The SEMEVAL English Lexical Substitution Task: Results

Diana McCarthy and Roberto Navigli

1 Results and Baselines

In this document we show precision (P) and recall (R) and mode precision (mode P) and mode recall (mode R) as described in our scoring documentation[1]. In tables 1 to 4 we have ordered systems according to recall on the **best** task, and in tables 5 to 8 according to recall on **oot**. In tables 3, 4, 7 and 8 we show further analysis of results using the subset of items which were i) NOT identified as multiwords (NMWT) ii) scored only on non multiword substitutes from both annotators and systems (i.e. no spaces) (NMWS) iii) items where the sentences were selected randomly (RAND) and iv) items where the sentences were selected manually (MAN). We retain the same ordering of systems for this further analysis. Although there are further differences in the systems which would warrant reranking on an individual analysis, since we combined the subanalyses in one table we keep the order as for 1 and 5 respectively for ease of comparison.

We produced baselines using WordNet 2.1 (Miller et al., 1993) and a number of distributional similarity measures. For the WordNet **best** baseline we found the best ranked synonym using the criteria 1 to 4 below in order. For WordNet **oot** we found up to 10 synonyms using criteria 1 to 4 in order until 10 were found:

1. Synonyms from the first synset of the target word, and ranked with frequency data obtained from the BNC (Leech, 1992).
2. synonyms from the hypernyms (verbs and nouns) or closely related classes (adjectives) of that first synset, ranked with the frequency data.
3. Synonyms from all synsets of the target word, and ranked using the BNC frequency data.
4. synonyms from the hypernyms (verbs and nouns) or closely related classes (adjectives) of all synsets of the target, ranked with the BNC frequency data.

Systems	P	R	Mode P	Mode R
KU	12.90	12.90	20.65	20.65
UNT	12.77	12.77	20.73	20.73
MELB	12.68	12.68	20.41	20.41
HIT	11.35	11.35	18.86	18.86
USYD	11.23	10.88	18.22	17.64
IRST1	8.06	8.06	13.09	13.09
IRST2	6.95	6.94	20.33	20.33
TOR	2.98	2.98	4.72	4.72

Table 1: **best** results

We also produced **best** and **oot** baselines using the distributional similarity measures l1, jaccard, cosine, lin (Lin, 1998) and αSD (Lee, 1999) [2]. We took the word with the largest similarity (or smallest distance for αSD and l1) for **best** and the top 10 for **oot**.

For **mw** detection and identification we used WordNet to detect if a multiword in WordNet which includes the target word occurs within a window of 2 words before and 2 words after the target word.

[1] Available at http://nlp.cs.swarthmore.edu/semeval/tasks/task10/task10documentation.pdf

[2] We used 0.99 as the parameter for α for this measure.

	NMWT		NMWS		RAND		MAN	
Systems	P	R	P	R	P	R	P	R
KU	13.39	13.39	14.33	13.98	12.67	12.67	13.16	13.16
UNT	13.46	13.46	13.79	13.79	12.85	12.85	12.69	12.69
MELB	13.35	13.35	14.19	13.82	12.50	12.50	12.89	12.89
HIT	11.97	11.97	12.55	12.38	11.81	11.81	10.81	10.81
USYD	11.68	11.34	12.48	12.10	11.47	11.01	10.95	10.73
IRST1	8.44	8.44	8.98	8.92	8.65	8.64	7.38	7.38
IRST2	7.25	7.24	7.67	7.66	6.71	6.68	7.23	7.23
TOR	3.22	3.22	3.32	3.32	3.10	3.10	2.84	2.84

Table 3: Further analysis for **best**

	NMWT		NMWS		RAND		MAN	
Systems	Mode P	Mode R	Mode P	Mode R	Mode P	Mode R	Mode P	Mode R
KU	21.20	21.20	21.88	21.42	20.34	20.34	21.01	21.01
UNT	21.63	21.63	21.59	21.59	20.18	20.18	21.35	21.35
MELB	21.29	21.29	21.74	21.33	19.72	19.72	21.18	21.18
HIT	19.81	19.81	19.93	19.65	20.03	20.03	17.53	17.53
USYD	18.46	17.90	19.25	18.63	19.14	18.35	17.20	16.84
IRST1	13.38	13.38	13.85	13.74	13.76	13.76	12.33	12.33
IRST2	20.76	20.76	21.50	21.50	22.17	22.17	18.23	18.23
TOR	5.04	5.04	4.90	4.89	5.20	5.20	4.17	4.17

Table 4: Further analysis for **best**: finding the mode

	NMWT		NMWS		RAND		MAN	
Systems	P	R	P	R	P	R	P	R
IRST2	72.04	71.90	76.19	76.06	66.94	66.72	71.46	71.46
UNT	51.13	51.13	54.01	54.01	51.71	51.71	46.26	46.26
KU	48.43	48.43	49.72	49.72	47.80	47.80	44.23	44.23
IRST1	43.11	43.08	45.13	45.11	42.14	42.09	40.17	40.17
USYD	37.26	36.17	40.13	38.89	35.67	34.26	36.52	35.78
SWAG2	39.95	36.51	40.97	37.75	39.74	36.26	35.56	32.79
HIT	35.60	35.60	36.63	36.63	33.95	33.95	33.81	33.81
SWAG1	37.49	34.64	38.36	35.67	36.94	34.52	33.83	30.85
TOR	11.77	11.77	12.22	12.22	9.98	9.98	12.61	12.61

Table 7: Further analysis for **oot**

	NMWT		NMWS		RAND		MAN	
Systems	Mode P	Mode R	Mode P	Mode R	Mode P	Mode R	Mode P	Mode R
IRST2	60.38	60.38	61.97	61.97	58.26	58.26	58.85	58.85
UNT	68.03	68.03	70.15	70.15	68.04	68.04	64.24	64.24
KU	63.42	63.42	63.74	63.74	62.84	62.84	59.55	59.55
IRST1	56.82	56.82	58.26	58.26	55.50	55.50	55.03	55.03
USYD	44.71	43.35	46.25	44.77	42.90	41.13	44.50	43.58
SWAG2	52.28	47.78	52.25	47.98	53.61	48.78	46.34	42.88
HIT	48.48	48.48	49.33	49.33	47.25	47.25	46.53	46.53
SWAG1	49.11	45.35	49.41	45.70	48.94	45.72	45.63	41.67
TOR	15.03	15.03	15.26	15.26	13.00	13.00	16.49	16.49

Table 8: Further analysis for **oot**: finding the mode

Systems	P	R	Mode P	Mode R
WordNet	9.95	9.95	15.28	15.28
lin	8.84	8.53	14.69	14.23
l1	8.11	7.82	13.35	12.93
lee	6.99	6.74	11.34	10.98
jaccard	6.84	6.60	11.17	10.81
cos	5.07	4.89	7.64	7.40

Table 2: **best** baseline results

Systems	P	R	Mode P	Mode R
IRST2	69.03	68.90	58.54	58.54
UNT	49.19	49.19	66.26	66.26
KU	46.15	46.15	61.30	61.30
IRST1	41.23	41.20	55.28	55.28
USYD	36.07	34.96	43.66	42.28
SWAG2	37.80	34.66	50.18	46.02
HIT	33.88	33.88	46.91	46.91
SWAG1	35.53	32.83	47.41	43.82
TOR	11.19	11.19	14.63	14.63

Table 5: **oot** results

Systems	P	R	Mode P	Mode R
WordNet	29.70	29.35	40.57	40.57
lin	27.70	26.72	40.47	39.19
l1	24.09	23.23	36.10	34.96
lee	20.09	19.38	29.81	28.86
jaccard	18.23	17.58	26.87	26.02
cos	14.07	13.58	20.82	20.16

Table 6: **oot** baseline

	system HIT		WordNet BL	
	P	R	P	R
detection	45.34	56.15	43.64	36.92
identification	41.61	51.54	40.00	33.85

Table 9: MW results

References

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Geoffrey Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.

George Miller, Richard Beckwith, Christine Fellbaum, David Gross, and Katherine Miller, 1993. *Introduction to WordNet: an On-Line Lexical Database*. ftp://clarity.princeton.edu/pub/WordNet/5papers.ps.