CIS 530 — Literature Review

Emily Hilman, Dylan Zuniga Cardenas, Roberta Nin Feliz, Haroni Amare, John Mcgahay
ehilman@sas.upenn.edu, dzu@sas.upenn.edu, ninfeliz@seas.upenn.edu, hamare21@seas.upenn.edu, jmcgahay@sas.upenn.edu

April 22, 2019

Xu et al (2013) approached the problem of gathering and generating paraphrases from Twitter. They first present a way of automatically collecting large paraphrase corpus of tweets by first extracting relevant events from tweets, and then extracting paraphrases within events. In order to extract paraphrases within events, they used Jaccard distance metrics to identify sentences that were similar at the lexical level. They also used their model to normalize noisy text by biasing their model using the New York Times to represent grammatical English. The model showed improvements over various other state-of-the-art paraphrase and normalization models in BLEU-PINC metric curves.

Ganitkevitch et al. (2013) presents the release of a paraphrase database called the PPDB (ParaPhrase DataBase), which includes components for both English and Spanish paraphrases. The paraphrases were extracted using Bannard and Callison-Burch (2005)'s bilingual pivoting method. In essence, given two English strings *e1* and *e2* covered by the same nonterminal symbol C that are both translated with the same string *f* in some foreign language, this method assumes *e1* and *e2* to be paraphrases of one another. For instance, English *thrown into jail* and *imprisoned* may both be translated into German with the pivot word *festgenommen*, so this method takes *thrown into jail* and *imprisoned* to be paraphrases. For any given paraphrase tuple (*e1, e2*), a paraphrase probability could be found like so:

$$p(e_2|e_1) \approx \sum_f p(e_2|f)p(f|e_1).$$

A number of parallel bilingual corpora were used to generate paraphrase databases for both English and Spanish yielding over 160 million paraphrases in both languages. Paraphases were categorized into three types: lexical (for individual word paraphrases, i.e. synonyms), phrasal (for continuous strings of words), and syntactic (expressions containing both words and non-terminal symbols). After Paraphrase extraction was complete, paraphrase pairs were evaluated based on paraphrase probability and cosine

similarity of context signature vectors based on a variety of features (e.g. words seen to left and right, incoming and outgoing dependency link features, part of speech features, etc.). Pruning of the less highly-ranked paraphrases was undertaken until optimal performance was reached on Propbank predicates (recall of 52%, accounting for 97.5% of tokens)

In the paper (Xu et al, 2014), they proposed a new approach to identify paraphrases that performed far better than at the time state-of-the-art models. The new approach these researchers used focused on the concept of an “anchor”, which is a nontrivial word that appears in both sentences with the same topic. Topic means the main subject (say, “Ezekiel”), and the anchor is some other word that appears in both sentences (say, “3D glasses”). This anchor turned out to be a strong indicator that the two sentences were indeed paraphrases. When combined with other state of the art models “a product of experts” led to significantly better results. The previous state of the art had had an fscore of 0.645, whereas their new model boosted to the fscore to 0.724, a significant improvement. This paper also creates a new dataset of annotated paraphrases that they provide for the research community.

Zanzotto et al (2011) present a way to systematically approach the problem of the high level of information redundancy within micro-blogs and social media platforms. They are interested in exploring the this redundancy within the Textual Entailment Recognition. Two tweets are considered redundant if they either convey the same information (paraphrase) or if the information of one tweet is contained in the information of another (textual entailment). In this paper, they describe the system they’ve built which successfully solves the redundancy detection task. They experiment with a few systems but overall, the system that performed the best, uses syntactic feature spaces as effective tools for modeling redundancy, especially when used in first-order rules. They also found that showing that redundancy is pervasive in Twitter, and that methods for its detection will be essential in future Twitter-based applications and data science.

Works Cited

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (NAACL 2013). PPDB: The Paraphrase Database. https://www.aclweb.org/anthology/N13-1092

Xu, W., Ritter, A., and Grishman, R. (ACL Workshop BUCC 2013). Gathering and Generating Paraphrases from Twitter with Application to Normalization. https://www.aclweb.org/anthology/W13-2515

Xu, W., Ritter, A., Callison-Burch, C., Dolan, W. B. and Ji, Y. (TACL 2014). Extracting Lexically Divergent Paraphrases from Twitter. https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacl_a_00194

Zanzotto, F. M., Pennacchiotti, M., and Tsioutsiouliklis, K. (EMNLP 2011). Linguistic Redundancy in Twitter. https://aclweb.org/anthology/D11-1061