University of Ljubljana
Faculty of Computer and
Information Science

David Miškić, Kim Ana Badovinac, Sabina Matjašič

Abstract

Keywords
sense, disambiguation

Advisors: Slavko Žitnik

Introduction

Many words have multiple meanings, which is determined by their context. We can determine the specific meaning with the help of word-sense disambiguation (WSD). The problem has been long active in the field of natural language processing, as those languages are context-dependent and understanding is not possible without knowing the meaning [1]. The hardness of the problem has been described as equivalent to solving central AI problems [2]. WSD relies on knowledge, but those can vary considerably depending on corpora, and there are different approaches with their own standardization. But the need for automatic knowledge is growing with the amount of unstructured data and use of machine translation. There are multiple approaches we will explore in the following section, but one of those aims to automatically disambiguate a text in one language by exploiting its differences with another language in a parallel corpus. This is cross-lingual word sense disambiguation (CLWSD).

Related Work

Manual encoding

Wilks, 1972

Preference semantics assumed that each sense has a formula that could determine the meaning. The formula was composed of hierarchical selectional constraints, but those could be loosened within some conditions. In addition, verbs and adjectives listed the preferred sense of the words they modified. Disambiguation involved choosing a formula that was satisfied the most.

One example is the use of the verb to drink in sentences ”The adder drank from the pool” and ”My car drinks gasoline”. The verb prefers animate subject. In the first case, the adder can have the meaning of machine or animate object, so the animate meaning satisfies the verb's condition. In the second case, the car is not animate so it does not fit the verb. The meaning of the verb has to be extended and a new usage is added. In this way, the meanings can be dynamically evolved over the text [1].

Small 1980

Word expert parsing presumes the lexicon to be the source of information, but the representation is build up from disambiguation as well. Small's unconfirmed hypothesis was that human knowledge is more word-based rather than rule-based, so each such word would need its own disambiguator [1].

Dictionary based

Lesk 1986

Optimizing dictionary definition overlap with the help of the dictionary, where the sense can be determined depending on the sense of the words close to it, based on dictionary definitions. One example is the word ”pine cone” where pine is either a tree or a verb, while the cone is either a body or a fruit. The overlap is in the words evergreen and a tree in the dictionary definition, so those meaning were picked (pine as tree and cone as a fruit) [1].

Connectionist based

Waltz and Pollack 1985

The approach was to carry out syntactic and semantic processing in parallel, based on psycholinguistics observation. Such an approach attempted to model the errors as well, one example is the sentence ”The astronomer married the star” where humans make a temporary mistake of regarding the word star as a celestial body before the more probable sense of celebrity is picked because the semantic effect of the word astronomer is stronger than case selection of the verb to marry. The model is composed of a network of nodes, representing case frames, semantic priming, and syntactic preferences [1].

Cross Lingual

Sense projection (SP) approach can bootstrap the creation of sense-annotated parallel corpora by exploiting existing resources in well-represented languages, with word alignment and connected sense inventories as the only requirements. [3]

Unlike SP, multilingual sense intersection (SI) approach does not require any of the texts in a parallel corpus to be sense-annotated, so it can be applied to a wider range of existing resources. Its logical foundation is in that a polysemous word in a language is likely to be translated into different words in other languages, so the comparison with the semantic space of each translation help select the sense actually intended. [3]

Idea

Idea is to explore two approaches of annotating a multilingual parallel corpus as authors Bond and Bonasasinga described in the article [3]. Both approaches can be applied to any multilingual parallel corpus, as long as large inter-linked sense inventories exist for both of the languages involved. We intend to use Slovenian and English language.

The first approach is sense projection which uses word alignment as a bridge. Assuming that translations preserve the meaning of a text, if a sense annotated source text is aligned to its translation, then the annotations can be transferred, as long an inter-linked sense inventory is used by all languages.

The second approach is multilingual sense intersection (SI). Given an ambiguous target word, in the SI approach, each of its aligned translations in the parallel sentences contributes to the disambiguation process by bringing in all its ”set of senses” retrieved from the inter-linked sense inventory. The intersection is then performed over each nonempty set retrieved. If the overlap only consists of one sense, then the target word is disambiguated. If the overlap contains more than one sense, then it is further intersected with the set of most frequent senses available for the target lemma.

Data

We need text where we can rely on a more literal translation instead of a faithful one, so books, which aim to translate meaning and not the text itself, are less appropriate. One such source can be official documents, the selection of Slovenian laws and regulations with English translation is available at http://www.pisrs.si/Pis.web/cm?idStrani=prevodi.

References

[1] *The Oxford handbook of computational linguistics*. Oxford University Press, 2004. Bibliografija ob posameznih poglavjih Kazali.

[2] Roberto Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2), feb 2009.

[3] Francis Bond and Giulia Bonansinga. *Exploring Cross-Lingual Sense Mapping in a Multilingual Parallel Corpus*, pages 56–61. 01 2015.