Lexical Features We use two lexical features. The first feature is based on TF*IDF cosine similarity. The IDF's were computed over all sentences for each source paper, thus the IDF values differed across each of the 10 source papers. For any citing sentence, we computed the TF*IDF cosine similarity with all the sentences in the source paper and use them as a feature. The second lexical feature is based on the LCS (Longest Common Subsequence) between the citing sentence (C) and source sentence S and is computed as:

$$\frac{|LCS|}{min(|C|, |S|)}$$

Knowledge Based Features We also compute a set of features based on Wordnet similarity. We use six wordnet based word similarity measures and combine these word similarities to obtain six knowledge based sentence similarity features using the method proposed in [2]. The wordnet based word similarity measures we use are path similarity, WUP similarity [7] , LCH similarity [4], Resnik similarity [6], Jiang-Conrath similarity [3], and Lin similarity [5].

Given each of these similarity measures, the similarities between two sentences is computed by first creating a set of senses for each of the words in each of the sentences. Given these two sets of senses, the similarity score between citing sentence C and source sentence S is calculated as follows:

$$sim_{wn}(C, S) = \frac{(\omega + \sum_{i=1}^{|\phi|} \phi_i) * (2|C||S|)}{|C| + |S|}$$

Here ω is the number of shared senses between C and S. The list ϕ contains the similarities of non-shared words in the shorter text, ϕ_i is the highest similarity score of the ith word among all the words of the lower text [1].

Syntactic Features We compute an additional feature based on similarity of dependency structures using the method described in [1] . We use the Stanford parser to obtain dependency parse all the citing sentences and source sentences. Given a candidate sentence pair, two syntactic dependencies are considered equal if they have the same dependency type, govering lemma, and dependent lemma. If R_c and R_s are the set of all dependency relations in C and S, the dependency overlap score is computed using the formula:

$$sim_{dep}(C, S) = \frac{2 * |R_c \cap R_s| * |R_c||R_s|}{|R_c| + |R_s|}$$

References

[1] *ECNUCS: Measuring Short Text Semantic Equivalence Using Multiple Similarity Measurements*, Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. Association for Computational Linguistics, 2013.

[2] Carmen Banea, Samer Hassan, Michael Mohler, and Rada Mihalcea. Unt: A supervised synergistic approach to semantic text similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 635–642. Association for Computational Linguistics, 2012.

[3] J.J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pages 19–33, 1997.

[4] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. In Christiane Fellfaum, editor, *MIT Press*, pages 265–283, Cambridge, Massachusetts, 1998.

[5] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[6] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[7] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.