Sanjana Agarwal

ASSIGNMENT 2 - MACHINE LEARNING IN COMPUTATIONAL LINGUISTICS

1)
Title: Random Walks for Knowledge-Based Word Sense Disambiguation
Authors: Eneko Agirre, Oier Lopez de Lacalle, Aitor Soroa
Year: 2011
Conference/Journal: Volume 40, Number 1
Issue/Conference Venue: 2014 Association for Computational Linguistics
Pages: 57-84
Algorithm used: Personalized Page rank with Lexical Knowledge Base

Summary:
They have built a Word Sense Disambiguation (WSD) method based on random walks over large Lexical Knowledge Bases (LKB). The PageRank random walk algorithm and LKB combination compares favorably to the state-of-the-art in knowledge-based WSD on a wide variety of data sets, including four English and one Spanish data set. They have used the full WordNet graph.

2)
Title: A Large-Scale Pseudoword-Based Evaluation Framework for State-of-the-Art Word Sense Disambiguation
Authors: Mohammad Taher Pilehvar, Roberto Navigli
Year: 2013
Conference/Journal: Volume 40, Number 4
Issue/Conference Venue: Publications by Affiliated Organizations: Computational Linguistics
Pages: 837-881
ML Algorithm used: Supervised and knowledge based WSD

Summary:
They have created new types of artificial words that model real words by preserving their semantics as much as possible. These semantically aware pseudowords can be used to model any word in the lexicon. However, they have focused only on nouns. They leveraged these pseudowords to create a large-scale evaluation framework for WSD. Using this framework they performed an experimental comparison of state-of-the-art systems for supervised and knowledge-based WSD on a very large data set made up of millions of sense-tagged sentences.

3)
Title: Entity Linking meets Word Sense Disambiguation: a Unified Approach
Authors: Andrea Moro, Alessandro Raganato, Roberto Navigli
Year: 2014
Conference/Journal: Volume 2
Issue/Conference Venue: Transactions of the Association for Computational Linguistics
Pages: 231-244
ML Algorithm used: Graphs

Summary:
They disambiguate and link all nominal and named entity mentions occurring within a text. The linking task is performed by associating each mention with the most suitable entry of a given knowledge base. So they have made a system called: Babelfy. Given a lexicalized semantic network, they associate with each vertex, i.e., either concept or named entity, a semantic signature, that is, a set of related vertices. Given a text, they extract all the linkable fragments from this text and, for each of them, list the possible meanings according to the semantic network. Then they create a graph-based semantic interpretation of the whole text by linking the candidate meanings of the extracted fragments using the previously-computed semantic signatures. Finally, they extract a dense subgraph of this representation and select the best candidate meaning for each fragment.

4)
Title: An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model
Authors: Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro
Year: 2014
Issue/Conference Venue: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics
Pages: 1591–1600
Non-ML Algorithm used: Lesk Algorithm

Summary:
A Word Sense Disambiguation (WSD) algorithm that extends two well-known variations of the Lesk WSD method. Given a word and its context, Lesk algorithm exploits the idea of maximum number of shared words (maximum overlaps) between the context of a word and each definition of its senses (gloss) in order to select the proper meaning. The main part relies on the use of a word similarity function defined on a distributional semantic space to compute the gloss-context overlap.

5)
Title: A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation
Authors: Ted Pedersen
Year: 2000
Conference/Journal: NAACL 2000 Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference
Pages: 63-69
ML Algorithm used: Naïve Bayes

Summary:
This paper presents a corpus-based approach that results in high accuracy by combining a number of very simple classifiers into an ensemble that performs disambiguation via a majority vote. The contextual features used in this paper are binary and indicate if a given word occurs within some number of words to the left or right of the ambiguous word. No additional positional information is contained in these features; they simply indicate if the word occurs within some number of surrounding words. The context is a variation on the bag-of-words feature set, where a single window of context includes words that occur to both the left and right of the ambiguous word. Once all the parameters have been estimated, the model has been trained and can be used as a classifier to perform disambiguation by determining the most probable sense for an ambiguous word, given the context in which it occurs.

6)
Title: Word Sense Disambiguation for Cross-Language Information Retrieval
Authors: Mary Xiaoyong Liu, Ted Diamond, and Anne R. Diekema
Year: 2000
Conference/JournalACM International Conference Proceeding Series, Proceedings of the Workshop on Student Research
Issue/Conference Venue: Volume 5
Pages: 35-40
ML Algorithm used: Bayes Theorem, WSD Algorithm

Summary:
This conceptual interlingua is a hierarchically organized multilingual concept lexicon, which is structured following WordNet. By representing query and document terms by their WordNet synset numbers they arrived at essentially a language neutral representation consisting of synset numbers representing concepts. This representation facilitates cross-language retrieval by matching tea-m synonyms in English as well as across languages. However, many terms are polysemous and belong to multiple synsets, resulting in spurious matches in retrieval. Their algorithm aimed at improving the precision of a cross-language retrieval system.