CSE 841: Project Proposal

Grant King
kinggra1@msu.edu

1. PROBLEM

For my CSE 841 Project, I am interested in pursuing methods of calculating semantic similarity in short texts. Short texts in particular being a constrained problem where traditional larger document comparison methods underperform[1]. I want to apply and compare a variety of methods for semantic comparison between short sentences, including traditional methods, but with a primary focus on word embeddings discussed in recent literature[2]. It is my belief that this is a reasonably approachable topic with a wide breadth of literature and is suitable for a course project. I plan on using existing implementations for word embeddings (e.g. word2vec) and for some of the resulting text similarity calculations.

2. DATA

I plan on performing my evaluation on the Microsoft Research Paraphrase Corpus [3]. This will involve a process of thresholding similarity comparisons to classify whether or not short descriptions are paraphrases of one another. I may also possibly work with a similar corpus collected from Twitter data[4].

3. REFERENCES

[1] Cedric De Boom, Steven Van Canneyt, q Steven Bohez, Thomas Demeester, and Bart Dhoedt. Learning semantic similarity for very short texts. *CoRR*, abs/1512.00765, 2015.

[2] Tom Kenter and Maarten de Rijke. Short text similarity with word embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1411–1420, New York, NY, USA, 2015. ACM.

[3] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 190–200, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[4] Wei Xu, Chris Callison-Burch, and Bill Dolan. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 1–11, 2015.