cTAKES 3.1.2 - SenseDisambiguatorAnnotator

Introduction

Terms in a natural language may be ambiguous, i.e. can be mapped to multiple distinct concepts. For example, the word ‘cold’ can refer to the viral infection ‘common cold’ or the ‘sensation of cold’. YTEX implements the 'adapted lesk' method that uses semantic similarity measures to quantify how well a concept ‘fits’ in a given context. This page describes the WSD algorithm, the configuration for the SenseDisambiguatorAnnotator, and describes how to reproduce the results of our evaluation on the NLM WSD and MSH WSD data sets.

The adapted Lesk algorithm works as follows: for each term to be disambiguated (target term) it finds all possible senses; it selects all words within a window surrounding the target term and maps them to concepts (context concepts); and it scores senses by summing the semantic relatedness between each sense and all context concepts; and finally it selects the sense with the highest score. In its original formulation, the adapted Lesk calculates the relatedness of concepts using the cosine of their profile vectors; however, this can be replaced with a semantic similarity measure.

For a high-level overview of the WSD method we've implemented, refer to our paper: Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification.

Note that you must perform the additional YTEX installation tasks to use this component.

SenseDisambiguatorAnnotator

The SenseDisambiguatorAnnotator is an UIMA annotator integrated with cTAKES. cTAKES identifes named entities (EntityMention Annotations), which in turn can contain multiple concepts (OntologyConcept Feature Structures). The SenseDisambiguatorAnnotator disambiguates each ambiguous term (i.e. EntityMention with multiple OntologyConcepts) in a document as follows:

- Takes all EntityMentions within a window around the ambiguous term
- Scores candidate concepts using the semantic similarity with context concepts; the score is stored in the score attribute of the OntologyConcept.
- Picks the candidate concept with the highest score: sets the OntologyConcept.disambiguated attribute to true for the best concept, and false for others.

The SenseDisambiguatorAnnotator is configured via CTAKES_HOME/resources/org/apache/ctakes/ytex/ytex.properties:

- ytex.sense.windowSize - context window size. concepts from named entities +- windowSize around the target named entity are used for disambiguation. defaults to 50
- ytex.sense.metric - measure to use. defaults to INTRINSIC_PATH. See Semantic Similarity for valid values.
- ytex.conceptGraph - concept graph to use. Defaults to sct-rxnorm (SNOMED-CT, RXNORM).

The optimal measure and concept graph depends on the application. These defaults achieved the best score on the MSH WSD data set; you might want to experiment with the LCH measure and umls concept graph: this configuration achieved the best performance on the NLM WSD data set.