			Correlation
MSRpar	LFT score	Basic	0.5240
		SemRels	0.4318
		Full	0.5074
	LFT scores + features		0.5522
	WN scores		0.5052
	All		0.5852
	(Bär et al. 2012)		0.6830
MSRvid	LFT score	Basic	0.7295
		SemRels	0.6459
		Full	0.6665
	LFT scores + features		0.7716
	WN scores		0.8504
	All		0.8602
	(Bär et al. 2012)		0.8739
SMTeuroparl	LFT score	Basic	0.4695
		SemRels	0.4728
		Full	0.4978
	LFT scores + features		0.4724
	WN scores		0.5111
	All		0.5180
	(Bär et al. 2012)		0.5280

Table 5: Correlations obtained for the test split by our approach and the top-performer at SemEval-2012 Task 6.

including concepts that are not arguments of relations (Full) is better than disregarding them (SemRels), and that building a system grounded exclusively on semantic relations (SemRels) performs worse than simpler approaches that only account for concepts (Basic).

Combining LFT scores and features derived from the proofs, the main novelty of our approach, brings substantial improvements with MSRpar and MSRvid, but worse performance with SMTeuroparl. The fact that most pairs in SMTeuroparl include one ungrammatical sentence makes our NLP tools perform poorly, greatly affecting overall performance. We note, though, that when sentences are easier to parse (MSRpar, MSRvid), the benefits are clear.

Only using scores from WN-based word similarity measures performs astonishingly well. *WN scores* outperforms *LFT scores + features* except in MSRpar. We believe that this is due to the fact that sentences in MSRvid are very short (13 words on average per pair), and the grammar issue in SMTeuroparl pointed out above.

Finally, best results are obtained when all scores and features are combined. This suggests that while WN-based scores provide a strong baseline, it can be improved by incorporating features capturing the semantic structure of sentences. Also, semantic information brings improvements only when combined with simpler methods, as the results obtained by *All*, *LFT scores + features* and *LFT score* with the three LFT modes show.

Comparison with Previous Work Our best system, *All*, performs worse than the top performer (Table 5): -0.0978 (MSRpar), -0.0137 (MSRvid) and -0.100 (SMTeuroparl). We note, though, that (1) the differences are small for MSRvid and SMTeuroparl, and (2) our approach does not require knowledge from Wikipedia or other large corpora.

References

AbdelRahman, S., and Blake, C. 2012. Sbdlrhmn: A rule-based human interpretation system for semantic textual similarity task. In *Proceedings of SemEval 2012*, 536–542.

Agirre, E.; Cer, D.; Diab, M.; and Gonzalez-Agirre, A. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of SemEval 2012*, 385–393.

Banea, C.; Hassan, S.; Mohler, M.; and Mihalcea, R. 2012. Unt: A supervised synergistic approach to semantic text similarity. In *Proceedings of SemEval 2012*, 635–642.

Bär, D.; Biemann, C.; Gurevych, I.; and Zesch, T. 2012. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of SemEval 2012*, 435–440.

Dolan, W. B., and Brockett, C. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Giampiccolo, D.; Magnini, B.; Dagan, I.; and Dolan, B. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 1–9.

Glinos, D. 2012. Ata-sem: Chunk-based determination of semantic text similarity. In *Proceedings of SemEval 2012*, 547–551.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.* 11(1):10–18.

McCune, W., and Wos, L. 1997. Otter: The cade-13 competition incarnations. *Journal of Automated Reasoning* 18:211–220.

Mihalcea, R.; Corley, C.; and Strapparava, C. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st national conference on Artificial intelligence*, AAAI'06, 775–780.

Moldovan, D., and Blanco, E. 2012. Polaris: Lymba's semantic parser. In *Proceedings of the Eight Int. Conference on Language Resources and Evaluation (LREC'12)*.

Moldovan, D.; Harabagiu, S.; Girju, R.; Morarescu, P.; Lacatusu, F.; Novischi, A.; Badulescu, A.; and Bolohan, O. 2002. Lcc tools for question answering. In Voorhees, and Buckland., eds., *Proceedings of the 11th Text REtrieval Conference (TREC-2002)*.

Quinlan, R. J. 1992. Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, 343–348.

Rios, M.; Aziz, W.; and Specia, L. 2012. Uow: Semantically informed text similarity. In *Proceedings of SemEval 2012*, 673–678.

Šarić, F.; Glavaš, G.; Karan, M.; Šnajder, J.; and Dalbelo Bašić, B. 2012. Takelab: Systems for measuring semantic text similarity. In *Proceedings of SemEval 2012*, 441–448.

Wang, Y., and Witten, I. H. 1997. Induction of model trees for predicting continuous classes. In *Poster papers of the 9th European Conference on Machine Learning*. Springer.