Errata for SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses

February 27, 2015

David Jurgens
School of Computer Science
McGill University
jurgens@cs.mcgill.ca

Ioannis Klapaftis
Search Technology Center Europe
Microsoft
ioannisk@microsoft.com

1 Summary

SemEval-2013 Task 13 measures the performance of Word Sense Induction (WSI) and unsupervised Word Sense Disambiguation (WSD) systems. Given a word in context, the systems must label the word with one or more senses, indicating valid interpretations of the word, and where each sense annotation may be accompanied by a weight indicating how likely that interpretation is. Before performing the annotation task, WSI systems first induce the different meanings of a word by examining usages of the word in text; in contrast, the WSD systems were asked to use the WordNet 3.1 sense inventory.

After the completion of Task 13 and the publication of the task description paper (Jurgens and Klapaftis, 2013), a software bug was discovered in the evaluation program that affected the scores in a limited set of circumstances.[1] Specifically, the bug resulted in an incorrect calculation of Recall for a WSI or WSD system when not all instances were labeled with senses. In most cases, a system does label all instances with senses and thus, the bug does not occur. However, the task report includes a follow-up experiment that tested systems using only those instances that were labeled with multiple senses; in this setting, many WSI systems ultimately reported fewer instances and, due to the bug, had incorrect scores in the task report.[2]

Jaccard Index		K_{δ}^{sim}	
Old	New	Old	New
0.244	0.245	0.642	0.641

Table 1: Scoring changes for AI-KU (remove5-add1000) in the all-instances setting (cf. Table 3 in the task paper)

Precision	
Old	New
0.628	0.630

Table 2: Scoring changes for AI-KU (remove5-add1000) in the single-sense instance setting (cf. Table 4 in the task paper)

Importantly, this bug did not affect any calculation of B-Cubed or Normalized Mutual Information (NMI) scores, nor did the changes in magnitude affect general findings of the task.

2 Corrigendum

Apart from the multi-sense instance setting, only the scores for AI-KU (remove5-add1000) were affected by the bug and changed slightly from those reported in the paper. Tables 1 and 2 show the score corrections for this system in the all-instances and single-sense instances settings.

In the multi-sense setting, multiple systems had

[1] Evaluation code is available at https://code.google.com/p/cluster-comparison-tools/.

[2] The reason for WSI systems not labeling all instances was likely due to the small size of the multi-sense data. For a WSI system to label an instance with WordNet senses, a mapping function is trained that transforms an annotation with induced senses to one with WordNet senses. When little training data is available, the mapping function may not observe all induced senses during training, and thus during testing, when presented with an instance with a novel induced sense, the annotation cannot be transformed and the instance is never labeled.

		WSD F1		
Team	System	Jac. Ind.	K_{δ}^{sim}	WNDCG
AI-KU	remove5-add1000	**0.444**	0.573	0.297
Unimelb	5p	0.430	0.586	0.289
Unimelb	50k	0.417†	0.598	0.301
UoS	#WN Senses	0.387	**0.627**	0.313
UoS	top-3	0.431	0.565	0.309
La Sapienza	system-1	0.263	0.492	0.288
La Sapienza	system-2	0.263	0.531†	**0.365**†

Table 3: Corrected system performances on all instances labeled with multiple senses. Top system performances are marked in bold, with † indicating the system that was formerly marked as top-performing for each metric (cf. Table 5 in the task paper)

their scores impacted by the bug. Table 3 shows the corrected performance numbers for affected systems. Even with the corrections, the general trend remains that all systems outperform the baselines. While, the correct scores do re-order the top performing systems for the Jaccard Index or K_{δ}^{sim}, the relative magnitude of score differences between systems are largely consistent with those in the original task description paper.

Acknowledgments

We thank Kavek for first identifying the scoring discrepancy and help in verifying the result.

References

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 task 10: Word Sense Induction for Graded and Non-Graded Senses. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics.