

System	English			Farsi			German			Italian			Spanish		
	r	$\rho$	Final												
Luminoso_run2	<b>0.78</b>	<b>0.80</b>	<b>0.79</b>	0.51	0.50	0.50	<b>0.70</b>	<b>0.70</b>	<b>0.70</b>	<b>0.73</b>	<b>0.75</b>	<b>0.74</b>	<b>0.73</b>	<b>0.75</b>	<b>0.74</b>
Luminoso_run1	<b>0.78</b>	0.79	<b>0.79</b>	0.51	0.50	0.50	0.69	0.69	0.69	<b>0.73</b>	<b>0.75</b>	<b>0.74</b>	<b>0.73</b>	<b>0.75</b>	<b>0.74</b>
QLUT_run1*	<b>0.78</b>	0.78	0.78	-	-	-	-	-	-	-	-	-	-	-	-
hhu_run1*	0.71	0.70	0.70	0.54	0.59	0.56	-	-	-	-	-	-	-	-	-
HCCL_run1*	0.68	0.70	0.69	0.42	0.45	0.44	0.58	0.61	0.59	0.63	0.67	0.65	0.69	0.72	0.70
NASARI (baseline)	0.68	0.68	0.68	0.41	0.40	0.41	0.51	0.51	0.51	0.60	0.59	0.60	0.60	0.60	0.60
hhu_run2*	0.66	0.70	0.68	0.61	0.60	0.60	-	-	-	-	-	-	-	-	-
QLUT_run2*	0.67	0.67	0.67	-	-	-	-	-	-	-	-	-	-	-	-
RUFINO_run1*	0.65	0.66	0.66	0.38	0.34	0.36	0.54	0.54	0.54	0.48	0.47	0.48	0.53	0.57	0.55
Citius_run2	0.60	0.71	0.65	-	-	-	-	-	-	-	-	-	0.44	0.64	0.52
I2f_run2 (a.d.)	0.64	0.65	0.65	-	-	-	-	-	-	-	-	-	-	-	-
I2f_run1 (a.d.)	0.64	0.65	0.64	-	-	-	-	-	-	-	-	-	-	-	-
Citius_run1*	0.57	0.65	0.61	-	-	-	-	-	-	-	-	-	0.44	0.63	0.51
MERALI_run1*	0.59	0.60	0.59	-	-	-	-	-	-	-	-	-	-	-	-
Amateur_run1*	0.58	0.59	0.59	-	-	-	-	-	-	-	-	-	-	-	-
Amateur_run2*	0.58	0.59	0.59	-	-	-	-	-	-	-	-	-	-	-	-
MERALI_run2*	0.57	0.58	0.58	-	-	-	-	-	-	-	-	-	-	-	-
SEW_run2 (a.d.)	0.56	0.58	0.57	0.38	0.40	0.39	0.45	0.45	0.45	0.57	0.57	0.57	0.61	0.62	0.62
jmp8_run1*	0.47	0.69	0.56	-	-	-	0.26	0.51	0.35	0.41	0.64	0.50	-	-	-
Wild_Devs_run1	0.46	0.48	0.47	-	-	-	-	-	-	-	-	-	-	-	-
RUFINO_run2*	0.39	0.40	0.39	0.25	0.26	0.26	0.38	0.36	0.37	0.30	0.31	0.31	0.40	0.41	0.41
SEW_run1	0.37	0.41	0.39	0.38	0.40	0.39	0.45	0.45	0.45	0.57	0.57	0.57	0.61	0.62	0.62
hjpwhuer_run1	-0.04	-0.03	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.05	0.05	0.05	-0.06	-0.06	0.00
Mahtab_run2*	-	-	-	<b>0.72</b>	<b>0.71</b>	<b>0.71</b>	-	-	-	-	-	-	-	-	-
Mahtab_run1*	-	-	-	<b>0.72</b>	<b>0.71</b>	<b>0.71</b>	-	-	-	-	-	-	-	-	-

Table 6: Pearson ( $r$ ), Spearman ( $\rho$ ) and official (Final) results of participating systems on the five monolingual word similarity datasets (subtask 1).

across languages NASARI relies on the lexicalizations provided by BabelNet (Navigli and Ponzetto, 2012) for the concepts and entities in each language. Then, the final score was computed through the conventional closest senses strategy (Resnik, 1995; Budanitsky and Hirst, 2006), using cosine similarity as the comparison measure.

### 3.2 Results

We present the results of subtask 1 in Section 3.2.1 and subtask 2 in Section 3.2.2.

#### 3.2.1 Subtask 1

Table 6 lists the results on all monolingual datasets.<sup>10</sup> The systems which made use of the shared Wikipedia corpus are marked with \* in Table 6. Luminoso achieved the best results in all languages except Farsi. Luminoso couples word embeddings with knowledge from ConceptNet (Speer et al., 2017) using an extension of Retrofitting (Faruqui et al., 2015), which proved highly effective. This system additionally proposed two fallback strategies to handle

<sup>10</sup>Systems followed by (a.d.) submitted their results after the official deadline.

System	Score	Official Rank
Luminoso_run2	0.743	1
Luminoso_run1	0.740	2
HCCL_run1*	0.658	3
NASARI (baseline)	0.598	-
RUFINO_run1*	0.555	4
SEW_run2 (a.d.)	0.552	-
SEW_run1	0.506	5
RUFINO_run2*	0.369	6
hjpwhuer_run1	0.018	7

Table 7: Global results of participating systems on subtask 1 (multilingual word similarity).

out-of-vocabulary (OOV) instances based on loanwords and cognates. These two fallback strategies proved essential given the amount of rare words or domain-specific words which were present in the datasets. In fact, most systems fail to provide scores for all pairs in the datasets, with OOV rates close to 10% in some cases.

The combination of corpus-based and knowledge-based features was not unique to

System	German-Spanish			German-Farsi			German-Italian			English-German			English-Spanish		
	r	$\rho$	Final	r	$\rho$	Final	r	$\rho$	Final	r	$\rho$	Final	r	$\rho$	Final
Luminoso_run2	<b>0.72</b>	<b>0.74</b>	<b>0.73</b>	<b>0.59</b>	<b>0.59</b>	<b>0.59</b>	<b>0.74</b>	<b>0.75</b>	<b>0.74</b>	<b>0.76</b>	<b>0.77</b>	<b>0.76</b>	<b>0.75</b>	<b>0.77</b>	<b>0.76</b>
Luminoso_run1	<b>0.72</b>	0.73	0.72	<b>0.59</b>	<b>0.59</b>	<b>0.59</b>	0.73	0.74	0.73	0.75	<b>0.77</b>	<b>0.76</b>	<b>0.75</b>	<b>0.77</b>	<b>0.76</b>
NASARI (baseline)	0.55	0.55	0.55	0.46	0.45	0.46	0.56	0.56	0.56	0.60	0.59	0.60	0.64	0.63	0.63
OoO_run1	0.54	0.56	0.55	-	-	-	0.54	0.55	0.55	0.56	0.58	0.57	0.58	0.59	0.58
SEW_run2 (a.d.)	0.52	0.54	0.53	0.42	0.44	0.43	0.52	0.52	0.52	0.50	0.53	0.51	0.59	0.60	0.59
SEW_run1	0.52	0.54	0.53	0.42	0.44	0.43	0.52	0.52	0.52	0.46	0.47	0.46	0.50	0.51	0.50
HCCL_run2* (a.d.)	0.42	0.39	0.41	0.33	0.28	0.30	0.38	0.34	0.36	0.49	0.48	0.48	0.55	0.56	0.55
RUFINO_run1 <sup>†</sup>	0.31	0.32	0.32	0.23	0.25	0.24	0.32	0.33	0.33	0.33	0.34	0.33	0.34	0.34	0.34
RUFINO_run2 <sup>†</sup>	0.30	0.30	0.30	0.26	0.27	0.27	0.22	0.24	0.23	0.30	0.30	0.30	0.34	0.33	0.34
hjpwhu_run2	0.05	0.05	0.05	0.01	0.01	0.01	0.06	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.04
hjpwhu_run1	0.05	0.05	0.05	0.01	0.01	0.01	0.06	0.05	0.05	-0.01	-0.01	0.00	0.04	0.04	0.04
HCCL_run1*	0.03	0.02	0.02	0.03	0.02	0.02	0.03	-0.01	0.00	0.34	0.28	0.31	0.10	0.08	0.09
UniBuc-Sem_run1*	-	-	-	-	-	-	-	-	-	0.05	0.06	0.06	0.08	0.10	0.09
Citius_run1 <sup>†</sup>	-	-	-	-	-	-	-	-	-	-	-	-	0.57	0.59	0.58
Citius_run2 <sup>†</sup>	-	-	-	-	-	-	-	-	-	-	-	-	0.56	0.58	0.57

  

System	English-Farsi			English-Italian			Spanish-Farsi			Spanish-Italian			Italian-Farsi		
	r	$\rho$	Final	r	$\rho$	Final	r	$\rho$	Final	r	$\rho$	Final	r	$\rho$	Final
Luminoso_run2	<b>0.60</b>	<b>0.59</b>	<b>0.60</b>	<b>0.77</b>	<b>0.79</b>	<b>0.78</b>	<b>0.62</b>	<b>0.63</b>	<b>0.63</b>	<b>0.74</b>	<b>0.77</b>	<b>0.75</b>	<b>0.60</b>	<b>0.61</b>	<b>0.60</b>
Luminoso_run1	<b>0.60</b>	<b>0.59</b>	<b>0.60</b>	0.76	0.78	0.77	<b>0.62</b>	<b>0.63</b>	<b>0.63</b>	<b>0.74</b>	0.76	<b>0.75</b>	<b>0.60</b>	0.60	<b>0.60</b>
hhu_run1	0.49	0.54	0.51	-	-	-	-	-	-	-	-	-	-	-	-
NASARI (baseline)	0.52	0.49	0.51	0.65	0.65	0.65	0.49	0.47	0.48	0.60	0.59	0.60	0.50	0.48	0.49
hhu_run2	0.43	0.58	0.49	-	-	-	-	-	-	-	-	-	-	-	-
SEW_run2 (a.d.)	0.46	0.49	0.48	0.58	0.60	0.59	0.50	0.53	0.52	0.59	0.60	0.60	0.48	0.50	0.49
HCCL_run2* (a.d.)	0.44	0.42	0.43	0.50	0.49	0.49	0.37	0.33	0.35	0.43	0.41	0.42	0.33	0.28	0.30
SEW_run1	0.41	0.43	0.42	0.52	0.53	0.53	0.50	0.53	0.52	0.59	0.60	0.60	0.48	0.50	0.49
RUFINO_run2 <sup>†</sup>	0.37	0.37	0.37	0.24	0.23	0.24	0.30	0.30	0.30	0.28	0.29	0.29	0.21	0.21	0.21
RUFINO_run1 <sup>†</sup>	0.26	0.25	0.25	0.34	0.34	0.34	0.25	0.26	0.26	0.35	0.36	0.36	0.25	0.25	0.25
HCCL_run1*	0.02	0.01	0.01	0.12	0.07	0.09	0.05	0.05	0.05	0.08	0.06	0.06	0.02	0.00	0.00
hjpwhu_run1	0.00	-0.01	0.00	-0.05	-0.05	0.00	0.01	0.00	0.01	0.03	0.03	0.03	0.02	0.02	0.02
hjpwhu_run2	0.00	-0.01	0.00	-0.05	-0.05	0.00	0.01	0.00	0.01	0.03	0.03	0.03	0.02	0.02	0.02
OoO_run1	-	-	-	0.58	0.59	0.58	-	-	-	0.57	0.57	0.57	-	-	-
UniBuc-Sem_run1*	-	-	-	0.08	0.10	0.09	-	-	-	-	-	-	-	-	-

Table 8: Pearson ( $r$ ), Spearman ( $\rho$ ) and the official (Final) results of participating systems on the ten cross-lingual word similarity datasets (subtask 2).

Luminoso. In fact, most top performing systems combined these two sources of information. For Farsi, the best performing system was Mahtab, which couples information from Word2Vec word embeddings (Mikolov et al., 2013) and knowledge resources, in this case FarsNet (Shamsfard et al., 2010) and BabelNet. For English, the only system that came close to Luminoso was QLUT, which was the best-performing system that made use of the shared Wikipedia corpus for training. The best configuration of this system exploits the Skip-Gram model of Word2Vec with an additive compositional function for computing the similarity of multiwords. However, Mahtab and QLUT only performed their experiments in a single language (Farsi and English, respectively).

For the systems that performed experiments in at least four of the five languages we computed a global score (see Section 3.1.1). Global rank-

ings and results are displayed in Table 7. Luminoso clearly achieves the best overall results. The second-best performing system was HCCL, which also managed to outperform the baseline. HCCL exploited the Skip-Gram model of Word2Vec and performed hyperparameter tuning on existing word similarity datasets. This system did not make use of external resources apart from the shared Wikipedia corpus for training. RUFINO, which also made use of the Wikipedia corpus only, attained the third overall position. The system exploits PMI and an association measure to capture second-order relations between words based on the Jaccard distance (Jimenez et al., 2016).

### 3.2.2 Subtask 2

The results for all ten cross-lingual datasets are shown in Table 8. Systems that made use of the shared Europarl parallel corpus are marked with \* in the table, while systems making use of

System	Score	Official Rank
Luminoso_run2	0.754	1
Luminoso_run1	0.750	2
NASARI (baseline)	0.598	-
OoO_run1*	0.567	3
SEW_run2 (a.d.)	0.558	-
SEW_run1	0.532	4
HCCL_run2* (a.d.)	0.464	-
RUFINO_run1†	0.336	5
RUFINO_run2†	0.317	6
HCCL_run1*	0.103	7
hjpwhu_run2	0.039	8
hjpwhu_run1	0.034	9

Table 9: Global results of participating systems in subtask 2 (cross-lingual word similarity).

Wikipedia are marked with †. Luminoso, the best-performing system in Subtask 1, also achieved the best overall results on the ten cross-lingual datasets. This shows that the combination of knowledge from word embeddings and the ConceptNet graph is equally effective in the cross-lingual setting.

The global ranking for this subtask was computed by averaging the results of the six datasets on which each system performed best. The global rankings are displayed in Table 9. Luminoso was the only system outperforming the baseline, achieving the best overall results. OoO achieved the second best overall performance using an extension of the Bilingual Bag-of-Words without Alignments (BilBOWA) approach of Gouws et al. (2015) on the shared Europarl corpus. The third overall system was SEW, which leveraged Wikipedia-based concept vectors (Raganato et al., 2016) and pre-trained word embeddings for learning language-independent concept embeddings.

## 4 Conclusion

In this paper we have presented the SemEval 2017 task on *Multilingual and Cross-lingual Semantic Word Similarity*. We provided a reliable framework to measure the similarity between nominal instances within and across five different languages (English, Farsi, German, Italian, and Spanish). We hope this framework will contribute to the development of distributional semantics in general and for languages other than English in particular, with a special emphasis on multiling-

ual and cross-lingual approaches. All evaluation datasets are available for download at <http://alt.qcri.org/semeval2017/task2/>.

The best overall system in both tasks was Luminoso, which is a hybrid system that effectively integrates word embeddings and information from knowledge resources. In general, this combination proved effective in this task, as most other top systems somehow combined knowledge from text corpora and lexical resources.

## Acknowledgments

The authors gratefully acknowledge the support of the MRC grant No. MR/M025160/1 for PheneBank and ERC Starting Grant MultiJEDI No. 259234. Jose Camacho-Collados is supported by a Google Doctoral Fellowship in Natural Language Processing.

We would also like to thank Ángela Collados Ais, Claudio Delli Bovi, Afsaneh Hojjat, Ignacio Iacobacci, Tommaso Pasini, Valentina Pyatkin, Alessandro Raganato, Zahra Pilehvar, Milan Gritta and Sabine Ullrich for their help in the construction of the datasets. Finally, we also thank Jim McManus for his suggestions on the manuscript and the anonymous reviewers for their helpful comments.

## References

- Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşa, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL*, pages 19–27.
- Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Sofia, Bulgaria, pages 183–192.
- Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925* .
- Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of EMNLP*, pages 278–289.
- Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*. Berlin, Germany, pages 7–12.