Leveraging Large Language Models to Build a Cutting-Edge French Word Sense Disambiguation Corpus

Authors : Mouheb Mehdoui, Amel Fraisse, Mounir Zrigui

Abstract : With the increasing amount of data circulating over the Web, there is a growing need to develop and deploy tools aimed at unraveling semantic nuances within text or sentences. The challenges in extracting precise meanings arise from the complexity of natural language, while words usually have multiple interpretations depending on the context. The challenge of precisely interpreting words within a given context is what the task of Word Sense Disambiguation meets. It is a very old domain within the area of Natural Language Processing aimed at determining a word's meaning that it is going to carry in a particular context, hence increasing the correctness of applications processing the language. Numerous linguistic resources are accessible online, including WordNet, thesauri, and dictionaries, enabling exploration of diverse contextual meanings. However, several limitations persist. These include the scarcity of resources for certain languages, a limited number of examples within corpora, and the challenge of accurately detecting the topic or context covered by text, which significantly impacts word sense disambiguation. This paper will discuss the different approaches to WSD and review corpora available for this task. We will contrast these approaches, highlighting the limitations, which will allow us to build a corpus in French, targeted for WSD.