CORWA: A Citation-Oriented Related Work Annotation Dataset

Xiangci Li, Biswadip Mandal & Jessica Ouyang
University of Texas at Dallas

Introduction

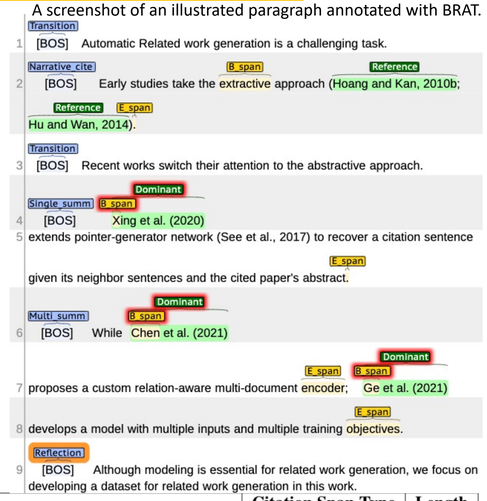
- Ultimate goal: Automatically generating related work sections.
- The task of automatic related work generation is challenging and not standardized yet. Each prior work has different task settings, inputs and outputs. Most prior works only introduce their ad-hoc datasets as a byproduct of their modeling work.
- Previous tasks all generated related work by sentences or paragraphs, which neglects the natural structure of the related work section that consists of multiple variable lengths of citation spans and other supporting sentences.
- We collect Citation-Oriented Related Work Annotation (CORWA) dataset, which is a dataset designed for related work generation task. CORWA consists of 3 tasks: discourse tagging, citation type recognition and citation span detection.
- We propose a strong Transformer-based baseline tagger to jointly tag 3 tasks of CORWA. Using this tagger, we can easily generate massive training data for related work generation from unlabeled texts of related work sections.
- Applications of CORWA
  - Serving as one of the pre-training tasks of the language model for related work generation.
  - Citation span is a more natural generation objective than citation sentences (In progress).
  - Condition the writing style (summarization vs. narrative) of the generated citation text (In progress).

CORWA Dataset

Annotation Scheme

- Discourse tagging
  - Sentence-level tagging
  - Summarization vs. narrative
  - Distinguish citation vs. non-citation
- Citation type recognition
  - Reference
    - Cited for reference.
  - Dominant
    - Main citation of the sentence.
- Citation span detection
  - The influenced span of the cited work.
  - A few words ~ a few sentences.
  - Dominant vs. Reference

A screenshot of an illustrated paragraph annotated with BRAT.

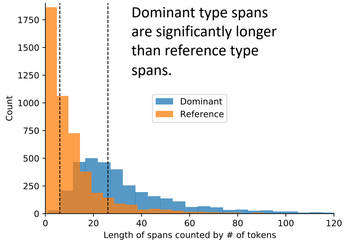


Discourse label	Definition	Citation Span Type	Length
single_summ	Summarization of single prior work	*dominant*	Long
multi_summ	Summarization of multiple prior works	*dominant*	Long
narrative_cite	Narrative sentence with citations	*reference*	Short
reflection	Sentence that focuses on the current work	N/A	N/A
transition	Non-citation topic or transition sentence	N/A	N/A
other	Mistakenly included in related work section	N/A	N/A

Table 2: A summry of discourse label types in CORWA, and their concurring citation span types.

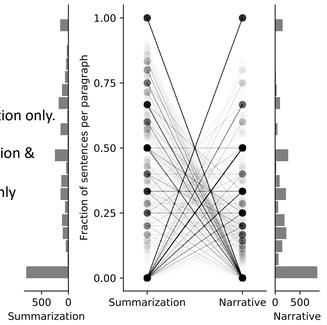
CORWA Dataset

Statistics



Histogram of the length of dominant (n=3303) & reference (n=4738) type citation spans excluding citation marks counted by number of tokens. The dashed vertical lines are the means of reference and dominant spans with the values of 11.1 and 33.4 respectively.

- Summarization only.
- Mixture of summarization & narrative
- Narrative only



Parallel plot of the fraction of summarization and narrative sentences in each paragraph (n=2434). The histograms show the distributions of the fraction of summarization and narrative sentences respectively. Paragraphs that contain no summarization or narrative sentences are excluded.

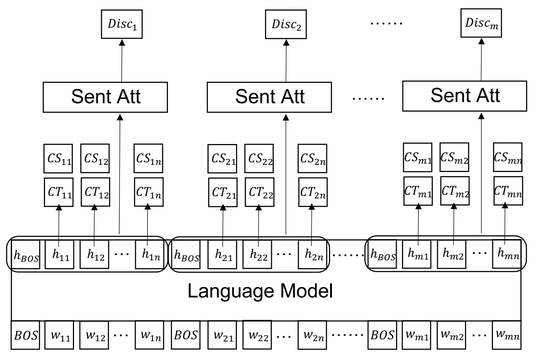
Disc. Label (d)	$n(d)$	$p(d)$	$p(d \mid D)$	$p(d \mid R)$	$p(D \mid d)$	$p(R \mid d)$	$p(D, d)$	$p(R, d)$
single_summ	4388	30.7%	72.8%	3.5%	91.0%	9.0%	**29.9%**	2.1%
transition	3495	24.5%	0	0.1%	0	100.0%	0	0
narrative_cite	2622	18.3%	0.4%	88.5%	0.2%	99.8%	0.2%	**52.2%**
reflection	2576	18.0%	0.1%	3.7%	1.3%	98.7%	0	2.2%
multi_summ	698	4.9%	26.7%	4.1%	76.2%	23.9%	**11.0%**	2.4%
other	513	3.6%	0.5%	0	0	0	0	0

Table 4: Distributions of discourse labels and citation spans in CORWA dataset. d: Discourse labels. D/R: Dominant/reference type citation span. $n(D) = 3303, n(R) = 4738$.

Joint Tagger for CORWA

Model Architecture

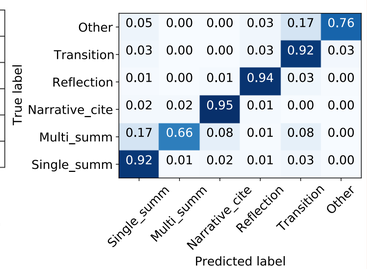
Discourse tagging, citation type recognition and citation span detection share the same language model as the encoder of the input paragraph. Citation type and span labels are predicted separately token-by-token. Discourse labels are predicted on top of a sentence-level attention.



Experiments

Language Model	Disc	CT	CS
SciBERT	**0.916**	**0.967**	**0.935**
Roberta-base	0.902	0.963	0.929
BERT-base	0.896	0.962	0.924
LED-base (Pretrained)	0.886	0.956	0.921
LED-base	0.882	0.954	0.917

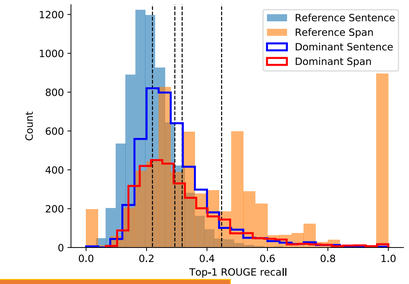
Table 3: Test micro-F1 of joint related work tagger with different language models on discourse tagging (disc), citation type tagging (CT) and citation span tagging (CS).



Related Work Writing Analysis

Citation Span for ROUGE-based Retrieval

Histogram of top-1 ROUGE recall scores of retrieved sentences from cited papers using different types of queries. The dashed vertical lines are the means of reference sentence (0.220), dominant sentence (0.294), dominant span (0.318) and reference spans (0.448).



Frequent Discourse Label Subsequences

Functionalities	Discourse Sequence	Examples
Introducing an approach and providing background knowledge.	Transition, Narrative_cite, Single_summ	1. Joint POS tagging with parsing is not a new idea. 2. In PCFG-based parsing (Collins, 1999; Charniak, 2000; Petrov et al., 2006), POS tagging is considered as a natural step of parsing by employing lexical rules. 3. For transition-based parsing, Hatori et al. (2011) proposed to integrate POS tagging with dependency parsing.
Comparing the prior work to the current work.	Single_summ, Reflection	1. Haghighi et al. (2009) confirm and extend these results, showing BLEU improvement for a hierarchical phrasebased MT system on a small Chinese corpus. 2. As opposed to ITG, we use a linguistically motivated phrase-structure tree to drive our search and inform our model.
Supporting the current work with a previous work	Reflection, Single_summ	1. Our baseline semisupervised model can be viewed as an extension of these approaches to a reading comprehension setting. 2. Dai et al. (2015) also explore initialization from a language model, but find that the recurrent autoencoder is superior, which is why we do not consider language models in this work.
Topic sentence, narration of prior work followed by critique.	Transition, Narrative_cite, Transition	1. Traditional work on relation classification can be categorized into feature-based methods and kernelbased methods. 2. The former relies on a large number of human-designed features (Zhou et al., 2005; Jiang and Zhai, 2007; Li and Ji, 2014) while the latter leverages various kernels to implicitly explore a much larger feature space (Bunescu and Mooney, 2005; Nguyen et al., 2009). 3. However, both methods suffer from error propagation problems and poor generalization abilities on unseen words.
Commenting previous works summarized.	Single_summ, Single_summ, Transition	1. Walker et al. (2012) extract rules representing characters from their annotated movie subtitle corpora. 2. Miyazaki et al. (2015) propose a method of converting utterances using rewriting rules automatically derived from a Twitter corpus. 3. These approaches have a fundamental problem to need some manual annotations, which is a main issue to be solved in this work.
• Criticizing the previously cited work and citing an improved work. • Describing an idea following by a comment and then citations implementing the idea.	Narrative_cite, Transition, Single_summ	1. There have also been several classical studies based on nonneural approaches to headline generation (Woodsend et al., 2010; Alfonseca et al., 2013; Colmenares et al., 2015) , but they basically addressed sentence compression after extracting important linguistic units such as phrases. 2. In other words, their methods can still yield erroneous output, although they would be more controllable than neural models. 3. One exception is the work of Alotaiby (2011) , where fixed-sized substrings were considered for headline generation. 1. One of the classes of errors in the Helping Our Own (HOO) 2011 shared task (Dale and Kilgarriff, 2011) was punctuation. 2. Comma errors are the most frequent kind of punctuation error made by learners. 3. Israel et al. (2012) present a model for detecting these kinds of errors in learner texts.

Frequent discourse label subsequences detected by applying PrefixSpan and Gap-Bide algorithm.