Discussion for the Shared Task at TAC 2015

Kokil Jaidka*, Muthu Kumar Chandrasekaran

* Nanyang Technological University, Singapore
koki0001@e.ntu.edu.sg

TAC 2015: CL-Summ task rollout

We propose to have a full-fledged official shared task for an expanded CL corpus

- 20 training topics
- 10 test topics
- 4 annotators per summary

We will need 120 sets annotated!

Checklist

- Human Resources
  - Coordinator(s)
  - Assignment / Responsibilities
  - Funding
- Timeline
- Dataset Preparation
- Evaluation Plan

- Kokil's slides go here

TAC 2015: CL-Summ task rollout

Important dates:

- Expression of interest: End Jan
- Go / no go decision: Early Feb
- Release of training data: End May, 2015
- Participant registration closes: June, 2015
- Release of test data: Early Sep, 2015

TAC 2015: We need your help

- We welcome all volunteer efforts
- We'd appreciate resources – human or software
- It would be great if all participating teams contribute
- Universities may get funding to run annotation efforts
- We seek support from the summarization community, the CL community in particular
- Because, together we can!

How you can help

Option 1. Would you like share the responsibility for annotating?

- What would this involve:
  "every team provides number of annotations"
- Pooled evaluation, à la Cranfield.
- Estimation of effort:
  - 4 annotators each for 30 sets = 120 sets to be annotated
  - 4 man-hours for annotating 1 set (10-15 citances per set)

How you can help

Option 2: Run the annotation task at your institution

- We, and volunteers from your team use instructions to school Master's / PhD students for annotating; Aim: train and select annotators
- Total 4 to 6 hours' training spread over a few days
  - Some walkthroughs
  - Iterative rounds of coding/discussion
  - A "test" coding task and selection of final coders
  - Monthly status update meetings to supervise

Evaluation Metric

- Task 1A: Overlap vs Rouge
- Task 2: Evaluate over human summaries
  - from the paper vs from the citances
  - difficulties in getting the annotators to do this

REPEATED / BACKUP SLIDES

CL-Summ Pilot: highlights

First corpus in the computational linguistics community incorporating prior research on citation based summaries

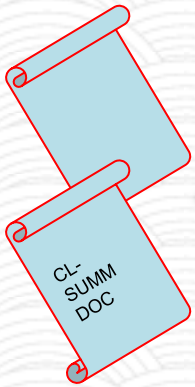
- 10 teams registered
- 3 teams participated in the evaluation
- 2 teams submitted their systems' performance
- 1 more proposed algorithms to solve the tasks

CL training Corpus

- 10 reference papers or topics randomly sampled from the ACL live anthology
- Up to 10 citing papers per reference paper including those outside ACL live anthology
- Annotated corpus publicly available

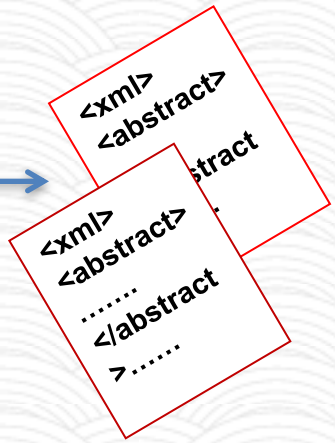
https://github.com/WING-NUS/scisumm-corpus/

Annotation Pipeline

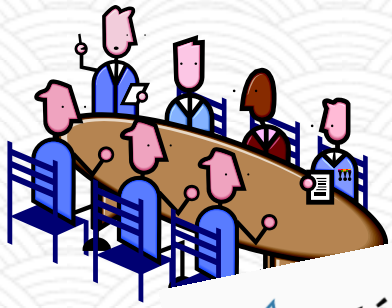


OCR & Section Parse

ParsCit's SectLabel module



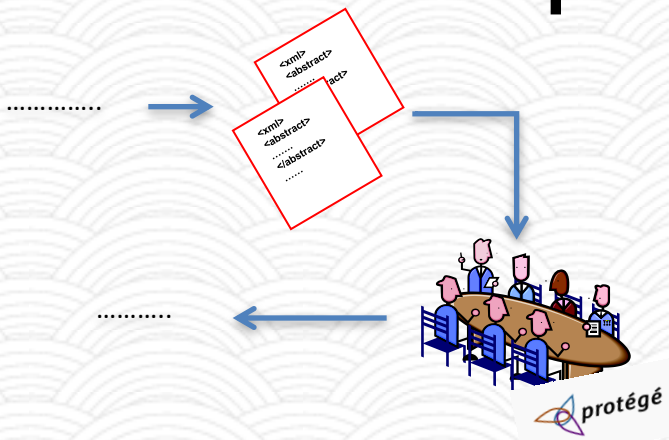
Annotation!



protégé

Post Processing to BiomedSumm format

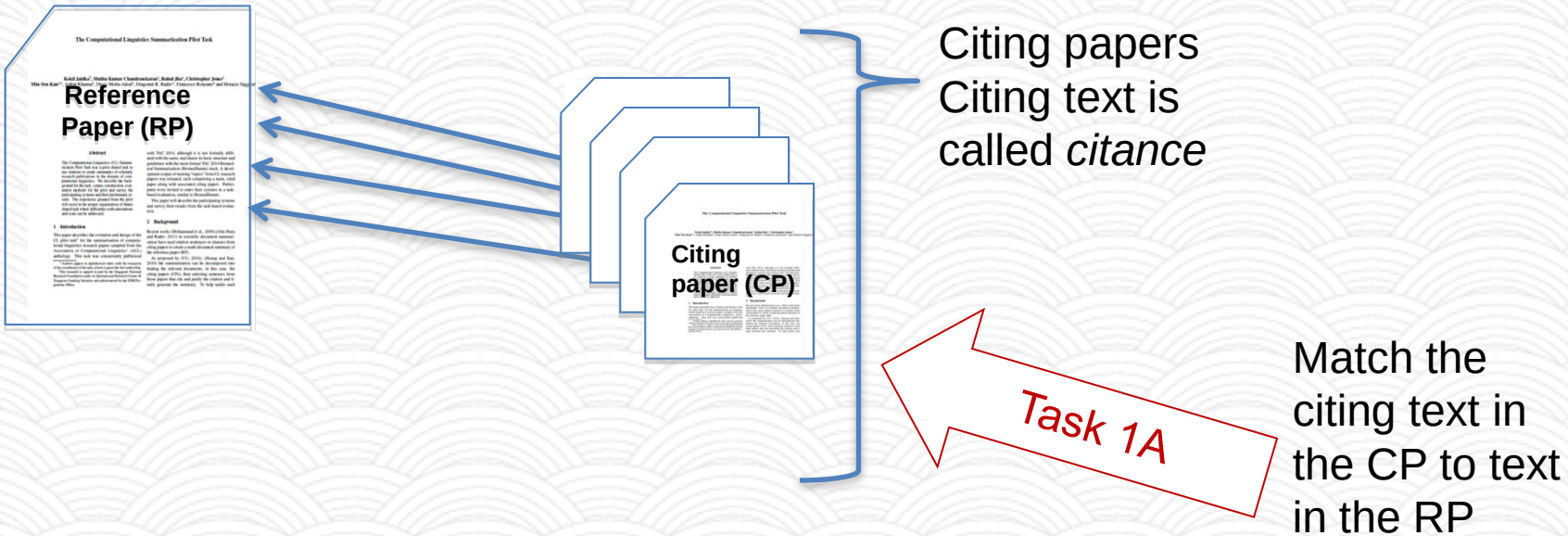
Annotating the SciSumm corpus



- 3 annotators
- Released data has one gold standard annotation per topic or reference paper
- Discourse facet has a minor change from Biomedsumm's categories

Tasks

Task 1A: Identify the text span in the RP which corresponds to the *citances* from the CP.

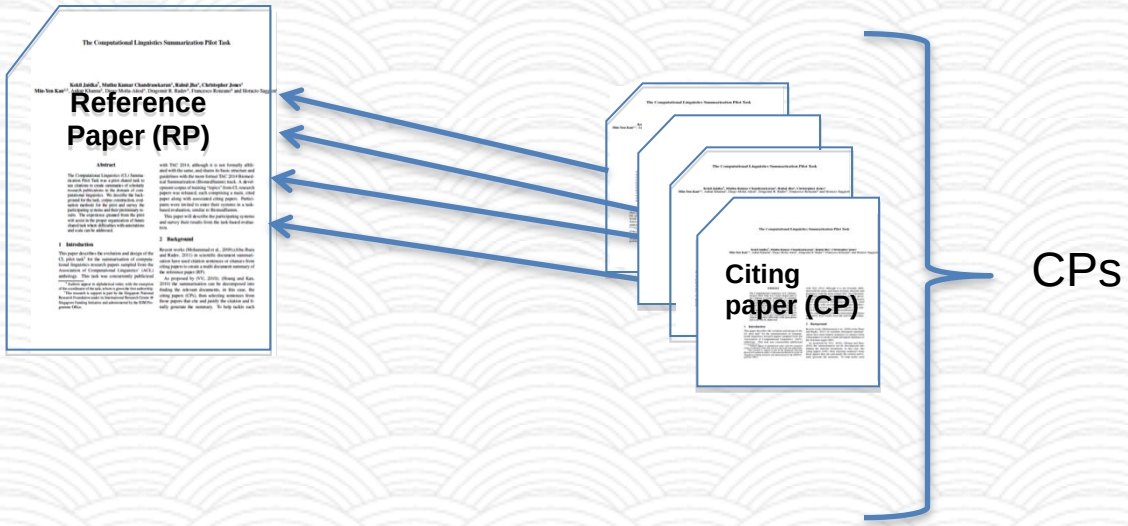


Tasks

Task 1B: Identify the discourse facet for every cited text span from a predefined set of facets.

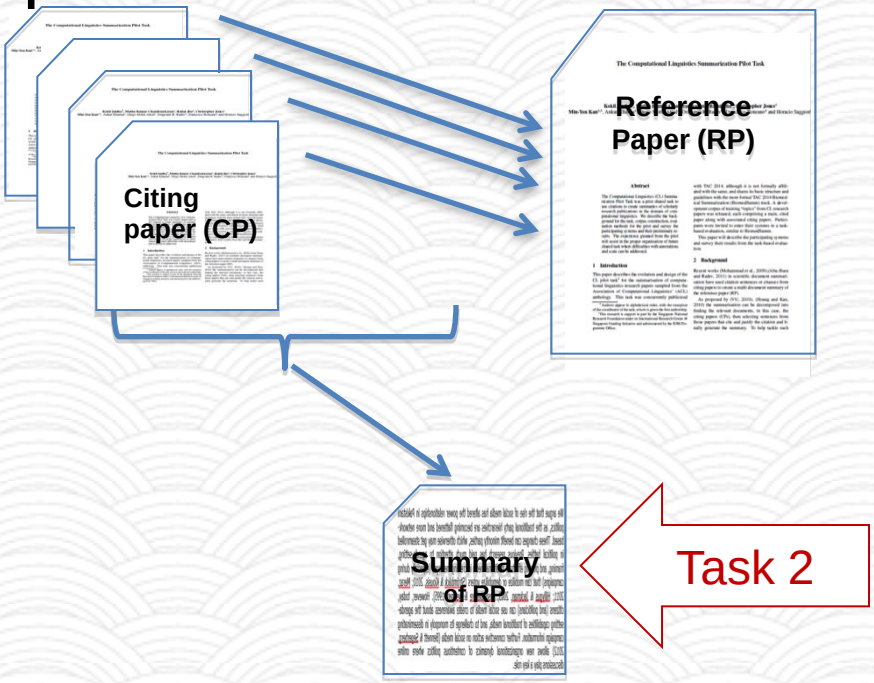
Classify the cited text in RP into one of several facets

Task 1B



Tasks

Task 2: Generate a faceted summary of up to 250 words, of the reference paper, using itself and the citing papers.



Use citances and the RP to create a summary

Evaluation

Small corpus: 10 fold cross validated evaluation over the 10 documents

- Task 1A scored by ROUGE-L metric
- Task 1B scored by classification metrics: Precision, Recall and F_1
- Task 2 also scored by ROUGE-L metric

Results – Task 1A

MQ			Clair_UMich		
Precision	Recall	F_1	Precision	Recall	F_1
0.212	0.335	0.223	0.444	0.574	0.487

- MQ was unsupervised while Clair_Umich was supervised
- Challenging classification problem: Task seeks to map each citation sentence with a few out of 100s of potential matches in the Reference paper (RP)
- Lexical, semantic and structural similarities between citances and RP sentences somewhat help

Results – Task 1A

Paper ID	MQ	Clair_UMich
C90_2039	0.235	0.635
C94_2154	0.288	0.536
E03_1020	0.239	0.478
H05_1115	0.350	0.375
H89_2014	0.332	0.546
J00_3003	0.196	0.559
J98_2005	0.101	0.344
N01_1011	0.221	0.498
P98_1081	0.200	0.367
X96_1048	0.248	0.535

Large deviation in scores, across topics, from both systems

Results – Task 2

Paper ID	MQ (using Task 1A MMR)
C90_2039	0.293
C94_2154	0.120
E03_1020	0.196
H05_1115	0.321
H89_2014	0.320
J00_3003	0.367
J98_2005	0.233
N01_1011	0.284
P98_1081	0.206
Average	**0.260**

ROUGE-L scores here measure overlap over the abstract since we did not have human summaries

Low scores could be due to deviation between summary of citances and the abstract of the paper

Errors – Task 1A

Citing text: "The line of our argument below follows a proof provided in… for the maximum likelihood estimator based on nite tree distributions."

Clair_UMich

False negative: "We will show that in both cases the estimated probability is tight."

MQ

Target text from RP: "*The work described here also makes use of hidden Markov model.*"

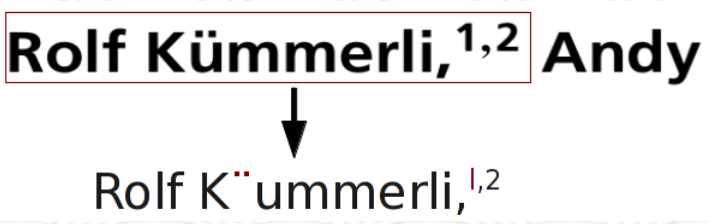
*False positive: "*The statistical methods can be described in terms of Markov models."

Learning from the Pilot Task

- Offset mismatch between the text file and the XML that annotators used
  - Corpus sentence segmented and sentences assigned a sentence ID
- Problems with handling non-contiguous spans in Protégé
- Character offsets can be miscounted by different parsers
- Handling non-UTF8 characters

Limitations of this corpus

- No gold standard citation based summaries
- OCR errors:



- The use of "..." where text spans are snippets
- Errors in citation/reference offset numbers
- Different text encodings
- Errors in file construction
- Small size of corpus!