicography. Our goal is to prepare a solution that will be able to determine if the word X is used in the same context or not in 2 different sentences. This can be achieved by using different methods: Supervised Machine Learning, Semi-Supervised Learning, Dictionary Methods...

Related work

To better understand the task itself, we reviewed similar work in the research area. We came across different approaches for solving our problem.

WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representation [1]

Word-in-Context is a dataset made for the binary classification task of determining whether a word is used in the same context in two different sentences. It was built from usage examples from three lexical resources: WordNet, VerbNet and Wiktionary. Each instance of the database consists of the target word and two sentences in which the word is used. The instances can be positive, meaning the word is used in the same context or negative, meaning the words are used in different contexts. It is already divided into the development and test subsets of sizes 1600 and 800 instances respectively. The authors made sure that the splits were balanced in terms of positive and negative instances and that there were at most 3 instances of the same target word, as well as not duplicating the context sentences. As some words have a lot of different meanings, the database does not include senses which are very similar to each other. The database was used with a series of most performing WSD methods, with the BERT method being the best performing on the new dataset with the accuracy of 65,5 %.

SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems [2]

The GLUE benchmark is a single-number metric that summarizes progress on a set of language understanding tasks. Since performance of recent methods (ELMo, OpenAI GPT, BERT) on the GLUE benchmark has recently surpassed the level of non-expert humans, the authors present SuperGLUE, with which they aim to pose a more rigorous test of language understanding. SuperGLUE consists of a public leaderboard built around eight language understanding tasks, drawing on existing data, accompanied by a single-number performance metric and an analysis toolkit. The eight tasks are the following: Boolean Questions, Commitment Bank, Choice of Plausible Alternatives, Multi-Sentence Reading Comprehension, Reading Comprehension with Commonsense Reasoning

Dataset, Recognizing Textual Entailment, Words-in-context, Winograd Schema Challenge, and they are evaluated equally to produce a single-number score. They are compared to a human performance. The authors evaluated BERT-based baselines and found that they still perform worse than humans by nearly 20 points. Given the difficulty of SuperGLUE for BERT, they expect that further progress in multi-task, transfer, and unsupervised/self-supervised learning techniques will be necessary to approach human-level performance on the benchmark.

Knowledge-based Word Sense Disambiguation using Topic Models [3]

Most of the word sense disambiguation systems use the sentence in which the word is used to determine its sense. But we know that the whole text which includes this sentence is very likely to cover a specific topic of some wider domain and the sense of words used in this domain are likely to be similar. That is why Chaplot and Salakhutdinov introduced a new word sense disambiguation method, which incorporates topic modeling to detect the topics used inside the document and use these topics to connect words to their senses. The topic modeling method used in their research is the Latent Dirichlet Allocation (LDA), which is an unsupervised method which we can initialize with specific priors. In this case the LDA method is initialized with priors from the synset set of synonyms from the WordNet lexical database. The model then returns the sense of the word which is the most probable in each document in the corpus. It assumes that all occurrences of a word in a document are used in the same sense, which is an unrealistic assumption.

Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations [4]

This article presents different strategies of integrating pre-trained contextualized word representations for WSD. Contextualized word representations are effective in downstream natural language processing (NLP) tasks, such as question answering, etc., but on word sense disambiguation (WSD), it does not outperform the state-of-the-art approach that uses noncontextualized word embeddings. The contextualized word representation used is BERT (Devlin et al., 2019), a bidirectional transformer encoder model (Vaswani et al., 2017) pre-trained on billions of words of texts. The model is trained on two tasks, masked word and next sentence prediction, where in both, prediction accuracy is determined by the ability of the model to understand the context. The authors describe different techniques of leveraging BERT for WSD, broadly categorized into nearest neighbor matching and linear projection of hidden layers. The tasks for evaluating are the English all-words task, English lexical sample task, and Chinese OntoNotes WSD task. The best results were consistently obtained by linear projection models. The final results of the authors' best BERT WSD models greatly outperformed previous neural WSD models.

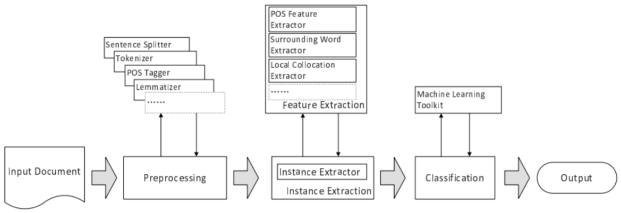

Figure 1. Workflow of the method ITS This is an example of a figure that spans only across one of the two columns.

It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text [5]

It Makes Sense (IMS) is a freely available English all-words Word Sense Disambiguation (WSD) system, built on supervised learning methods. The authors describe that the motivation for such a system is the unavailability of public, open-source WSD systems, useful for applications outside of WSD. They explain the reason for using a supervised learning approach is better performance on SensEval and SemEval tasks, which is what the system was evaluated on. The architecture of the system consists of three independent modules: preprocessing, feature and instance extraction, and classification, see Figure 1. For applications, where WSD is only a component, it is also useful to provide classification models, so they provide training sets. Training instances are collected from sense-annotated corpus SEMCOR (Miller et al., 1994) and DSO corpus (Ng and Lee, 1996), but also extracted from parallel texts, following the approach of Chan and Ng (2005). The system is evaluated with SensEval and SemEval tasks, the benchmark data sets for WSD and in the article, results are provided for lexical-sample tasks, where both the training and test data sets are provided, and all-words tasks where no training data is provided. Overall, IMS achieves good WSD accuracies on both all-words and lexical-sample tasks. The performance of IMS shows that it is a state-of-the-art WSD system.

Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods [6]

This article represents the use of two different approaches for WSD: knowledge-based method and a corpus-based method. Their approach combines various sources of knowledge with both methods. A specification marks method (SM) is used as a knowledge-based method, and a maximum entropy-based method (ME) as a corpus-based method. Three different schemes for combining both approaches are presented, where each of them outperforms each of the methods individually.

Methods

As there is no slovene database similar to the WiC database for evaluation of word sense disambiguation methods, a new database will be constructed. The steps to create this database will be similar to those proposed in the WiC article. The usage examples for target words will be obtained from the sloWnet

database [7], which is a slovene version of the english WordNet database, and the Dictionary of the Slovenian Standard Language. Similar to WiC, the database will consist of positive and negative instances, consisting of the target word and two sentences containing the target word in the same context or in a different context.

According to the articles described in the Related Work section, the BERT method wouble be suitable for our analysis part of project. BERT, which stands for Bidirectional Encoder Representations from Transformers, is an ML framework for NLP. BERT was pretrained on two tasks: language modelling (15 % of tokens were masked and then it was trained to predict them from context) and next sentence prediction (BERT was trained to predict if a chosen next sentence was probable or not given the first sentence). As a result of the training process, BERT learns contextual embeddings for words. After pretraining BERT can be fine-tuned on smaller datasets to optimize its performance on specific language processing tasks. BERT was developed by Google and is also used in Google search engine to optimize the interpretation of user search queries.

Results

We've compiled a list of words that can be used in multiple senses. Our desire is to classify words by their meanings, so we started by collecting words from a dictionary of synonyms. The dictionary of Slovenian synonyms is available at viri.cjvt.si/sopomenke/eng/ and was developed within the Centre for language resources and technologies at the University of Ljubljana. One version of this dictionary is available on the clarin.si website (Common Language Resources and Technology Infrastructure, Slovenia) under the name Thesaurus of Modern Slovene 1.0. It was published in 2018 and is available for download as an xml file. We then selected about 2000 words from the dictionary that could be classified into several different categories/meanings (more than 4).

Next we needed sentences containing words from the list generated earlier. We retrieved them using the Slovenian web corpus by Tomaž Erjavec and Nikola Ljubešić (slWaC 2.1), which we found on the Sketch Engine corpus platform. Using provided API interface we searched for the uses of each word. For each of them we stored 10 use cases. So far, for each use case, we store a total of 3 sentences before the observed word and 3 sentences after the observed word. Due to API data limitations, we have so far only done this for about 500 words. From the extracted sentences, we have also prepared pairs of sentences, which will then be compared with each other using an appropriate algorithm.

We also added some functions which will later be used for text preprocessing. Using the Stanza Python library, we can preprocess text in slovene language. A simple pipeline of tokenization, lemmatization and stop word removal was implemented, which will be later used to prepare the corpus for further analysis.

Discussion

Acknowledgments

References

[1] Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representation. 2019.

[2] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537, 2019.

[3] Devendra Singh Chaplot and Ruslan Salakhutdinov. Knowledge-based Word Sense Disambiguation using Topic Models. 2018.

[4] Hadiwinoto Christian, Ng Hwee Tou, and Wee Chung Gan. Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, page 5297–5306, 2019.

[5] Zhi Zhong and Hwee Tou Ng. It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. *Proceedings of the ACL 2010 System Demonstrations*, pages 78—-83, 2010.

[6] Andres Montoyo, Armando Suarez, German Rigau, and Manuel Palomar. Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods. *Journal of Artificial Intelligence Research*, 23:299–330, 2005.

[7] Darja Fier and Jernej Novak. slownet 3.0: development, extension and cleaning. 2011.