Cornell Statement Strength Dataset v1.0 (released April 2014)

Distributed together with:

A Corpus of Sentence-level Revisions in Academic Writing: A Step towards Understanding Statement Strength in Communication
Chenhao Tan, Lillian Lee
In Proceedings of ACL (short papers), 2014

The paper, data, and associated materials can be found at:
http://chenhaot.com/pages/statement-strength.html

If you use this data, please cite:

```
@inproceedings{tan+lee:14,
  author = {Chenhao Tan and Lillian Lee},
  title = {A Corpus of Sentence-level Revisions in Academic Writing: A Step towards Understanding Statement Strength in Communication},
  year = {2014},
  booktitle = {Proceedings of ACL (short papers)}
}
```

Files description:

This dataset contains two files.

matched_sentences.csv:
108,678 aligned sentence pairs from scientific paper abstracts or introductions where the similarity score for the pair was larger than 0.5. Each line contains the arXiv id of the paper, which section the pair is from, and the two sentences of the pair. Arxiv papers can be retrieved from http://arxiv.org .

turker_labels.txt:
Blank lines separate 500 groups of 11 lines.
Each 11-line group has the following format:
  First two lines: the two sentences in a pair
  Next nine lines each have the following format:
    id: (label) comment
The id is a anonymized numerical id for the labelers, employed through Amazon Mechanical Turk.

Please email any questions to: chenhao@chenhaot.com