Brief description of the citation classification dataset

Cailing Dong, Ulrich Schäfer

1. Data source

This dataset is built based on the corpus coming from the ACL (Association for Computational Linguistics) Anthology (http://aclweb.org/anthology). We randomly chose papers from proceedings of the ACL conference in 2007 and 2008. Specific information is listed in the following table:

ACL paper ID	# of distinct citation sentences	Dataset file name
P08-1009 ~ P08-1050	731	P08_1009-50.txt
P07-1001 ~ P07-1050	784	P07_1001-50.txt
P08-2001 ~ P08-2030	253	P08_2001-30.txt

2. File description

All three files have the same format. Each line corresponds to a unique citation item and contains 7 different fields, separated by tab. Following is the detailed information on each field:

1) **CitationID** with the format **year-paperID_sequenceID**. The citationID of one line can be identical with that of another line only if both describe the same reference. (A reference might be cited more than once in the same paper in different citation sentences) .
2) **Citation sentence.** The detailed citation marks are replaced by an empty pair of parentheses.
3) **Part-of-Speech tags sequence.** Each tag corresponds to the word in the citation sentence, in the same order.
4) **Level1 classification label.** {BackGround, Fundamental, Compare}
5) **Level 2 classification label.** {GRelated, SRelated, MRealted, Idea, Basis, Compare}
6) **Sentimental label.** {Positive, Negative, Neutral}

3. Labels (fields 4–6)

We assigned three kinds of labels to each citation sentence. For the specific definition on these labels and the basic annotation guidelines, please read **Annotation_Guidelines.pdf** in this folder.

You can flexibly adjust these labels from different levels. For instance, in our work in IJCNLP'2011, we used four labels {BackGround, Fundamental Idea, Technical Basis, Comparison }.

4. Citation

This dataset is public for research usage. Please use the following reference when you use it in your publication (bibtex):

Cailing Dong, Ulrich Schäfer: Ensemble-style Self-training on Citation Classification. Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP2011), pages 623-631, 2011. ISBN 978-974-466-564-5. Chiang Mai, Thailand. URL http://aclweb.org/anthology/I11-1070.pdf.