

SenseClusters

Clustering Similar Contexts in Natural Language Text

Project Information: <http://senseclusters.sourceforge.net>

Source Code Download: <http://sourceforge.net/projects/senseclusters>

Overview

SenseClusters is a word sense discrimination package that differentiates among the possible meanings of an ambiguous word. For example, given various contexts in which *sharp* is used, it separates the contexts that refer to intellectual sharpness versus those that refer to cutting edge sharpness or sharp criticism. SenseClusters is based on the strong contextual hypothesis of Miller and Charles (1991) which states that "two words are semantically related to the extent that their contextual representations are similar." Hence, the overall objective of this package is to identify contextually similar text units.

Features

- Uses a purely unsupervised, knowledge-lean clustering approach that does not rely on any external knowledge resources like dictionaries, ontologies or annotated corpora.
- Selects lexical features like n-grams and co-occurrences using statistical measures of association such as the log-likelihood ratio, mutual information, chi-squared test, etc.
- Represents contexts of ambiguous words using 1st and 2nd order context vectors.
- Measures pair-wise similarities among contexts using cosine, matching, overlap, Jaccard, and Dice coefficients.
- Resolves polysemy and synonymy in contexts by performing dimensionality reduction via Singular Value Decomposition.
- Supports a variety of clustering algorithms via CLUTO, the Clustering Toolkit.
- Performs extensive evaluation of discovered clusters using internal and external evaluation metrics.
- Open source code that is freely distributed under the GNU Public License (GPL).

Applications

- Synonymy Identification
- Email Classification
- Text Summarization
- Word Sense Disambiguation

Visit us at AAI-04

- I.S. Demo : Tues, July 27 – Weds, July 28 at Table #212 (10:00 am – 6:00 pm)
- Student Poster Session : Tues, July 27 (3:00 pm – 5:30 pm)

Amruta Purandare (pura0010@d.umn.edu) and Ted Pedersen (tpederse@d.umn.edu)
University of Minnesota, Duluth