The Computational Linguistics Summarization Pilot Task

Introduction

- Provides resources aimed to encourage research on scientific document summarization specially summarization utilizing the citances
  - Citence means “sentence(s) surrounding the citation within a document (Nakov et al. [1])”
- Constructed by sampling papers from ACL
- Has released training corpus of CL research papers

[1] Nakov, Preslav I., Ariel S. Schwartz, and Marti Hearst. "Citances: Citation sentences for semantic analysis of bioscience text." In Proceedings of the SIGIR, vol. 4, pp. 81-88. Citeseer, 2004.

Introduction > Existing approaches

Scientific article summarization systems

- Automatically generating related work sections for a target paper via hierarchical topic (Hoang and Kan [2])
- Generating model citation sentences (Mohammad et al., 2009)
- Implementing a literature review framework

[2] Hoang, Cong Duy Vu, and Min-Yen Kan. "Towards automated related work summarization." In Coling 2010: Posters, pp. 427-435. 2010.
[3] Mohammad, Saif, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishnan, Vahed Qazvinian, Dragomir Radev, and David Zajic. "Using citations to
[4] generate surveys of scientific paradigms." In Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, pp. 584-592. 2009.Jaidka, Kokil, Christopher Khoo, and Jin-Cheon Na. "Deconstructing human literature reviews–a framework for multi-document summarization." In *proceedings of the 14th European workshop on natural language generation*, pp. 125-135. 2013.

Introduction > Existing approach limitations

- However, limited evaluation resources – i.e., human-created summaries means that the efficacy of these approaches can not be verified by others, hurting the replicability of works in this domain.

Introduction > Main steps in existing approaches

1. Finding relevant documents (CPs)
2. Selecting sentences which justify the citation in the reference paper (RP)
3. Generating summary

Corpus Construction

- Source : ACL Anthology
- 25k publications in Anthology (on 18th September 2014)
- Randomly sampled RPs using the following procedure.
  - Published on or after 2006. => leaving only 13.8k publications. Randomized the order.
  - Used google web and google scholar searches to approximate the number of CPs for RP. Retained any paper as an RP if it was reported over 10 citations.
  - Vetted the citations to ensure that the citation spread was at least a window of three years(previous work indicates that citations over different time periods (with respect to the publication date of the RP) exhibit different tendencies (Abu-Jbara et al. [5])

[5] Abu-Jbara, Amjad, Jefferson Ezra, and Dragomir Radev. "Purpose and polarity of citation: Towards nlp-based bibliometrics." In Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies, pp. 596-606. 2013.

Corpus Construction

- For every RP, they aimed to provide at least three CPs
  - Non-list citation
  - The oldest and newest citations within AAN
  - Citations from different years

CL-Summ Task

- Proposes to solve the same problem as [6]

Define the task as:

- A topic, comprising of the PDF and extracted text of an reference paper (RP) and up to 10 citing papers (CPs). In each provided CP, the citations to the RP (or citances) have been identified and manually annotated. The information referenced in the RP is also annotated.
- Systems are required to do,
    - Identify the text span in the RP. It may be a sentence fragment
        - Evaluate using rouge
    - Identify the discourse facet for every cited text span from a predefined set of facets.
    - Generate a facet summary of reference paper of up to 250 worlds.

[6] Cohan, Arman, Luca Soldaini, and Nazli Goharian. "Matching citation text and cited spans in biomedical literature: a search-oriented approach." In proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies, pp. 1042-1048. 2015.

Corpus Construction

- Original text was not sentence-segmented
- Perform the following
  - Automatic pdf-to-text conversion
  - Manual verification of output
  - Sentence splitter and sentence sanitizer
  - Mapping annotations to clean textual versions

Comparing overlap of word synsets

Baseline system based on the basic information retrieval measure:

Baseline:

- Term frequency * inverse document frequency (TF.IDF)

Comparing overlap of word synsets

Supervised System:

- **Lexical features**: 2 lexical features were used- TF.IDF and LCS (longest common subsequence) between the citing sentence (C) and reference sentence (S) which is computed as $\frac{|LCS|}{min(|C|,|S|)}$
- **Knowledge base features**: 6 wordnet based similarity measures were combined to obtain a six sentence similarity features.

$$sim_{wn}(C,R) = \frac{(\omega + \sum_{i=1}^{|\phi|} \phi_i) * (2|C||R|)}{|C| + |R|}$$

Here ω is the number of shared senses between C and R. The list φ contains the similarities of non-shared words in the shorter text, φi is the highest similarity score of the ith word among all the words of the lower text

Comparing overlap of word synsets

Supervised System:

- Syntactic Features: given a candidate sentence pair, two syntactic dependencies were considered equal if they had the same dependency type. if Rc and Rr are the set of all dependency relations in C and R, the dependency overlap score was computed using the formula:

$$sim_{dep}(C, R) = \frac{2 * |R_c \cap R_r| * |R_c||R_r|}{|R_c| + |R_r|}$$

Finding the best fit to citance

- Given the text of a citance, the MQ system ranked the sentences of the reference paper according to its similarity to the citance. Every sentence and its ci- tance was modeled as a vector and compared using cosine similarity.

$$\text{MMR} = \arg \max_{D_i \in R \setminus S} \left[\lambda(\text{sim}(D_i, Q)) - (1 - \lambda) \max_{D_j \in S} \text{sim}(D_i, D_j) \right]$$

Where:

- Q is the citance text.
- R is the set of sentences in the document.
- S is the set of sentences that haven been chosen in the summary so far.

Figure 1: Maximal Marginal Relevance (MMR) algorithm, as used in the *MQ* system.

Identifying the discourse facet of the cited text spans

- This is a sentence classification problem.
- This resulted in a set of 266 CP sentences as distributed

Docset	Citing papers
Aim	46
Hypothesis	1
Implication	25
Results	29
Method	165
TOTAL:	266

Table 1: Discourse facet of the sentences of cited papers belonging to a manually annotated reference text span.

Evaluation and Results

	P	R	F_1
MQ	0.212	0.335	0.223
clair_umich	0.444	0.574	0.487
TALN.UPF	0.194	0.344	0.225

Table 2: Task 1A performance for the participating systems expressed as ROUGE-L score averaged over all topics.

Evaluation and Results

Paper ID	MQ	clair_umich	TALN.UPF
C90-2039	0.235	0.635	0.180
C94-2154	0.288	0.536	0.200
E03-1020	0.239	0.478	0.198
H05-1115	0.350	0.375	0.233
H89-2014	0.332	0.546	0.275
J00-3003	0.196	0.559	0.263
J98-2005	0.101	0.344	0.196
N01-1011	0.221	0.498	0.254
P98-1081	0.200	0.367	0.211
X96-1048	0.248	0.535	0.240

Table 3: Task 1A ROUGE-L F_1 scores for individual topics.

Evaluation and Results

Paper ID	MQ	clair_umich	TALN.UPF
C90-2039	0.235	0.635	0.180
C94-2154	0.288	0.536	0.200
E03-1020	0.239	0.478	0.198
H05-1115	0.350	0.375	0.233
H89-2014	0.332	0.546	0.275
J00-3003	0.196	0.559	0.263
J98-2005	0.101	0.344	0.196
N01-1011	0.221	0.498	0.254
P98-1081	0.200	0.367	0.211
X96-1048	0.248	0.535	0.240

Table 3: Task 1A ROUGE-L F_1 scores for individual topics.

Evaluation and Results

Discourse facet	NB	SVM	**LR**
Aim	0.725	0.734	**0.732**
Method	0.706	0.826	**0.828**
Implication	0.049	0.000	**0.200**
Results	0.509	0,533	**0.533**
Hypothesis	0.024	0.000	**0.000**
WEIGHED AVG. F_1	0.623	0.698	**0.719**

Table 4: Task 1B self-evaluation for TALN.UPF: F_1 classification performance comparison.

Evaluation and Results

Paper ID	TF.IDF	Task 1A TF.IDF	Task 1A MMR	Paper ID	TF.IDF	Task 1A TF.IDF	Task 1A MMR
C90-2039_TRAIN	0.347	0.315	0.293	J00-3003_TRAIN	0.221	0.382	0.367
C94-2154_TRAIN	0.095	0.123	0.120	J98-2005_TRAIN	0.221	0.216	0.233
E03-1020_TRAIN	0.189	0.189	0.196	N01-1011_TRAIN	0.187	0.268	0.284
H05-1115_TRAIN	0.134	0.306	0.321	P98-1081_TRAIN	0.241	0.210	0.206
H89-2014_TRAIN	0.294	0.319	0.320	Average	0.214	0.259	0.260

Table 5: ROUGE-L F_1 results for summaries generated by the MQ system.

	CL-Summ				BiomedSumm			
Run	P	R	F_1	CI	P	R	F_1	CI
TF.IDF	0.198	0.316	0.211	0.185–0.240	0.326	0.273	0.279	0.265–0.293
topics	0.201	0.324	0.217	0.191–0.245	0.357	0.288	0.300	0.285–0.316
context	0.214	0.339	0.225	0.197–0.255	0.372	0.291	0.308	0.293–0.323
MMR	0.212	0.335	0.223	0.195–0.251	0.375	0.290	0.308	0.293–0.323

Table 6: ROUGE-L results of the MQ system runs for Task 1A.

Shortcomings and Limitations

- With different text formats (UTF-8, Windows-1252, GB18030), thus making difficult the implementation of an automated homogeneous textual processing pipeline.
- Some of the older PDF files, when parsed to text or XML, presented several text formatting issues: hyphenation problems, words not separated by blank spaces, page headers and footnotes included in the textual flow, misspelled words, spaces within words, sentences in the wrong place and so on.
- Errors in citation / reference offsets: In the original annotations
- Discontiguous texts
- Small corpus
- Errors in file construction: An automatic, open-source software was used to map the citation annotations from the adopted annotation software, Protege, to a text file. However, participants identified several errors in the output

Conclusion

- 3 Systems were participated in the CL pilot task consisting of,
  - Task 1A
  - Task 1B
  - Task 2
- All teams used version of TF.IDF as the baseline.
- For the citation span identification task, MQ and TALN.UPF implemented unsupervised algorithms
- clair_umich system decided on a supervised approach.
- in this first task, clair_umich's supervised algorithm performed best, using lexical, syntactic and knowledge-based features to calculate the overlap between sentences in the citation span and the reference paper
- TALN.UPF attempted the second part of the task (Task 1B) for identifying the discourse facet being cited. They compared the performance of three sentence classifiers and found that the best performance was obtained using logistic regression on lemmatized word features
- Task 2 was attempted by the MQ team.In comparison with MQ's results on the BioMedSumm task, the results were inconclusive to state whether or not the system's features were actually aiding in generating better gold standard summaries.
- Methods that worked in the biomedical domain do not seem to have fared well in computational linguistics

References

[1] Nakov, Preslav I., Ariel S. Schwartz, and Marti Hearst. "Citances: Citation sentences for semantic analysis of bioscience text." In Proceedings of the SIGIR, vol. 4, pp. 81-88. Citeseer, 2004.
[2] Hoang, Cong Duy Vu, and Min-Yen Kan. "Towards automated related work summarization." In Coling 2010: Posters, pp. 427-435. 2010.
[3] Mohammad, Saif, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishnan, Vahed Qazvinian, Dragomir Radev, and David Zajic. "Using citations to generate surveys of scientific paradigms." In Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, pp. 584-592. 2009.
[4] generate surveys of scientific paradigms." In Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, pp. 584-592. 2009.Jaidka, Kokil, Christopher Khoo, and Jin-Cheon Na. "Deconstructing human literature reviews–a framework for multi-document summarization." In proceedings of the 14th European workshop on natural language generation, pp. 125-135. 2013.
[5] Abu-Jbara, Amjad, Jefferson Ezra, and Dragomir Radev. "Purpose and polarity of citation: Towards nlp-based bibliometrics." In Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies, pp. 596-606. 2013.
[6] Cohan, Arman, Luca Soldaini, and Nazli Goharian. "Matching citation text and cited spans in biomedical literature: a search-oriented approach." In proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies, pp. 1042-1048. 2015.

Thank You