Applying Core Scientific Concepts to context-based citation recommendation

Daniel Duma¹, Maria Liakata², Amanda Clare³, James Ravenscroft², Ewan Klein¹

2. Core Scientific Concepts

CoreSC: a sentence-based functional rhetorical classification scheme for scientific documents. Sapientia classifier: 51.9% accuracy over all classes, trained and evaluated on biomedical articles.

4. Methodology

1. Split annotated corpus into document collection and test set for query generation
   a) Index document collection
      All sentences in a document of a same CoreSC type are indexed into the same Lucene document field
   b) Generate queries
      From each citation to a document that is in the collection, generate a query:
      • Extracted query terms (1 sentence up + citing sentence + 1 down, excluding stopwords)
      • CoreSC class of citing sentence = query type
      • Original citation = ground truth

2. Split queries into 4 folds. For each fold:
   a) re-run queries (3/4) adjusting weights one by one until no improvements are found (hill climbing)
   b) test those weights on held-out set (1/4)

6. Results

We show here only the query types for which there is consistent improvement across folds.

We propose that these consistent links between citing and cited sentences can be exploited to increase the relevance of citation recommendation, as well as for scientometrics and summarization.