SUDOKU: Treating Word Sense Disambiguation & Entity Linking as a Deterministic Problem – via an Unsupervised & Iterative Approach

Steve L. Manion
University of Canterbury, Christchurch, New Zealand
steve.manion @pg.canterbury.ac.nz

Abstract

SUDOKU's submissions to SemEval Task 13 treats Word Sense Disambiguation and Entity Linking as a deterministic problem that exploits two key attributes of open-class words as constraints – their *degree of polysemy* and their *part of speech*. This is an extension and further validation of the results achieved by Manion and Sainudiin (2014). SUDOKU's three submissions are incremental in the use of the two aforementioned constraints. Run1 has no constraints and disambiguates all lemmas in one pass. Run2 disambiguates lemmas at increasing degrees of polysemy, leaving the most polysemous until last. Run3 is identical to Run2, with the additional constraint of disambiguating all named entities and nouns first before other types of open-class words (verbs, adjectives, and adverbs). Over all-domains, for English Run2 and Run3 were placed second and third. For Spanish Run2, Run3, and Run1 were placed first, second, and third respectively. For Italian Run1 was placed first with Run2 and Run3 placed second equal.

1 Introduction & Related Work

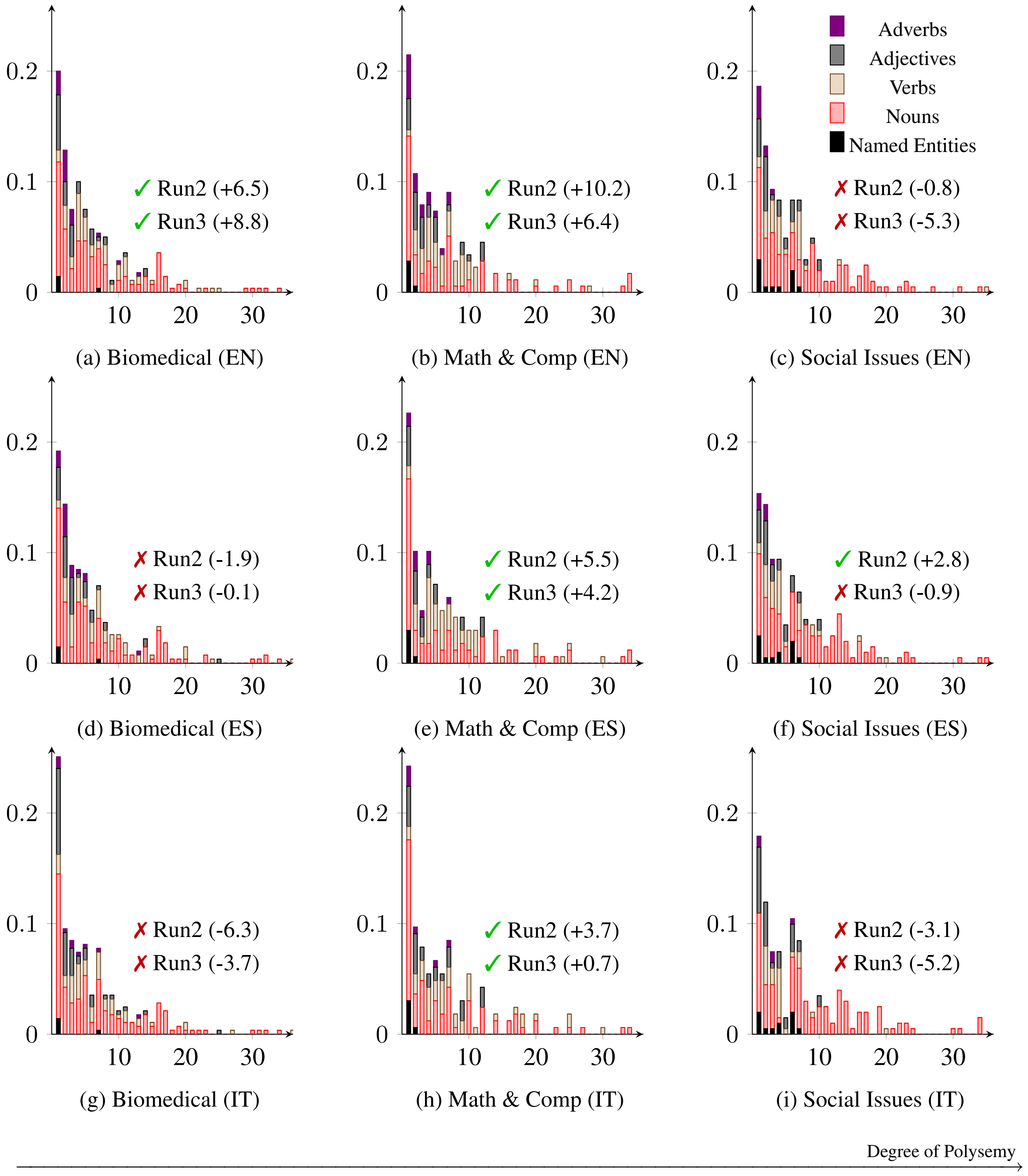
Almost a decade ago, Agirre and Edmonds (2007) suggested the promising potential for WSD that could exploit the interdependencies between senses in an interactive manner. In other words, this would be a WSD system which allows the disambiguation of word a to directly influence the *consecutive* disambiguation of word b. This is analogous to treating WSD as a deterministic problem, much like the Sudoku puzzle in which the final solution is reached by adhering to a set of pre-determined constraints. *Conventional* approaches to WSD often overlook the potential to exploit sense interdependencies, and simply disambiguate all senses in one pass based on a context window (e.g. a sentence or document). For this task the author proposes an *iterative* approach which makes several passes based on a set of constraints. For a more formal distinction between the *conventional* and *iterative* approach to WSD, please refer to this paper (Manion and Sainudiin, 2014).

Yr	%NE	%N	%V	%R	%A	F	ΔF
'04	-	37.7	34.0	12.6	15.6	27.1	+16.8
'10	-	73.8	26.2	-	-	26.8	+11.1
'13	17.1	82.9	-	-	-	58.3	+6.1
'15	*6.0*	*44.9*	*28.9*	*6.5*	*13.7*	*55.8*	*+5.8*

Table 1: Parts of Speech disambiguated (as percentages) for each SemEval Task (denoted by its year). In-Degree Centrality as implemented in (Manion and Sainudiin, 2014) observes F-Score improvement (F + ΔF) by applying the *iterative* approach.

The author found in the investigations of his thesis (Manion, 2014) that the iterative approach performed best on the SemEval 2013 Multilingual WSD Task (Navigli et al., 2013), as opposed to earlier tasks such as SensEval 2004 English All Words WSD Task (Snyder and Palmer, 2004) and the SemEval 2010 All Words WSD task on a Specific Domain (Agirre et al., 2010). While these earlier tasks also experienced improvement, F-Scores remained lower overall. Table 1 above and Figures 1(a) to (i) help highlight what changed between these tasks.

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 365–369,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics



(a) Biomedical (EN)

(b) Math & Comp (EN)

(c) Social Issues (EN)

(d) Biomedical (ES)

(e) Math & Comp (ES)

(f) Social Issues (ES)

(g) Biomedical (IT)

(h) Math & Comp (IT)

(i) Social Issues (IT)

Figure 1: Depicted above are distributions for each domain and language, detailing the probability (y-axis) of specific parts of speech at increasing degrees of polysemy (x-axis). These distributions were produced from the gold keys (or synsets) of the test documents by querying BabelNet for the polysemy of each word. Each distribution was normalised with one sense per discourse assumed, therefore duplicate synsets were ignored. Lastly the difference in F-Score between the *conventional* Run1 and the *iterative* Run2 and Run3 is listed beside each distribution.

Firstly WSD tasks before 2013 generally relied on only a lexicon, such as WordNet (Fellbaum, 1998) or an alternative equivalent, whereas SemEval 2013 Task 12 WSD and this task (Moro and Navigli, 2015) included Entity Linking (EL) using the encyclopaedia Wikipedia via BabelNet (Navigli and Ponzetto, 2012). Secondly, as shown by Manion and Sainudiin (2014) with a simple linear regression, the iterative approach increases WSD performance for documents that have a higher degree of *document monosemy* - the percentage of unique monosemous lemmas in a document. As seen in Figures 1(a) to (i) on the previous page, named entities (or *unique* rather than *common* nouns) are more monosemous compared to other parts of speech, especially for more technical domains. Lastly, the SemEval 2013 WSD task differs in that only nouns and named entities required disambiguation. This simplifies the WSD task, as shown in the experiments on local context by Yarowsky (1993), nouns are best disambiguated by directly adjacent nouns (or modifying adjectives). Based on these observations, the author hypothesized the following implementations of the iterative approach should perform well.

2 System Description & Implementation

Run1 (SUDOKU-1) is the *conventional* approach – *no constraints* are applied. Formalised in (Manion and Sainudiin, 2014), this run can act as a baseline to gauge any improvement for Run2 and Run3 that apply the *iterative* approach. Run2 (SUDOKU-2) has the constraint of words being disambiguated in order of increasing polysemy, leaving the most polysemous to last. Run3 (SUDOKU-3) is an untested and unpublished version of the *iterative* approach. It includes Run2's constraint plus a second constraint – that all nouns and named entities must be disambiguated before other parts of speech.

For each run, a semantic subgraph is constructed from BabelNet (version 2.5.1). Then for disambiguation the graph centrality measure PageRank (Brin and Page, 1998) is used in conjunction with a surfing vector that biases probability mass to certain sense nodes in the semantic subgraph. This idea is taken from Personalised PageRank (PPR) (Agirre and Soroa, 2009), which applies the method put forward by Haveliwala (2003) to the field of WSD. In the previous SemEval WSD task (Navigli et al., 2013) team UMCC_DLSI (Gutierrez et al., 2013) implemented this method and achieved the best performance by biasing probability mass based on SemCor (Miller et al., 1993) sense frequencies. As the winning method for this task, PPR was selected to test the iterative approach on. For SUDOKU's implementation to be *unsupervised*, all runs biased probability mass towards senses from monosemous lemmas. Additionally for Run2 and Run3, once a lemma is disambiguated it is considered to be monosemous. Therefore with each iteration of Run2 and Run3, probability mass is redistributed across the surfing vector to acknowledge these *newly appointed* monosemous lemmas.

All system runs are applied at the document level, across all languages and domains, for all named entities, nouns, verbs, adverbs, and adjectives. Semantic subgraphs are constructed from BabelNet via a Depth First Search (DFS) up to 2 hops in path length. PageRank's damping factor is set to 0.85, with a maximum of 30 iterations[1]. In order to avoid masking the effect of using the iterative approach, a *back-off* strategy (see (McCarthy et al., 2004)) was *not* used. Multiword units were found by finding lemma sequences that contained at least one noun and at the same time could return a result from BabelNet. Lemma sequences beginning with definite/indefinite articles (e.g. *the*, *a*, *il*, *la*, and *el*) were removed as they induced too much noise, given they almost always returned a result from BabelNet (such as a book or movie title).

3 Results, Discussions, & Conclusions

As seen in Figures 1(a) to (i) on the previous page, the Biomedical and Math & Computers domains include a substantial degree of monosemy, no doubt increased by the monosemous technical terms and named entities present. Given the importance of document monosemy for the iterative approach, it is of no surprise that Run2 and Run3 in most cases performed much better than Run1 for these technical domains. Equally so, Run2 and Run3 were outperformed by Run1 for the less technical Social Issues

[1] PageRank iterations remain at the atomic level, i.e. they do not influence the construction of the semantic subgraph, see (Manion and Sainudiin, 2014) Section 3.1 for more details.

	All Domains			Biology			Math & Comp			Social Issues		
Part of Speech	(1)	Δ(2-1)	Δ(3-1)	(1)	Δ(2-1)	Δ(3-1)	(1)	Δ(2-1)	Δ(3-1)	(1)	Δ(2-1)	Δ(3-1)
Named Ents	16.8	+70.2	+70.2	4.1	+94.8	+94.8	0.0	+56.3	+56.3	60.9	+20.6	+20.6
Nouns	53.4	+9.1	+9.3	62.8	+9.1	+13.0	28.5	+22.9	+20.4	56.4	-3.6	-8.2
Verbs	52.2	-2.6	-6.2	52.5	-5.2	-1.9	51.4	-2.3	-9.1	52.9	+3.9	-12.0
Adverbs	48.9	+21.5	+22.8	50.7	+27.2	+24.6	52.0	+4.6	+12.2	36.4	+39.5	+39.5
Adjectives	74.4	-2.7	-6.3	82.3	+1.0	-4.5	75.0	-7.5	-17.5	63.6	-4.3	-0.6

Table 2: The difference in F-Scores over each Domain and Part of Speech for English SUDOKU Runs.

domain in which many of the named entities are polysemous rather than monosemous.

While the iterative approach achieved reasonably competitive results in English, this success did not translate as well to Spanish and Italian. The Italian Biomedical domain had the highest document monosemy, observable in Figure 1 (g), yet this did not help the *iterative* Run2 and Run3. Yet it is worth noting the results of the task paper (Moro and Navigli, 2015) report that SUDOKU Run2 and Run3 achieved very low F-Scores for named entity disambiguation (<28.6) in Spanish and Italian. Given that more than half of the named entities were monosemous in Figure 1(d) and (g), the WSD system either did not capture them in text or filtered them out during subgraph construction (see BabelNet API). This underscores the importance of named entities being included in disambiguation tasks. To further support this evidence, while the iterative approach is suited to domain based WSD, recall that the 2010 domain based WSD task in Table 1 also had no tagged named entities (and thus scores were lower than for successive named entity *inclusive* WSD tasks).

As seen in Table 2, the iterative approach has a varied effect on different parts of speech. Always improved is the disambiguation of named entities and adverbs. This is also the case for nouns in technical domains (e.g. Biomedical as opposed to Social Issues). On the other hand the disambiguation of verbs and adjectives suffers under the iterative approach. In hindsight, the iterative approach could be restricted to the parts of speech it is known to improve, while remaining with the conventional approach on others. To the right in Table 3 the author's SUDOKU runs are compared against the team with the most competitive results – LIMSI. The author could not improve on their superior results achieved in English, however for Spanish and Italian the BabelNet First Sense (BFS) baseline was much lower since it often resorted to lexicographic sorting in the absence of WordNet synsets – see (Navigli et al., 2013). The author's *baseline-independent* submissions were unaffected by this, which on reviewing results in (Moro and Navigli, 2015) appears to have helped SUDOKU do best for these languages.

	Team Run	All	Bio	Mat	Soc
(EN)	LIMSI	**65.8**	**71.3**	**54.1**	**67.2**
	SUDOKU-2	61.6	68.9	53.2	55.6
	SUDOKU-3	60.7	71.2	49.4	51.1
	SUDOKU-1	55.8	62.4	43.0	56.4
	BFS	67.5	72.2	55.3	70.8
(ES)	SUDOKU-2	**57.1**	60.8	**49.7**	**57.0**
	SUDOKU-3	56.8	62.6	48.4	53.3
	SUDOKU-1	56.0	**62.7**	44.2	54.2
	LIMSI	45.0	51.0	34.8	43.1
	BFS	37.5	43.7	28.7	34.0
(IT)	SUDOKU-1	**59.9**	**65.1**	48.4	**61.0**
	SUDOKU-3	56.9	64.1	49.1	55.8
	SUDOKU-2	56.9	58.8	**52.1**	57.9
	LIMSI	48.4	53.1	44.6	42.9
	BFS	40.2	44.3	36.7	35.7

Table 3: F1 scores for each domain/language for SUDOKU and LIMSI.

In summary, the inclusion of named entities in disambiguation tasks certainly improves results, as well as the effectiveness of the iterative approach. Furthermore in Table 3 above, the *iterative* Run3 for the English Biomedical domain is 0.1 short of achieving the best result of 71.3. Investigating exactly which factors contributed to the success of this *unsupervised* result is a top priority for future work.

Resources

Codebase and resources are at the author's homepage: http://www.stevemanion.com.

Acknowledgments

This submission is an extension of the author's PhD thesis completed under the supervision of Dr Raazesh Sainudiin and with the help of the Korean Foundation Graduate Studies Fellowship.

References

Eneko Agirre and Philip Edmonds 2007. Chapter 1: Introduction. *Word Sense Disambiguation Algorithms and Applications*, pages 1-28. Springer, New York.

Eneko Agirre, Oier Lopez De Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. SemEval-2010 Task 17: All-words Word Sense Disambiguation on a Specific Domain. *In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, pages 75-80. Uppsala, Sweden.

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, pages 33-41. Athens, Greece.

Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30:107-117.

Christiane Fellbaum. 1998, ed. *WordNet: An Electronic Lexical Database.*, Cambridge, MA: MIT Press.

Yoan Gutirrez, Antonio Fernndez Orqun, Franc Camara, Yenier Castaeda, Andy Gonzlez, Andrs Montoyo, Rafael Muoz, Rainel Estrada, Dennys D. Piug, Jose I. Abreu, and Roger Prez. 2013. UMCC_DLSI: Reinforcing a Ranking Algorithm with Sense Frequencies and Multidimensional Semantic Resources to solve Multilingual Word Sense Disambiguation. *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013)*, pages 241-249. Atlanta, Georgia.

Taher H. Haveliwala. 2003. A Context-Sensetive Ranking Algorithm for Web Search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796.

Steve L. Manion and Raazesh Sainudiin. 2014. An Iterative Sudoku Style Approach to Subgraph-based Word Sense Disambiguation. *In Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics (*SEM'14)*, pages 40-50. Dublin, Ireland.

Steve L. Manion. 2014. Unsupervised Knowledge-based Word Sense Disambiguation: Exploration & Evaluation of Semantic Subgraphs. *Doctoral Thesis*. University of Canterbury.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding Predominant Word Senses in Untagged Text. *In Proceedings of the 42nd Annual Meeting for the Association for Computational Linguistics (ACL'04)*, pages 280-287. Barcelona, Spain.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A Semantic Concordance. *In Proceedings of the Workshop on Human Language Technology (HLT93)*, pages 303-308. Princeton, New Jersey.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*. Denver, Colorado.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual-Word Sense Disambiguation. *In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013)*, pages 222-231. Atlanta, Georgia.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Benjamin Snyder and Martha Palmer. 2004. The English All-Words Task. *In Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 41-43. Barcelona, Spain.

David Yarowsky. 1993. One Sense Per Collocation. *In Proceedings of the ARPA Workshop on Human Language Technology (HLT'93)*, pages 266-271. Morristown, New Jersey.