Context matters: Towards extracting a citation's context using linguistic features

THE UNIVERSITY of EDINBURGH
informatics
danielduma@gmail.com
Daniel Duma, Charles Sutton, Ewan Klein

TL;DR

Aim: recommend potential citations at a particular location in a draft paper.
Task: select the context for which to recommend citations
Evaluation: attempt to recover original citations in existing published papers from the whole document collection
Previous work: traditionally all contexts are extracted using symmetric windows over words or sentences
Approach: compare symmetrical methods for extracting a citation's context: window-of-words and window-of-sentences with a human oracle selecting relevant sentences
Corpus: ACL Anthology Corpus (AAC)

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.
- ...

- All previous work on citation recommendation uses symmetric methods to extract the context of a citation
- Are symmetric methods optimal?

2. Annotated citation contexts

Athar and Teufel (2012) – *Context-Enhanced Citation Sentiment Detection*
- **Corpus**: ACL Anthology
- **Annotated contexts**: ~1800 (citations to 20 selected papers)
- Per-sentence **annotations**:
  - **relevant** (3115 sentences)
  - **sentiment**:
    - (**p**)ositive (261)
    - (**n**)egative (365)
    - (**o**)bjective (2489)

• Most sentences containing a citation are labelled objective. (1929)

3. Evaluation

1. **Index document collection**
   AAC: ~28k documents, excluding annotated documents
2. **Generate queries**
   From each of the annotated citation contexts, remove stopwords and generate one query using:
   - Window of words (30, 50, 100, 500)
   - Window of sentences (1 only, 1 up, 1 down, 1up + 1down, 2up+2down, paragraph)
   - Oracle / human annotations (all relevant, combinations of positive, negative and objective)
3. **Evaluate queries**
   Run queries, attempt to retrieve original citation from document collection, measure Mean Reciprocal Rank (MRR)

4. Context extraction methods

Annotation	Sentence
X	This suggests that the performance which may be obtained for this task may be lower than has been achieved for standard text.
X	Further insight into the task can be gained from determining the degree to which the subjects agreed.
O	**Carletta (1996)** argues that the kappa statistic (a) should be adopted to judge annotator consistency for classification tasks in the area of discourse and dialogue analysis.
X	It is worth noting that the problem of sentence boundary detection presented so far in this paper has been formulated as a classification task in which each token boundary has to be classified as either being a sentence boundary or not.
O	Carletta argues that several incompatible measures of annotator agreement have been used in discourse analysis, making comparison impossible.
O	Her solution is to look to the field of content analysis, which has already experienced these problems, and adopt their solution of using the kappa statistic.

Extraction methods

Window of **tokens** (30 up, 30 down)

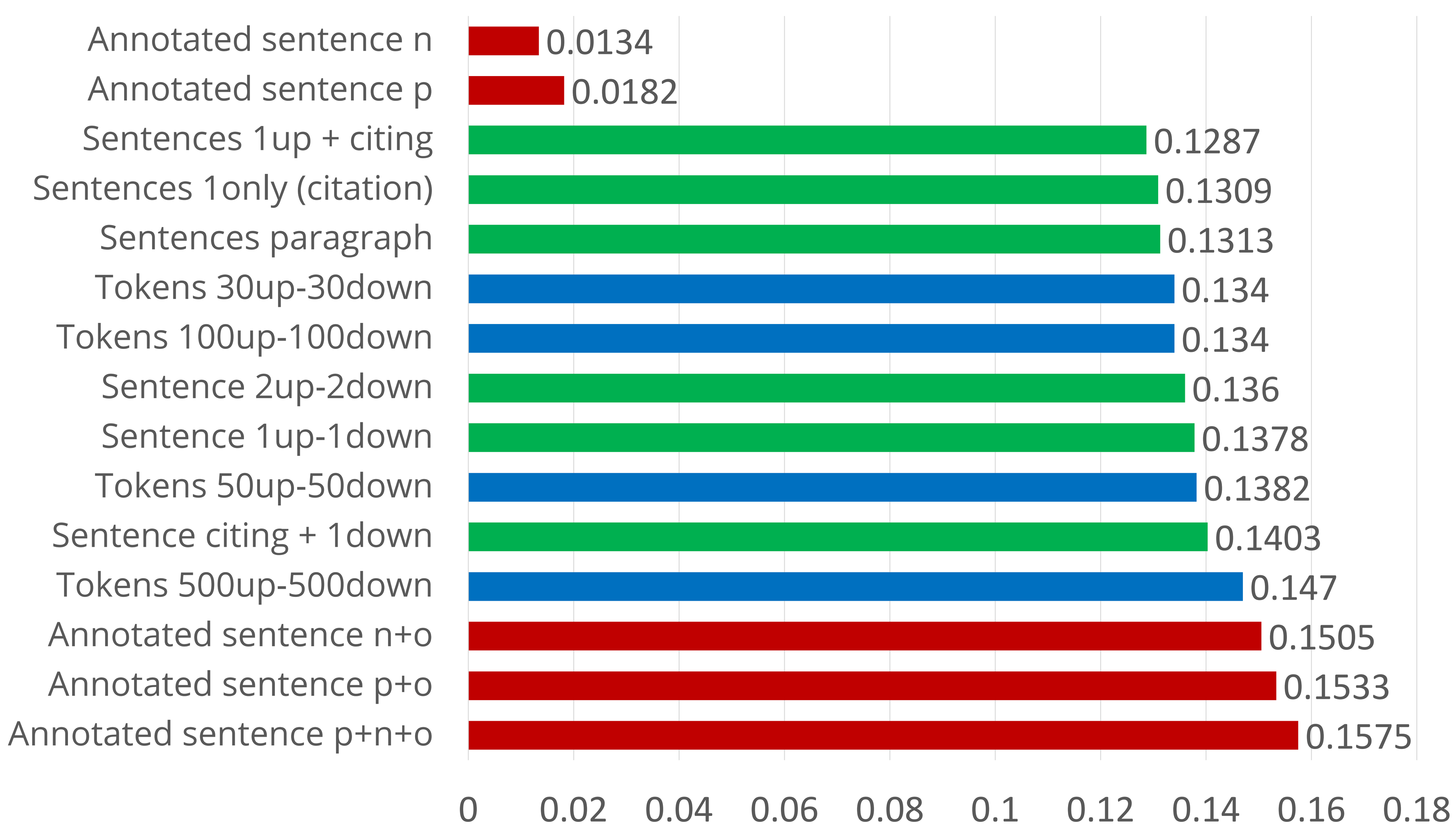
Window of **sentences** (2 up, 2 down)

Oracle - **annotated sentences** (p + n + o)

(from Stevenson and Gaizauskas (2000) - *Experiments on Sentence Boundary Detection*)

5. Results

Evaluation: Mean Reciprocal Rank



6. Discussion

Findings:
- Human oracle outperforms all symmetrical methods. Symmetrical windows of either tokens or sentences are therefore not optimal.
- The annotated sentiment of sentences was not useful for query extraction. The more sentences we include that were annotated as relevant, the higher the score.
- More query terms is not always better. Carefully selecting relevant text spans for context extraction improves results.

Future work: keyword extraction using linguistic features. Train a machine learning classifier to generate queries from sub-sentence-length spans.