Applying Core Scientific Concepts to context-based citation recommendation

Daniel Duma[1], Maria Liakata[2], Amanda Clare[3], James Ravenscroft[2], Ewan Klein[1]

THE UNIVERSITY of EDINBURGH informatics

danielduma@gmail.com

1. University of Edinburgh
2. University of Warwick
3. University of Aberystwyth

Key points

Task: recommend potential citations for each sentence of a draft paper.

Evaluation: attempt to recover original citations in existing published papers from the whole document collection.

Approach: automatically label each sentence with its rhetorical function in the document, find links between classes of *citing* sentences and *cited* sentences in the corpus. Index sentences by type for each document and learn per-class weights, conditional on the class of citing sentence.

Corpus: PubMed Central Open Access Subset (10^6 papers)

Annotation scheme: Core Scientific Concepts (CoreSC)

1. Motivation

> A variety of coherence theories have been developed over the years [...] and their principles have found application in many symbolic text generation systems (e.g. CITATION NEEDED)
>
> (Adapted from Barzilay and Lapata, 2005)

Recommendations

- D. Scott, C. S. de Souza. 1990. **Getting the message across in RST-based text generation.** In R. Dale, C. Mellish, M. Zock, eds., *Current Research in Natural Language Generation*, 47–73. Academic Press.
- R. Kibble, R. Power. 2004. **Optimising referential coherence in text generation.** *Computational Linguistics*, 30(4):401–416.
- E. H. Hovy. 1987. **Generating natural language under pragmatic constraints.** *Journal of Pragmatics*, *11*(6), 689-719.
- ...

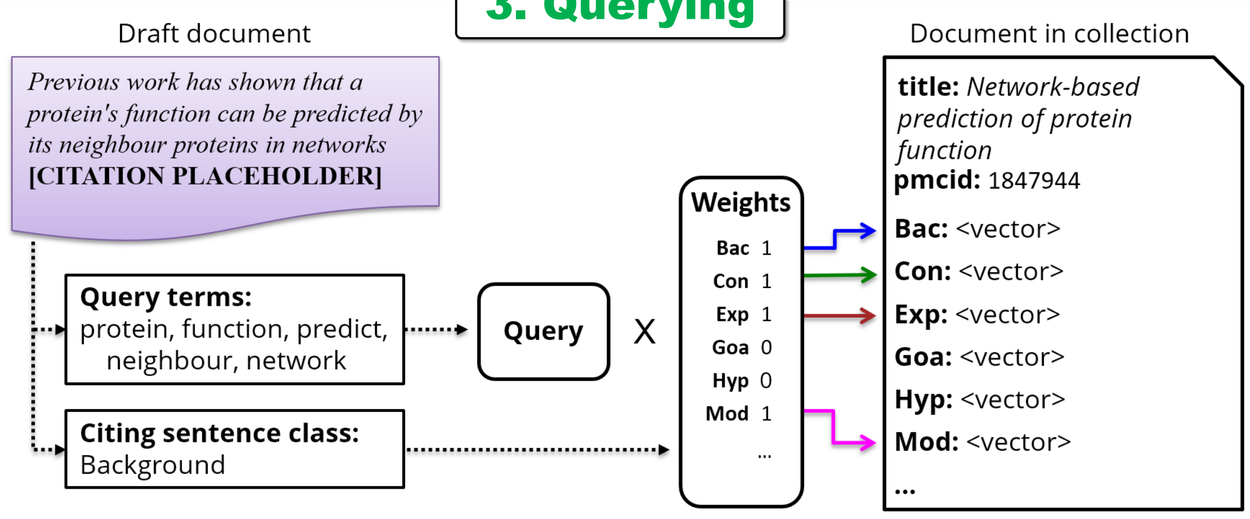
2. Core Scientific Concepts

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

CoreSC class	Description
Hypothesis	A statement not yet confirmed rather than a factual statement
Motivation	The reasons behind an investigation
Background	Generally accepted background knowledge and previous work
Goal	A target state of the investigation where intended discoveries are made
Object-New	An entity which is a product or main theme of the investigation
Method-New	Means by which authors seek to achieve a goal of the investigation
Method-Old	A method mentioned pertaining to previous work
Experiment	An experimental method
Model	A statement about a theoretical model or framework
Observation	The data/phenomena recorded in an investigation
Result	Factual statements about the outputs of an investigation
Conclusion	Statements inferred from observations & results relating to research hypothesis

Hypothesis: there are consistent links between the CoreSC class of *citing* sentences and classes of *cited* sentences.

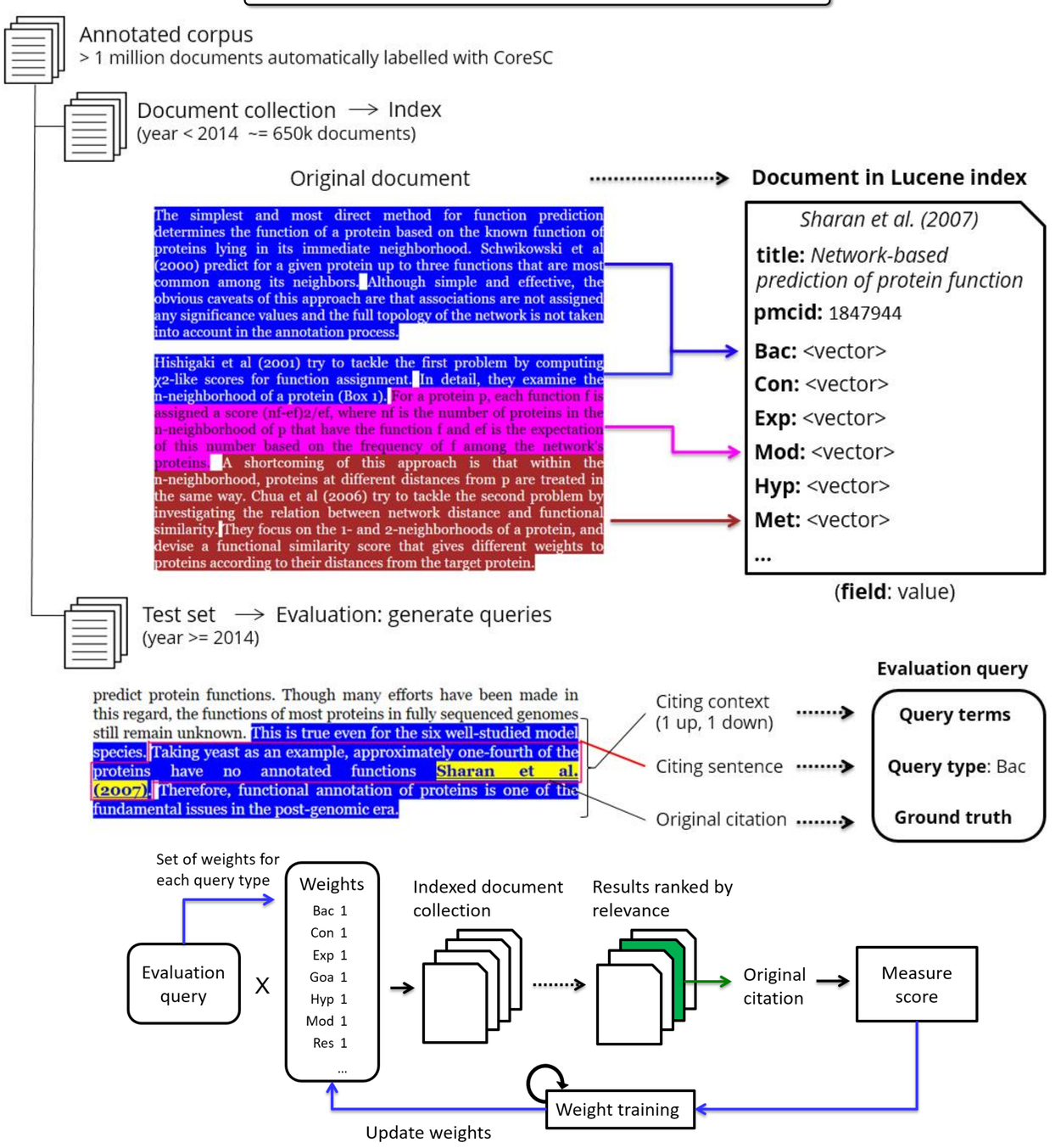
3. Querying



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)

5. Indexing and training



6. Results

Query type	Num queries	Fold	Weights Bac	Con	Exp	Goa	Hyp	Met	Mod	Mot	Obj	Obs	Res	Scores NDCG*	Accuracy*	% imp.
Bac	700	1	1	1	1	0	0	1	1	0	0	0	1	0.401	0.183	28.77
		2	1	0	1	0	0	1	0	0	0	1	1	0.440	0.160	7.56
		3	1	1	0	0	0	1	1	0	0	1	1	0.344	0.109	28.32
		4	1	1	1	0	0	1	1	0	0	0	1	0.437	0.143	24.44
Con	133	1	1	1	1	0	0	0	0	0	0	0	1	0.214	0.059	6.63
		2	1	1	0	0	0	1	1	0	0	0	0	0.528	0.212	16.96
		3	1	0	1	0	0	1	0	0	0	0	0	0.490	0.242	4.24
		4	1	1	1	0	0	1	1	0	0	0	0	0.339	0.152	9.84
Goa	44	1	1	0	1	0	0	1	0	0	0	0	0	0.459	0.182	73.74
		2	1	0	0	0	0	1	0	0	0	0	0	0.126	0.000	46.47
		3	1	0	0	0	0	1	0	0	0	1	0	0.309	0.182	32.38
		4	1	0	0	0	0	1	0	0	0	0	0	0.214	0.091	28.40
Met	602	1	1	0	0	0	0	1	0	0	0	0	1	0.404	0.146	23.70
		2	2	0	0	0	0	0	1	0	0	0	0	0.332	0.139	43.78
		3	2	0	0	0	0	1	0	0	0	0	1	0.467	0.173	18.44
		4	2	0	0	0	0	0	1	0	0	0	0	0.288	0.107	19.28
Obj	65	1	7	0	1	0	0	1	0	0	0	1	1	0.085	0.059	547.57
		2	7	0	1	1	0	1	0	0	0	1	1	0.122	0.063	19.32
		3	7	0	1	0	0	1	0	0	0	1	7	0.114	0.000	91.31
		4	13	0	1	1	0	4	0	0	0	1	0	0.235	0.063	18.30
Res	178	1	1	0	0	0	0	0	0	0	0	1	1	0.379	0.111	33.87
		2	1	0	0	0	0	0	0	0	0	1	1	0.350	0.133	54.18
		3	1	0	0	0	0	0	0	0	0	0	1	0.638	0.273	7.28
		4	1	0	0	0	0	0	0	0	0	0	1	0.362	0.091	30.95

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.

* averaged scores

Taking a majority vote over folds, we can visualize the links between citing and cited sentences: