Towards Automated Related Work Summarization

Cong Duy Vu Hoang and **Min-Yen Kan**
Department of Computer Science
School of Computing
National University of Singapore
{hcdvu,kanmy}@comp.nus.edu.sg

Abstract

We introduce the novel problem of automatic related work summarization. Given multiple articles (*e.g.*, conference/journal papers) as input, a related work summarization system creates a topic-biased summary of related work specific to the target paper. Our prototype **Re**lated **Wo**rk **S**ummarization system, **ReWoS**, takes in set of keywords arranged in a hierarchical fashion that describes a target paper's topics, to drive the creation of an extractive summary using two different strategies for locating appropriate sentences for general topics as well as detailed ones. Our initial results show an improvement over generic multi-document summarization baselines in a human evaluation.

1 Introduction

In many fields, a scholar needs to show an understanding of the context of his problem and relate his work to prior community knowledge. A related work section is often the vehicle for this purpose; it contextualizes the scholar's contributions and helps readers understand the critical aspects of the previous works that current work addresses. Creating such a summary requires the author to position his own work within the contextual research to showcase the advantages of his method.

We envision an NLP application that assists in creating a related work summary. We propose this *related work summarization* task as a challenge to the automatic summarization community. In its full form, it is a topic-biased, multi-document summarization problem that takes as input a target scientific document for which a related work section needs to be drafted. The output goal is to create a related work section that finds the relevant related works and contextually describes them in relationship to the scientific document at hand.

We dissect the full challenge as bringing together work of disparate interests; 1) in finding relevant documents; 2) in identifying the salient aspects of these documents in relation to the current work worth summarizing; and 3) in generating the final topic-biased summary. While it is clear that current NLP technology does not let us build a complete solution for this task, we believe that tackling the individual components will help bring us towards an eventual solution.

In fact, existing works in the NLP and recommendation systems communities have already begun work that fits towards the completion of the first two tasks. Citation prediction (Nallapati et al., 2008) is a growing research area that has aimed both at predicting citation growth over time within a community and at individual paper citation patterns. This year, an automatic keyphrase extraction task from scientific articles was first fielded in SemEval-2, partially addressing Task 1[1]. Also, automatic survey generation (Mohammad et al., 2009) is becoming a growing field within the summarization community. However, to date, we have not yet seen any work that examines topic-biased summarization of multiple scientific articles. For these reasons, we focus on Task 3 here – *the creation of a related work section, given a structured input of the topics for summary.*. The remaining contributions of our paper

[1] http://semeval2.fbk.eu/semeval2.php

consists of work towards this goal:

• We conduct a study of the argumentative patterns used in related work sections, to describe the plausible summarization tactics for their creation in Section 3.

• We describe our approach to generate an extractive related work summary given an input topic hierarchy tree, using two separate strategies to differentiate between summarizing shallow internal nodes from deep detailed leaf nodes of the topic tree in Section 4.

2 Related Work

Fully automated related work summarization is significantly different from traditional summarization. While there are no existing studies on this specific problem, there are closely related endeavors. The iOPENER[2] project works towards automated creation of technical surveys, given a research topic (Mohammad et al., 2009). Standard generic multi-document summarization algorithms were applied to generate technical surveys. They showed that citation information was effective in the generation process. This was also validated earlier in (Nakov et al., 2004), which showed that the citing sentences in other papers can give a useful description of a target work.

Other studies focus mainly on single-document scientific article summarization. The pioneers of automated summarization (Luhn, 1958; Baxendale, 1958; Edmundson, 1969) had envisioned their approaches being used for the automatic creation of scientific summaries. They examined various features specific to scientific texts (*e.g.*, frequency-based, sentence position, or rhetorical clues features) which were proven effective for domain-specific summarization tasks.

Further, Mei and Zhai (2008) and Qazvinian and Radev (2008) utilized citation information in creating summaries for a single scientific article in computational linguistics domain. Also, Schwartz and Hearst (2006) also utilized the citation sentences to summarize the key concepts and entities in bioscience texts, and argued that citation sentences may contain informative contributions of a paper that complement its original abstract.

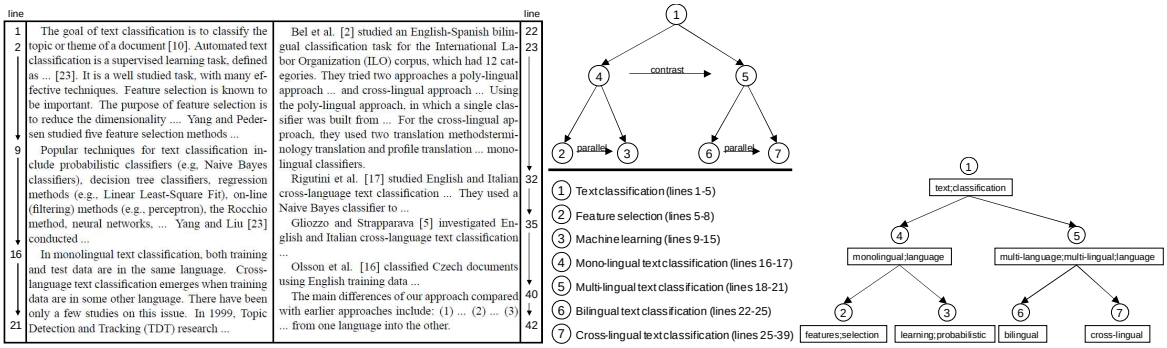
[2]http://clair.si.umich.edu/clair/iopener/

These works all center on the role of citations and their contexts in creating a summary, using citation information to rank content for extraction. However, they did not study the rhetorical structure of the intended summaries, targeting more on deriving useful content. For working along this vein, we turn to studies on the rhetorical structure of scientific articles. Perhaps the most relevant is work by (Teufel, 1999; Teufel and Moens, 2002) who defined and studied argumentative zoning of texts, especially ones in computational linguistics. While they studied the structure of an entire article, it is clear from their studies that a related work section would contain general background knowledge (BACKGROUND zone) as well as specific information credited to others (OTHER and BASIS zones). This vein of work has been followed by many, including Teufel et al. (2009; Angrosh et al. (2010).

3 Structure of Related Work Section

We first extend the work on rhetorical analysis, concentrating on related work sections. By studying examples in detail, we gain insight on how to approach related work summarization. We focus on a concrete related work example for illustration, an excerpt of which is shown in Figure 1a. Focusing on the argumentative progression of the text, we note the flow through different topics is hierarchical and can be represented as a topic tree as in Figure 1b.

This summary provides background knowledge for a paper on text classification, which is the root of the topic tree (node 1; lines 1–5). Two topics ("feature selection" and "machine learning") are then presented in parallel (nodes 2 & 3; lines 5–8 & 9–15), where specific details on relevant works are selected to describe two topics. These two topics are implicitly understood as subtopics of a more general topic, namely "mono-lingual text classification" (node 4; lines 16–17). The authors use the monolingual topic to contrast it with the subsequent subtopic "multi-lingual text classification" (node 5; lines 18–21). This topic is described by elaborating its details through two subtopics: "bilingual text classification" and "cross-lingual text classification" (nodes 6 & 7; lines 22–25 & 25–39) where again, various example works



(a) (b) (c)

Figure 1: a) A related work section extracted from (Wu and Oard, 2008); b) An associated topic hierarchy tree of a); c) An associated topic tree, annotated with key words/phrases.

are described and cited. The authors then conclude by contrasting their proposed approach with the introduced relevant approaches (lines 40–42).

This summary illustrates three important points. First, the topic tree is an essential input to the summarization process. The topic tree can be thought of as a high-level rhetorical structure for which a process then attaches content. While it is certainly non-trivial to build such a tree, modifications to hierarchical topic modeling (M. et al., 2004) or keyphrase extraction algorithms (Witten et al., 1999) we believe can be used to induce a suitable form. A resulting topic hierarchy from such a process would provide an associated set of key words or phrases that would describe the node, as shown in Figure 1c.

Second, while summaries can be structured in many ways, they can be viewed as moves along the topic hierarchy tree. In the example, nodes 2 and 3 are discussed before their parent, as the parent node (node 4) serves as a useful contrast to introduce its sibling (node 5). We find variants of depth-first traversal common, but breadth-first traversals of nodes with multiple descendants are more rare. They may be structured this way to ease the reader's burden on memory and attention. This is in line with other summary genres where information is ordered by high-level logical considerations that place macro level constraints (Barzilay et al., 2002).

Third, there is a clear distinction between sentences that describe a general topic and those that describe work in detail. Generic topics are often represented by background information, which is not tied to a particular prior work. These include definitions or descriptions of a topic's purpose. In contrast, detailed information forms the bulk of the summary and often describes key related work that is attributable to specific authors. Recently, Jaidka et al. (2010) also present the beginnings of a corpus study of related work sections, where they differentiate integrative and descriptive strategies in presenting discourse work. We see our differentiation between general and detailed topics as a natural parallel to their notion of integrative and descriptive strategies.

To introspect on these findings further, we created a related work data set (called **RWSData**[3]), which includes 20 articles from well-respected venues in NLP and IR, namely SIGIR, ACL, NAACL, EMNLP and COLING. We extracted the related work sections directly from those research articles as well as references the sections cited. References to books and Ph.D. theses were removed, as their verbosity would change the problem drastically (Mihalcea and Ceylan, 2007). Since we view each related work summary as a topic-biased summary originating from a topic hierarchy tree, annotation of such topical information for our data set is necessary. Each article's data consists of the reference related work summary, the collection of the input research articles

[3]To be made available at `http://wing.comp.nus.edu.sg/downloads/rwsdata`.

	SbL−RW	WbL−RW	No−RAs	SbL−RA	WbL−RA	TS	TD
average	17.9	522.4	10.9	2386.0	51739.6	3.3	1.8
stdev	7.9	216.5	5.6	1306.7	26682.3	1.7	0.6
min	6	179	2	348	8580	1	1
max	40	922	26	5549	112267	7	3

Table 1: **The demographics of RWSData**. No, RW, RA, SbL, WbL, TS, and TD are labeled as (N)umber (o)f, (R)elated (W)orks, (R)eferenced (A)rticles, (S)entence-(b)ased (L)ength of, (W)ord-(b)ased (L)ength of, (T)ree (S)ize, and (T)ree (D)epth, respectively.

that were referenced and a manually-constructed topic descriptions in a hierarchical fashion (topic tree). More details on the demographics of RWSData are shown in Table 1. RWSData summaries average 17.9 sentences, 522 words in length, citing an average of 10.9 articles. While hierarchical, the topic trees are simple, averaging 3.3 topic nodes in size and average depth of 1.8. Their simplicity furthers our claim that automated methods would be able to create such trees.

4 ReWoS: Paired General and Specific Summarization

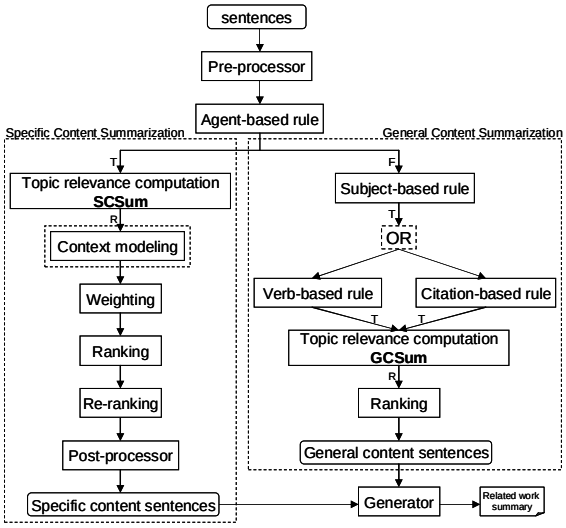


Figure 2: **The ReWoS architecture**. Decision edges labeled as **T**rue, **F**alse and **R**elevant.

Inspired by the above observations, we propose a novel strategy for related work summarization with respect to a given topic tree. Note that while the construction of the topic tree is central to the process, we consider this outside the scope of the current work (see §1); our investigation focuses on how such input could be utilized to construct a reasonable topic-biased related work summary.

We posit that sentences within a related work section come about by means of two separate processes – a process that gives general background information and another that describes specific author contributions. A key realization in our work is that these processes are easily mapped to the topic tree topologically: general content is described in tree-internal nodes, whereas leaf nodes contribute detailed specifics. In our approach, these two processes are independent, and combined to construct the final summary.

We have implemented our idea in **ReWoS** (**Re**lated **Wo**rk **S**ummarizer), whose general architecture is shown in Figure 2. ReWoS is a largely heuristic system, featuring both a **G**eneral **C**ontent **Sum**marization (**GCSum**) and a **S**pecific **C**ontent **Sum**marization (**SCSum**) modules, prefixed by preprocessing. A natural language template **generation** system fills out the end of the summary.

ReWoS first applies a set of preprocessing steps (shown in the top of Figure 2). Input sentences (*i.e.*, the set of sentences from each related/cited article) first removes sentences that are too short (< 7 tokens) or too long (> 80 tokens), ones that use future tense (possibly future work), and example and navigation sentences. This last category is filtered out by checking for the presence of a cue phrase among a lexical pattern database: *e.g.*, "in the section", "figure x shows", "for instance". Lowercasing and stemming are also performed.

We then direct sentences to either GCSum or SCSum based on whether it describes the author's own work or not, similar in spirit and execution to (Teufel et al., 2009). If sentence contains indicative pronouns or cue phrases (*e.g.*, "we", "this ap-

proach"), the sentence is deemed to describe own work and is directed to SCSum; otherwise the sentence is directed to the GCSum workflow.

4.1 (G)eneral (C)ontent (Sum)marization

GCSum extracts sentences containing useful background information on the topics of the internal node in focus. Since general content sentences do not specifically describe work done by the authors, we only take sentences that do not have the author-as-agent as input for GCSum.

We divide such general content sentences into two groups: indicative and informative. Informative sentences give detail on a specific aspect of the problem. They often give definitions, purpose or application of the topic (*"Text classification is a task that assigns a certain number of predefined labels for a given text."*). In contrast, indicative sentences are simpler, inserted to make the topic transition explicit and rhetorically sound (*"Many previous studies have studied monolingual text classification."*).

Indicative sentences can be easily generated by templates, as the primary information that is transmitted is the identity of the topic itself. Informative sentences, on the other hand, are better extracted from the source articles themselves, requiring a specific strategy. As informative sentences contain more content, our strategy with GCSum is to attempt to locate informative sentences to describe the internal nodes, failing which GCSum falls back to using predefined templates to generate an indicative placeholder.

To implement GCSum's informative extractor, we use a set of heuristics in a decision cascade to first filter inappropriate sentences (as shown on the RHS of Figure 2). Remaining candidates (if any) are then ranked by relevance and the top n are selected for the summary.

The heuristic cascade's purpose is to ensure sentences fit the syntactic structure of commonly-observed informative sentences. A useful sentence should discuss the topic directly, so GCSum first checks the subject of each candidate sentence, filtering sentences whose subject do not contain at least one topic keyword. We observed that background sentences often feature specific verbs or citations. GCSum thus also checks whether stock verb phrases (*i.e.*, "based on", "make use of" and 23 other patterns) are used as the main verb. Otherwise, GCSum checks for the presence of at least one citation – general sentences may list a set of citations as examples. If both the cue verb and citation checks fail, the sentence is dropped.

GCSum's topic relevance computation ranks remaining sentences based on keyword content. We state that the topic of an internal node is affected by its surrounding nodes – ancestor, descendants and siblings. Based on this idea, the score of a sentence is computed in a discriminative way using the following linear combination:

$$score_S \rightarrow score_S^{QA} + score_S^{Q} - score_S^{QR} \quad (1)$$

where $score_S$ is the final relevance score, and $score_S^{QA}$, $score_S^{Q}$, and $score_S^{QR}$ are the component scores of the sentence S with respect to the ancestor, current or other remaining nodes. We give positive credit to a sentence that contains keywords from an ancestor node, but penalize sentences with keywords from other topics (as such sentences would be better descriptors for those other topics). Component relevance scores are calculated using Term Frequency $\times$ Inverse Sentence Frequency (TF$\times$ISF) (Otterbacher et al., 2005):

$$\begin{aligned} score_S^{Q} &= \frac{rel(S,Q)}{\sum_{Q'} rel(S,Q')} \quad (2) \\ &= \frac{\sum_{w \in Q} \log(tf_w^S + 1) \times \log(tf_w^Q + 1) \times isf_w}{Norm} \end{aligned}$$

where $rel(S,Q)$ is the relevance of S with respect to topic Q, $Norm$ is a normalization factor of $rel(S,Q)$ over all input sentences, tf_w^S and tf_w^Q are the term frequencies of token w within S or sentences that discuss topic Q, respectively. isf_w is the inverse sentence frequency of w.

4.2 (S)pecific (C)ontent (Sum)marization

SCSum aims to extract sentences that contain detailed information about a specific author's work that is relevant to the input leaf node's topic from the set of sentences that exhibit the author-as-agent. SCSum starts by computing the topic relevance of each candidate sentence as shown in Equation (3). This process is identical to the step in GCSum, except that the term $score_S^{QR}$ in Equation (1) is replaced by $score_S^{QS}$, which is the relevance of S with respect to its sibling nodes. We

hypothesize that given a leaf node, sibling node topics may have an even more pronounced negative effect than other remaining nodes in the topic tree.

$$score_S \rightarrow score_S^{QA} + score_S^{Q} - score_S^{QS} \quad (3)$$

Context Modeling. We note that single sentences occasionally do not contain enough contexts to clearly express the idea mentioned in original articles. In fact, an agent-based sentence often introduces a concept but pertinent details are often described later. Extracting just the agent-based sentence may incompletely describe a concept and lead to false inferences. Consider the example in Figure 3. Here Sentences 0-5 are an contiguous extract of a source article being summarized, where Sentence 0 is an identified agent-based sentence. Sentence 6 shows a related work section sentence from a citing article that describes the original article. It is clear that the citing description is composed of information taken not only from the agent-based sentence but its context in the following sentences as well. This observation

0) To evaluate the quality of produced paraphrases, **we** picked at random 500 paraphrasing pairs from the lexical paraphrases produced by our algorithm.

---

1) These **pairs** were used as test data and also to evaluate whether **humans** agree on **paraphrasing judgments**.
2) The judges were given a page of guidelines, defining **paraphrase** as **approximate conceptual equivalence.**
3) The main dilemma in designing the evaluation is whether to include the context: should the **human judge** see only a paraphrase pair or should a pair of sentences containing these paraphrases also be given?
4) In a similar MT task evaluation of word-to-word translation context is usually included (Melamed, 2001).
5) Although paraphrasing is considered to be context dependent, there is no agreement on the extent.

---

6) (Barzilay and McKeown 2001) **evaluated** their **paraphrases** by asking **judges** whether **paraphrases** were "**approximately conceptually equivalent**".

Figure 3: A context modeling example.

motivates us to choose nearby sentences within a contextual window after the agent-based sentence to represent the topic. We set the contextual window to 5 and extract a maximum of 2 additional sentences. These additions are chosen based on their relevance scores to the topic, using Equation (3). Sentences with non-zero scores are then added as contexts of the anchor agent-based sentence, otherwise they are excluded. As a result, some topics may contain only a single sentence, but others may be described by additional contextual sentences.

Weighting. The score of a candidate content sentence is computed from topic relevance computation (SCSum) that includes contributions for keywords present in the current, ancestor and sibling nodes. We observe that the presence of one or more of current, ancestor and sibling nodes may affect the final score from the computation. Thus, to partially address this, we add a new weighting coefficient for the score computed from the topic relevance computation (SCSum) (Equation (3)) as follows:

$$score_S^* = w_S^{QA,Q,QS} \times score_S \quad (4)$$

where: $w_S^{QA,Q,QS}$ is a weighting coefficient that takes on differing values based on the presence of keywords in the sentence. Q, QA, and QS denote keywords from current, ancestor and sibling nodes. If the sentence contains keywords from other sibling nodes, we assign a penalty of 0.1. Otherwise, we assign a weight of 1.0, 0.5, or 0.25, based on whether keywords are present from both the ancestor node and current node, just the current node or just the ancestor node.

To build the final summary, ReWoS selects the top scoring sentence and iteratively adds the next most highly ranked sentence, until the n sentence budget is reached. We use SimRank (Li et al., 2008) to remove the next sentence to be added, if it is too similar to the sentences already in the summary.

4.3 Generation

ReWoS generates its summaries by using depth-first traversal to order the topic nodes, as in RWS-Data we observed this to be the most prevalent discourse pattern. It calls GCSum and SCSum to summarize individual nodes, distributing the total sentence budget equally among nodes.

ReWoS post-processes sentences to improve fluency where possible. We first replace agentive forms with a citation to the articles (*e.g.*, "we" → "(Wu and Oard, 2008)"). ReWoS also replaces found abbreviations with their corresponding long forms, by connecting abbreviation with their expansions by utilizing dependency relation output from the Stanford dependency parser.

System	ROUGE Recall Scores				Human Evaluation Scores			
	ROUGE-1	ROUGE-2	ROUGE-S4	ROUGE-SU4	Correctness	Novelty	Fluency	Usefulness
LEAD	0.501	0.096	0.116	0.181	3.027	2.764	3.082	2.745
MEAD	0.663	0.178	0.211	0.287	3.009	3.109	2.591	2.700
ReWoS−WCM	0.584	0.127	0.154	0.227	3.618	3.391	3.391	3.609
ReWoS−CM	0.698	0.183	0.218	0.298	3.691	3.618	2.955	3.573

Table 2: Evaluation results for ReWoS variants and baselines.

5 Evaluation

We wish to assess the quality of ReWoS, comparing to state-of-the-art generic summarization systems. We first detail our baseline systems used for performance comparison, and defined evaluation measures specific to related work summary evaluation. In our evaluation, we use our manually-compiled RWSData data set.

We benchmark ReWoS against two baseline systems: LEAD and MEAD. The LEAD baseline represents each of the cited article with an equal number of sentences. The first n sentences are drawn from the article, meaning that the title and abstract are usually extracted. The order of the article leads used in the resulting summary was determined by the order of articles to be processed. MEAD is a well-documented baseline extractive multi-document summarizer, developed in (Radev et al., 2004). MEAD offers a set of different features that can be parameterized to create resulting summaries. We conducted an internal tuning of MEAD to maximize its performance on the RWSData. The optimal configuation uses just two tuned features of *centroid* and *cosine similarity*. Note that the MEAD baseline does use the topic tree keywords in computing cosine similarity score. Our ReWoS system is the only system that leverages the topic tree *structure* which is central to our approach. In our experiments, we used MEAD toolkit[4] to produce the summaries for LEAD and MEAD baseline systems.

Automatic evaluation was performed with ROUGE (Lin, 2004), a widely used and recognized automated summarization evaluation method. We employed a number of ROUGE variants, which have been proven to correlate with human judgments in multi-document summarization (Lin, 2004). However, given the small size of our summarization dataset, we can only draw notional evidence from such an evaluation; it is not possible to find statistically significant conclusion from our evaluation.

To partially address this, we also conducted a human evaluation to assess more fine-grained qualities of our system. We asked 11 human judges to follow an evaluation guideline that we prepared, to evaluate the summary quality, consisting of the following evaluation measures:

• **Correctness:** Is the summary content actually relevant to the hierarchical topics given?

• **Novelty**: Does the summary introduce novel information that is significant in comparison with the human created summary?

• **Fluency**: Does the summary's exposition flow well, in terms of syntax as well as discourse?

• **Usefulness**: Is the summary useful in supporting the researchers to quickly grasp the related works given hierarchical topics?

Each judge was asked to grade the four summaries according to the measures on a 5-point scale of 1 (very poor) to 5 (very good). Summaries 1 and 2 come from LEAD-based and MEAD systems, respectively. Summaries 3 and 4 come from our proposed ReWoS systems, without (ReWoS−WCM) and with (ReWoS−CM) the context modeling in SCSum. All summarizers were set to yield a summary with the same length (1% of the original relevant articles, measured in sentences). Due to limited time, only 10 out of 20 evaluation sets were assessed by the evaluators. Each set was graded at least 3 times by 3 different evaluators; evaluators did not know the identities of the systems, which were randomized for each set examined.

6 Results

ROUGE results are summarized in Table 2. Surprisingly, the MEAD baseline system outperforms both LEAD baseline and ReWoS–WCM (without context modeling). Only ReWoS–CM (with

[4] `http://www.summarization.com/mead/`

context modeling) is significantly better than others, in terms of all ROUGE variants. Here are some possible reasons to explain this. First, ROUGE evaluation seems to work unreasonably when dealing with verbose summaries, often produced by MEAD. Second, related work summaries are multi-topic summaries of multi-article references. This may cause miscalculation from overlapping n-grams that occur across multiple topics or references.

Since automatic evaluation with ROUGE does not allow much introspection, we turn to our human evaluation. Results are also summarized in Table 2. They show that both ReWoS–WCM and ReWoS–CM perform significantly better than baselines in terms of correctness, novelty, and usefulness. This is because our system utilized features developed specifically for related work summarization. Also, our proposed systems compare favorably with LEAD, showing that necessary information is not only located in titles or abstracts, but also in relevant portions of the research article body.

ReWoS–CM (with context modeling) performed equivalent to ReWoS–WCM (without it) in terms of correctness and usefulness. For novelty, ReWoS–CM is better than ReWoS–WCM. It proved that the proposed component of context moding is useful in providing new information that is necessary for the related work summaries. For fluency, only ReWoS–CM is better than baseline systems. This is a negative result, but is not surprising because the summaries from the ReWoS–CM which uses context modeling seems to be longer than others. It makes the summaries quite hard to digest; some evaluators told us that they preferred the shorter summaries. A future extension is that using information fusion techniques to fuse the contextual sentences with its anchor agentive sentence.

A detailed error analysis of the results revealed that there are three main types of errors produced by our systems. The first issue is in calculating topic relevance. In the context of related work summarization, our heuristics-based strategies for sentence extraction cannot capture fully this issue. Some sentences that have high relevant scores to topics are not actually semantically relevant to the topics. The second issue of anaphoric expression is more addressable. Some extracted sentences still contain anaphoric expression (*e.g.*, "they", "these", "such", ...), making final generated summaries incoherent. The third issue is paraphrasing, where substituted paraphrases replace the original words and phrases in the source articles.

7 Conclusion and Future Work

According to the best of our knowledge, automated related work summarization has not been studied before. In this paper, we have taken the initial steps towards solving this problem, by dividing the task into general and specific summarization processes. Our initial results show an improvement over generic multi-document summarization baselines in human evaluation. However, our work shows that there is much room for additional improvement, for which we have outlined a few challenges.

A shortcoming of our current work is that we assume that a topic hierarchy tree is given as input. We feel that this is an acceptable limitation because we feel existing techniques will be able to create such input, and that the topic trees used in this study were quite simple. We plan to validate this by generating these topic trees automatically in our future work.

Exploring related work summarization comes at a timely moment, as scholars now have access to a preponderous amount of scholarly literature. Automated assistance in interpreting and organizing scholarly work will help build future applications for integration with digital libraries and reference management tools.

References

Angrosh, M. A., Stephen Cranefield, and Nigel Stanger. 2010. Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries. In *JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries*, pages 293–302. ACM.

Barzilay, Regina, Noemie Elhadad, and Kathleen R. McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summariza-

tion. In *Journal of Artificial Intelligence Research*, volume 17, pages 35–55.

Baxendale, P. B. 1958. Machine-made index for technical literature - an experiment. *IBM Journal of Research Development*, 2(4):354–361.

Edmundson, H. P. 1969. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285.

Jaidka, Kokil, Christopher S. G. Khoo, and Jin-Cheon Na. 2010. Imitating human literature review writing: An approach to multi-document summarization. In *ICADL*, pages 116–119.

Li, Wenjie, Furu Wei, Qin Lu, and Yanxiang He. 2008. PNR2: Ranking sentences with positive and negative reinforcement for query-oriented update summarization. In *Proceedings of 22nd International Conference on Computational Linguistics*, pages 489–496, Manchester, UK, August.

Lin, Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop Text Summarization Branches Out*, pages 74–81, Spain, July.

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159–165.

M., Blei D., Griffiths T. L., Jordan M. I., and Tenenbaum J. B. 2004. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems (NIPS)*.

Mei, Qiaozhu and ChengXiang Zhai. 2008. Generating impact-based summaries for scientific literature. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 816–824, Columbus, Ohio, June.

Mihalcea, Rada and Hakan Ceylan. 2007. Explorations in automatic book summarization. In *Proceedings of Empirical Methods in Natural Language Processing - Conference on Natural Language Learning (EMNLP-CoNLL)*, pages 380–389, Prague, Czech Republic, June.

Mohammad, S., B. Dorr, M. Egan, A. Hassan, P. Muthukrishan, V. Qazvinian, D. Radev, and D. Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies - North American Association for Computational Linguistics (HLT-NAACL)*, pages 584–592, Boulder, Colorado, June.

Nakov, Preslav I., Ariel S. Schwartz, and Marti A. Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Workshop on Search and Discovery in Bioinformatics*.

Nallapati, R. M., A. Ahmed, E. P. Xing, and W. W. Cohen. 2008. Joint latent topic models for text and citations. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery in Data and Data Mining*, pages 542–550.

Otterbacher, Jahna, Güneş Erkan, and Dragomir R. Radev. 2005. Using random walks for question-focused sentence retrieval. In *Proceedings of Human Language Technologies - Empirical Methods in Natural Language Processing (HLT-EMNLP '05)*, pages 915–922. ACL.

Qazvinian, Vahed and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of International Conference on Computational Linguistics (COLING)*, pages 689–696, Manchester, UK, August.

Radev, Dragomir R., Hongyan Jing, Malgorzata Sty, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management (IPM)*, 40(6):919–938.

Schwartz, Ariel S. and Marti Hearst. 2006. Summarizing key concepts using citation sentences. In *Proceedings of Natural language processing of biology text (BioNLP '06)*, pages 134–135. ACL.

Teufel, Simone and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Teufel, Simone, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore, August. Association for Computational Linguistics.

Teufel, Simone. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh.

Witten, Ian H., Gordon Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Kea: Practical automatic keyphrase extraction. In *Proceedings of Digital Libraries 99 (DL'99)*, pages 254–255. ACM Press.

Wu, Yejun and Douglas W. Oard. 2008. Bilingual topic aspect classification with a few training examples. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 203–210, New York, NY, USA. ACM.