UoS: A Graph-Based System for Graded Word Sense Induction

David Hope, Bill Keller
University of Sussex
Cognitive and Language Processing Systems Group
Brighton, Sussex, UK
davehope@gmail.com, billk@sussex.ac.uk

Abstract

This paper presents UoS, a graph-based Word Sense Induction system which attempts to find all applicable senses of a target word given its context, grading each sense according to its suitability to the context. Senses of a target word are induced through use of a non-parameterised, linear-time clustering algorithm that returns maximal quasi-strongly connected components of a target word graph in which vertex pairs are assigned to the same cluster if either vertex has the highest edge weight to the other. UoS participated in SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses. Two system were submitted; both systems returned results comparable with those of the best performing systems.

1 Introduction

Word Sense Induction (WSI) is the task of automatically discovering word senses from text. In principle, WSI avoids reliance on a pre-defined sense inventory.[1] Whereas the related task of Word Sense Disambiguation (WSD) can only assign pre-defined senses to words on the basis of context, WSI follows the dictum that "*The meaning of a word is its use in the language.*" (Wittgenstein, 1953) to discover senses through examination of context of use in large text corpora. WSI, therefore, may be applied to discover new, rare, or domain specific senses; senses undefined in existing sense inventories.[2]

Previous WSI evaluations (Agirre and Soroa, 2007; Manandhar et al., 2010) have approached sense induction in terms of finding the single most salient sense of a target word given its context. However, as shown in Erk and McCarthy (2009), a graded notion of sense may be more applicable, as multiple senses of the target word may be perceived by readers. The SemEval-2013 WSI evaluation described in this paper is designed to explore the possibility of finding all perceived senses of a target word in a single contextual instance. The aim for participants in the task is therefore to design a system that will induce a set of graded (weighted) senses of a target word in a particular context.

The paper is organised as follows: Section 2 introduces SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses; Section 3 presents UoS, the system that participated in the task; Section 4 reports evaluation results, showing that UoS returns scores comparable with those of the best performing systems.

2 SemEval-2013 Task 13

2.1 Aim

The aim for participants in SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses is to construct a system that will: (1) induce the senses of a given set of target words and (2), label each test set context (instance) of a target word with

[1] In practice, evaluation of a WSI system requires the use of a gold standard sense inventory such as WordNet (Miller et al., 1990) or OntoNotes (Hovy et al., 2006).

[2] Surveys of WSI and WSD approaches are found in Navigli (2009) and Navigli (2012).

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 689–694, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

all applicable target word senses. Candidate senses are drawn from the WordNet 3.1 sense inventory. Systems must therefore return a set of graded senses for each target word in a particular context, where a numeric weight signifies (grades) each sense's applicability to the context. A non-graded sense is simply the highest graded (weighted) sense out of all graded senses.

2.2 Test Set

The test set consists of 4806 instances of 50 target words: 20 verbs (1901 instances), 20 nouns (1908), and 10 adjectives (997).[3] Instances are extracted from the Open American National Corpus, being a mix of both written and spoken contexts of target words.[4] Only 542 instances are assigned more than one sense by annotators, thus have graded senses. This figure somewhat detracts from the task's aim as just 11.62% of the test set can be assigned graded senses.

2.3 Evaluation Measures

Systems are evaluated in two ways: (1) in a WSD task and (2), a clustering task. In the first evaluation, systems are assessed by their ability to correctly identify which WordNet 3.1 senses of the target word are applicable in a given instance, and to quantify, and so, rank, senses according to their level of applicability. The supervised evaluation method of previous SemEval WSI tasks (Agirre and Soroa, 2007; Manandhar et al., 2010) is applied to map induced senses to WordNet 3.1 senses, with the mapping function of Jurgens (2012) used to account for the applicability weights. Three evaluation metrics are used -

- *Jaccard Index*: measures the overlap between gold standard senses and those returned by a WSI system.

- *Positionally-Weighted Kendall's Tau*: measures the ability of a system to rank senses by their applicability.

- *Weighted Normalized Discounted Cumulative Gain (NDCG)*: measures the agreement in applicability ratings, accounting for both the ranking and difference in weights assigned to senses.

In the second evaluation, similarity between a participant's clustering solution and that of the gold standard set of senses is measured using two metrics -

- *Fuzzy Normalised Mutual Information* (NMI): extends the method of Lancichinetti et al. (2009) to compute NMI between overlapping (fuzzy) clusters. Fuzzy NMI measures the alignment of system and gold standard senses independently of the cluster sizes, so returns a measure of how well a WSI system would perform regardless of the sense distribution in a corpus.

- *Fuzzy B-Cubed*: adapts the overlapping B-Cubed measure defined in Amigó et al. (2009) to the fuzzy clustering setting. As an item-based, rather than cluster-based, measure, Fuzzy B-Cubed is sensitive to cluster size skew, thus captures the expected performance of a WSI system on a new corpus where the sense distribution is the same.

3 The UoS System

The UoS system uses a graph-based model of word co-occurrence to induce target word senses as follows:

3.1 Constructing a Target Word Graph

A graph $G = (V, E)$ is constructed for each target word. V is a set of vertices and $E \subseteq V \times V$ a set of edges. Each vertex $v \in V$ represents a word found in a dependency relation with the target word. Words are extracted from the dependency-parsed version of ukWaC (Ferraresi et al., 2008). In this evaluation V consists of the 300 highest ranked dependency relation words.[5] Words are ranked using the Normalised Pointwise Mutual Information

[3] Stated as 4664 instances on the task website. Note that the figure of 4806 is for the revised test set.

[4] `http://www.americannationalcorpus.org/OANC/index.html`.

[5] $|V| = 300$ was found to return the best results on the trial set over the range $|V| = [100, 200, 300, ..., 1000]$.

(NPMI) measure (Bouma, 2009)[6], defined for two words w_1, w_2 as:

$$NPMI(w_1, w_2) = \frac{\left(\log \frac{p(w_1, w_2)}{p(w_1)\, p(w_2)}\right)}{-\log p(w_1, w_2)}. \qquad (1)$$

An edge $(v_i, v_j) \in E$ is a pair of vertices. An edge represents a symmetrical relationship between vertices v_i and v_j; here, that words w_i and w_j co-occur in ukWaC contexts. Each edge (v_i, v_j) is assigned a weight $w(v_i, v_j)$ to quantify the significance of w_i, w_j co-occurrence, the weight being the value returned by $NPMI(w_i, w_j)$.

3.2 Clustering the Target Word Graph

A clustering algorithm is applied to the target word graph, partitioning it to a set of clusters. Each set of words in a cluster is taken to represent a sense of the target word. The clustering algorithm applied is MaxMax, a non-parameterised, linear-time algorithm shown to return good results in previous WSI evaluations (Hope and Keller, 2013). MaxMax transforms the weighted, undirected target word graph G into an unweighted, directed graph G', where edge direction in G' indicates a *maximal affinity* relationship between two vertices. A vertex v_i is said to have maximal affinity to a vertex v_j if the edge weight $w(v_i, v_j)$ is maximal amongst the weights of all edges incident on v_i. Clusters are identified by finding root vertices of quasi-strongly connected (QSC) subgraphs in G' (Thulasiraman and Swamy, 1992). A directed subgraph is said to be QSC if, for any vertices v_i and v_j, there is a root vertex v_k (not necessarily distinct from v_i and v_j) with a directed path from v_k to v_i and a directed path from v_k to v_j.[7]

3.3 Merging Clusters

MaxMax tends to generate many fine-grained sense clusters. Clusters are therefore merged using two measures: *cohesion* and *separation* (Tan et al., 2006). The cohesion of a cluster C_i is defined as:

$$cohesion(C_i) = \frac{\sum_{\substack{x \in C_i, \\ y \in C_i}} w(x, y)}{|C_i|}. \qquad (2)$$

Separation between two clusters C_i, C_j is defined as:

$$separation(C_i, C_j) = 1 - \left(\frac{\sum_{\substack{x \in C_i, \\ y \in C_j}} w(x, y)}{|C_i| \times |C_j|}\right). \qquad (3)$$

Cluster pairs with high cohesion and low separation are merged, the intuition being that words in such pairs will retain a relatively high degree of semantic similarity. High cohesion is defined as greater than average cohesion. Low separation is defined as a reciprocal relationship between two clusters: if a cluster C_i has the lowest separation to a cluster C_j (out of all clusters) and C_j the lowest separation to C_i, then the two (high cohesion) clusters are merged.[8]

3.4 Assigning Graded Word Senses to Target Words

Each test instance is labelled with graded senses of the target word. A score is computed for the test instance and each target word cluster as the reciprocal of the separation measure, where C_i is the set of content words in the instance (nouns, verbs, adjectives, and adverbs, minus the target word itself) and C_j, the words in the cluster. The cluster with the lowest separation score is taken to be the most salient sense of the target word, with all other positive separation scores taken to be perceived, graded senses of the target word in that particular instance.

4 Evaluation Results

Two sets of results were submitted. The first, UoS (top 3), returns the three highest scoring senses for each instance; the second, UoS (# WN senses), returns the $n =$ number of target word senses in WordNet 3.1 most cohesive clusters, as defined by Equation (2).

Results for the seven participating WSI systems are reported in Tables 1 and 2. The ten baselines, provided by the organisers of the task, are -

[6] Application of the Log Likelihood Ratio measure (Dunning, 1993) returned the same set of words. Though not required here, *NPMI* has the useful properties that: if w_1 and w_2 always co-occur *NPMI* $= 1$; if w_1 and w_2 are distributed as expected under independence *NPMI* $= 0$, and if w_1 and w_2 never occur together, *NPMI* $= -1$.

[7] MaxMax is described in detail in Hope and Keller (2013).

[8] The average number of WordNet 3.1 senses for target words is 8.58. MaxMax returns an average of 59.54 clusters for target words; merging results in an average of 21.86 clusters.

System/*Baseline*	**Jaccard Index F-Score**	**Positionally Weighted Tau F-Score**	**Weighted NDCG F-Score**
UoS (top 3)	0.232	0.625	0.374
AI-KU (r5-a1000)	0.244	0.642	0.332
AI-KU	0.197	0.620	0.387
Unimelb (50k)	0.213	0.620	0.371
Unimelb (5p)	0.218	0.614	0.365
UoS (# WN senses)	0.192	0.596	0.315
AI-KU (a1000)	0.197	0.606	0.215
Most Frequent Sense	**0.552**	0.560	**0.718**
Senses Eq. Weighted	0.149	**0.787**	0.436
Senses, Avg. Weight	0.187	0.613	0.499
One sense	0.192	0.609	0.288
1 of 2 random senses	0.220	0.627	0.287
1 of 3 random senses	0.244	0.633	0.287
1 of n random senses	0.290	0.638	0.286
1 sense per instance	0.000	0.945	0.000
SemCor, MFS	0.455	0.465	0.339
SemCor, All Senses	0.149	0.559	0.489

Table 1: Results for the WSD evaluation: all instances.

- *SemCor, Most Frequent Sense (MFS)*: labels each instance with the MFS in SemCor.[9]

- *SemCor, All Senses*: labels each instance with all SemCor senses, weighting each according to its frequency in SemCor.

- *1 sense per instance*: labels each instance with a unique induced sense, equivalent to the *1 cluster per instance* baseline of the SemEval-2010 WSI task (Manandhar et al., 2010).

- *One sense*: labels each instance with the same induced sense, equivalent to the *MFS* baseline of the SemEval-2010 WSI task.

- *Most Frequent Sense*: labels each instance with the sense that is most frequently selected by annotators for all target word instances.

- *Senses Avg.Weighted*: labels each instance with all senses. Each sense is scored according to its average applicability rating from the gold standard labelling.

- *Senses Eq. Weighted*: labels each instance with all senses, equally weighted.

- *1 of 2 random senses*: labels each instance with one of two randomly selected induced senses.

- *1 of 3 random senses*: labels each instance with one of three randomly selected induced senses.

- *1 of n random senses*: labels each instance with one of *n* randomly selected induced senses, where *n* is the number of senses for the target word in WordNet 3.1.[10]

As noted by the task's organisers[11], the *SemCor* scores are the fairest baselines for participating systems to compare against as they have no knowledge of the test set sense distribution; the other baselines are more challenging as they have knowledge of the test set sense distribution and annotator grading.

4.1 Summary Analysis of Evaluation Results

Given the number of evaluation metrics (16 in total on the task website), individual analysis of system results per metric is beyond the scope of this paper. However, a ranking of systems may be obtained by taking a summed ranked score; that is, by adding

[9] http://www.cse.unt.edu/~rada/downloads.html#semcor.

[10] For the random senses baselines, induced senses are mapped to WordNet 3.1 senses using the mapping procedure described in Agirre and Soroa (2007). The mapping is provided by the task organisers.

[11] http://www.cs.york.ac.uk/semeval-2013/task13/index.php?id=results

System/*Baseline*	**Fuzzy NMI**	**Fuzzy B-Cubed Precision**	**Fuzzy B-Cubed Recall**	**Fuzzy B-Cubed F-Score**
Unimelb (50k)	0.060	0.524	0.447	0.483
Unimelb (5p)	0.056	0.470	0.449	0.459
AI-KU	0.065	0.838	0.254	0.390
AI-KU (r5-a1000)	0.039	0.502	0.409	0.451
UoS (top 3)	0.045	0.479	0.420	0.448
UoS (# WN senses)	0.047	0.988	0.112	0.201
AI-KU (a1000)	0.035	0.905	0.194	0.320
One sense	0.000	**0.989**	0.455	**0.623**
1 of 2 random senses	0.028	0.495	**0.456**	0.474
1 of 3 random senses	0.018	0.329	0.455	0.382
1 of n random senses	0.016	0.168	0.451	0.245
1 sense per instance	**0.071**	0.000	0.000	0.000

Table 2: Results for the cluster-based evaluation: all instances.

up each system's rankings over all evaluation metrics. The summed ranking finds that UoS (top 3) is placed first. If the WSD and cluster-based evaluations are considered separately, then UoS (top 3) is ranked, respectively, first and fourth. However, this result is countered by the relatively poor performance of UoS (# WN senses), being ranked fifth overall. Considering baselines, UoS (top 3) equals or surpasses the SemCor baseline scores 67% of the time, and 54% for the more challenging baselines; UoS (# WN senses) scores, respectively, 50% and 44%.

All instances results were supplemented with single-sense (non-graded) and multi-sense (graded) splits at a later date.[12] These results show (again, using a ranked score) that for single-sense instances, AI-KU is the best performing system, with UoS (top 3) placed fifth, and UoS (# WN senses) last. Both UoS (top 3) and UoS (# WN senses) surpass the *SemCor MFS* baseline, with UoS (top 3) surpassing or equalling the harder baselines 79% of the time, and UoS (# WN senses) 68% of the time. For multi-sense instances, AI-KU is, again, the best performing system, with UoS (# WN senses) placed second and UoS (top 3) sixth. UoS (top 3) surpasses or equals the *SemCor* baseline scores 67% of the time; UoS (# WN senses) 83% of the time. UoS (top3) passes/equals, the harder baselines 63% of the time, with UoS (# WN senses) doing so 67% of the time. These results are somewhat confounding as one would expect a system that performs well in the main set of results (all instances), as UoS (top 3) does, to do so in at least one of the single-sense / multi-sense splits: this is clearly not the case. Indeed, the results suggest that UoS (# WN senses), found to perform poorly over all instances, is better suited to the task's aim of finding graded senses.

5 Conclusion

This paper presented UoS, a graph-based WSI system that participated in SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses. UoS applied the MaxMax clustering algorithm to find a set of sense clusters in a target word graph. The number of clusters was found automatically through identification of root vertices of maximal quasi-strongly connected subgraphs. Evaluation results showed the UoS (top 3) system to be the best performing system (all instances), if a simple ranking over all evaluation measures is applied. The second system, UoS (# WN senses), performed poorly, being ranked fifth out of the seven participating WSI systems. Note, however, that the number of evaluation metrics applied, and the wide variability in each system's performances over different metrics and different splits of instance types, make it difficult to judge exactly which system is the best performing. Future research therefore aims to carry out a detailed analysis of the results and to assess whether the measures applied in the evaluation adequately reflect the performance of WSI systems.

[12] `http://www.cs.york.ac.uk/semeval-2013/task13/index.php?id=results` (4/4/2013)

References

Eneko Agirre and Aitor Soroa. 2007. SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 7–12. Association for Computational Linguistics. Prague, Czech Republic.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints. *Information Retrieval*, 12(4):461–486.

Gerlof Bouma. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proceedings of GSCL*, pages 31–40.

T. Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Katrin Erk and Diana McCarthy. 2009. Graded Word Sense Assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 440–449, Singapore. Association for Computational Linguistics.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54. Marrakech, Morocco.

David Hope and Bill Keller. 2013. MaxMax: A Graph-Based Soft Clustering Algorithm Applied to Word Sense Induction. In A. Gelbukh, editor, *CICLing 2013, Part I, LNCS 7816*, pages 368–381. Springer-Verlag Berlin Heidelberg. to appear.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 57–60. Association for Computational Linguistics.

David Jurgens. 2012. An Evaluation of Graded Sense Disambiguation Using Word Sense Induction. *Proceedings of *SEM First Joint Conference on Lexical and Computational Semantics, 2012. Association for Computational Linguistics*, pages 189–198. Montreal,Canada.

Andrea Lancichinetti, Santo Fortunato, and János Kertész. 2009. Detecting the Overlapping and Hierarchical Community Structure in Complex Networks. *New Journal of Physics*, 11(3):033015.

Suresh Manandhar, Ioannis P. Klapaftis, Dmitriy Dligach, and Sameer S. Pradhan. 2010. SemEval-2010 Task 14: Word Sense Induction and Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68. Association for Computational Linguistics. Uppsala, Sweden.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235.

Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Roberto Navigli. 2012. A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. In *SOFSEM 2012: Theory and Practice of Computer Science*, volume 7147 of *Lecture Notes in Computer Science*, pages 115–129. Springer Berlin / Heidelberg.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. Pearson Addison Wesley.

K. Thulasiraman and N.S. Swamy. 1992. *Graphs: Theory and Algorithms*. Wiley.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Blackwell.