



SenseClusters - Clustering similar contexts ...

Amruta Purandare & Ted Pedersen

Project Webpage – <http://senseclusters.sourceforge.net>

Overview

Words in natural languages are often ambiguous and refer to different meanings when used in different contexts. SenseClusters achieves word sense discrimination by using the hypothesis that semantically related word meanings tend to occur in similar contexts. Thus, it solves the problem of discriminating among the possible meanings of an ambiguous word by clustering the contextually similar usages of that word.

Features

- Uses a purely unsupervised clustering approach that does not rely on any external knowledge sources like dictionaries, ontologies or sense-tagged corpora. Instances are clustered based on their mutual similarities that can be computed from the raw corpus itself.
- Supports various lexical features like n-grams and co-occurrences that can be selected by performing statistical tests of association like log-likelihood, mutual information, chi-squared test etc.
- Implements two types of context vector representations referred to as the 1st order and 2nd order context vectors.
- Supports a similarity space representation of the contexts by recording their pair-wise similarities in a similarity matrix structure.
- Allows feature space dimensionality reduction to address the problems like polysemy and synonymy in languages via an interface to SVDPACK.
- Supports variety of clustering algorithms by providing a seamless interface to the Clustering Toolkit, CLUTO.
- Performs extensive evaluation using external as well as internal evaluation metrics.
- Open source software that is freely distributed under the GNU Public License (GPL).

Applications

- Synonymy Identification
- Automatic Email Foldering/Indexing
- Text Summarization
- Word Sense Disambiguation

By: Amruta Purandare (pura0010@d.umn.edu) and Ted Pedersen (tpederse@d.umn.edu)