

Data for Paper

An Iterative 'Sudoku Style' Approach to Subgraph-based Word Sense Disambiguation

Last Updated

23rd July 2014

Description

Where appropriate, the following are provided:

- Output files (.txt) that contain system disambiguations
- Graph files (.dot) to visually see the subgraphs
- Path Index files to create the subgraphs using the BabelNet (or Daebak) API for both Subtree and Shortest paths
- The necessary files for SemEval 2013 Task 12 – Multilingual WSD

If you need more files or resources to reproduce the results of the aforementioned paper, please contact me (Steve L. Manion) at via my homepage – <http://stevemanion.com>. Also any questions regarding how the experiments were completed are welcome. Over time it is likely I will add more information to this file, along with extra data, based on feedback and questions.

For all experiments, BabelNet 1.1.1 was used, with path indexes also generated from this version. This BabelNet version was chosen to put results on equal footing with submissions made in SemEval 2013 Task 12 - Multilingual Word Sense Disambiguation.

Experiment 1 – Proof of Concept

Directory Structure

- Document Level (For Table 1)
 - Shortest Paths
 - Conventional Approach
 - Iterative Approach
 - SubTree Paths
 - Conventional Approach
 - Iterative Approach
- Sentence Level (For Table 2)
 - Shortest Paths
 - Conventional Approach
 - Iterative Approach
 - SubTree Paths
 - Conventional Approach

- Iterative Approach
- Cup Example (For Figure 5)

Important Notes

- For PageRank, alpha was set to 0.15, to ensure a damping factor of 0.85 (see <http://jung.sourceforge.net/doc/api/edu/uci/ics/jung/algorithms/scoring/PageRank.html>)
- All subgraphs are directed

Experiment 2 – Performance

Important Notes

- For PageRank, the BabelNet API version is used. Contrary to the previous experiment, alpha is set to 0.85 and the damping factor is 0.15 (check with `pr.getAlpha()`).
- The stats of the laptop these experiments were run on include:
 - Toshiba Satellite
 - Intel Core i7-3610QM CPU @ 2.30GHz
 - 16GB RAM
 - 64 bit OS

Experiment 3 – A Little Optimisation

Important Notes

- For SUDOKU_It-PPR[M](+), alpha is set to 0.4, to ensure a damping factor of 0.6. This alpha value produced the best result, for (It-)PR[U] runs alpha is set to 0.15
- The specific surfing vector used in this experiment, that is bias towards monosemy is called `dynamic_mono` (this is in the Daebak API), regular PageRank will of course have a uniform surfing vector
- Some extra runs are included in the folder titled “Extra runs”, they are not included in the paper but may be of interest

Path Indexes

The path indexes only hold the necessary paths required for the SemEval 2013 Task 12 English dataset and no more. Maximum path length is L=2, as used in the experiments. Make sure you select the right index files based on whether:

- you will disambiguate at the document or the sentence level
- you will construct shortest path or subtree subgraphs

Task

Follow the link below to download the test data set for SemEval 2013 Task 12. Note that the submissions from other teams are also included in this download:

<http://www.cs.york.ac.uk/semeval-2013/task12/data/uploads/datasets/semeval-2013-task12-test-data.zip>