Reducing polysemy in WordNet

Kanjana Jiamjitvanich, Mikalai Yatskevich

Department of Information and Communication Technology,
University of Trento, Italy
kanjana@disi.unitn.it, yatskevi@disi.unitn.it

1 WordNet

WordNet [4] is the lexical database for English language. A synset is a WordNet structure for storing senses of the words. Synset contains a set of synonym words and their brief description called gloss. For example, *well*, *wellspring* and *fountainhead* have the same meaning according to WordNet, so these three words are grouped in to one synset which is explained by a gloss "*an abundant source*".

A known problem of WordNet is that it is too fine-grained in its sense definitions. For instance, it does not distinguish between homographs (words that have the same spelling and different meanings) and polysemes (words that have related meanings). We propose to distinguish only between polysemes within WordNet while merging all homograph synsets. The ultimate goal is to compute a more coarse-grained version of linguistic database.

2 Meta matcher

Meta matcher is designed as a WordNet matcher, i.e., a matcher that is effective in matching WordNet with itself. It utilizes extensible set of element level matchers (see [1] for extensive discussion) and combines their results in hybrid manner, i.e., the final score is computed from the scores of independently executed matchers.

We implemented three element level matchers.

WordNet relation matcher (WNR). WNR takes two senses as an input and obtains two sets of senses connected to input senses by a given relation. Then these two sets are compared exploiting well-known Dice coefficient formula.

Part of speech context (POSC). POSC matcher exploits part of speech (POS) and sense tagged corpora for similarity computation. In particular, for each WordNet sense occurrence within corpora a set POS tags in the immediate vicinity of sense is memorized. Given multiple occurrence of a sense within corpora each sense is associated with a set of POS contexts. Then, the similarity between two senses is computed as set similarity between sets of POS contexts associated with them.

Inverted sense index inexact (ISII). ISII matcher exploits sense tagged WordNet 3.0 glosses for similarity computation. In particular, for each WordNet sense occurrence within sense tagged glosses, the synset of a tagged gloss is memorized. Than, senses are compared by comparing sets of synsets associated with them. We compare synsets exploiting well known Resnik similarity measure [6].

Matching process is organized in two steps.

2.1 Element level matchers threshold learning

The necessary prerequisite for this step is a training dataset or (a part of) the matching task for which human alignment *H* is known. All element level matchers then are executed on the training dataset, i.e., we obtain complete set of correspondences *M* for all matchers. Then the threshold learning procedure is executed. It performs exhaustive search through all threshold combinations for all element level matchers. Thus, we can select threshold that maximizes a given matching quality metric, e.g., Recall.

In the case of several matchers system result set *S* is obtained from their results through a combination strategy, namely a function that takes matchers results in input and produces a binary decision of whether the given correspondence holds. In this paper we used union of all matchers results as a combination strategy, i.e., if a given correspondence is returned by at least one matcher it is included in *S*.

2.2 Hybrid matching

On this step meta matcher is executed on testing dataset. Element level matchers results are combined using thresholds and the combination strategy exploited in the previous step. For union combination strategy positive result is produced only if confidence score, as computed by element level matchers, is higher than threshold learned on the previous step.

3 Evaluation results

We used a dataset exploited in SemEval[1] evaluation. The dataset contains 1108 nouns, 591 verbs, 262 adjectives and 208 adverbs. We split it into two equal parts: training and testing datasets.

We compared results of meta matcher with 3 other sense merging methods. In particular, we re-implemented sense merging algorithm [2], Genclust algorithm [5] and MiMo algorithm [3]. Meta matcher outperforms the other methods in terms of F-Measure.

References

1. F. Giunchiglia and M. Yatskevich. Element level semantic matching. In The Semantic Web: ISWC 2004: Third International Semantic Web Conference: Proceedings, 2004.
2. W. Meng, R. Hemayati, and C. Yu. Semantic-based grouping of search engine results using wordnet. In 9th Asia-Pacific Web Conference (AP-Web/WAIM'07), 2007.
3. R. Mihalcea and D. Moldovan. Automatic generation of a coarse-grained wordnet. In NAACL Workshop on WordNet, 2001.
4. G. Miller. WordNet: An electronic Lexical Database. MIT Press, 1998.
5. W. Peters, I. Peters, and P. Vossen. Automatic sense clustering in eurowordnet. In Proceedings of LREC'1998, 1998.
6. P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, 11:95{130, 1999.

[1] http://lcl.di.uniroma1.it/coarse-grained-aw/index.html