Neural Related Work Summarization with a Joint Context-driven Attention Mechanism

Yongzhen Wang[1][*], **Xiaozhong Liu**[2,3], **Zheng Gao**[2]

[1]School of Maritime Economics and Management, Dalian Maritime University, Dalian, China
[2]School of Informatics, Computing and Engineering, Indiana University Bloomington,
Bloomington, IN, USA
[3]Alibaba Group, Hangzhou, China

[*]kuadmu@163.com   liu237@indiana.edu   gao27@umail.iu.edu

Abstract

Conventional solutions to automatic related work summarization rely heavily on human-engineered features. In this paper, we develop a neural data-driven summarizer by leveraging the seq2seq paradigm, in which a joint context-driven attention mechanism is proposed to measure the contextual relevance within full texts and a heterogeneous bibliography graph simultaneously. Our motivation is to maintain the topic coherency between a related work section and its target document, where both the textual and graphic contexts play a big role in characterizing the relationship among scientific publications accurately. Experimental results on a large dataset show that our approach achieves a considerable improvement over a typical seq2seq summarizer and five classical summarization baselines.

1 Introduction

In scientific fields, scholars need to contextualize their contribution to help readers acquire an understanding of their research papers. For this purpose, the related work section of an article serves as a pivot to connect prior domain knowledge, in which the innovation and superiority of current work are displayed by a comparison with previous studies. While citation prediction can assist in drafting a reference collection (Nallapati et al., 2008), consuming all these papers is still a laborious job, where authors must read every source document carefully and locate the most relevant content cautiously.

As a solution in saving authors' efforts, automatic related work summarization is essentially a topic-biased multi-document problem (Cong and Kan, 2010), which relies heavily on human-engineered features to retrieve snippets from the references. Most recently, neural networks enable a data-driven architecture *sequence-to-sequence* (seq2seq) for natural language generation (Bahdanau et al., 2014, 2016), where an encoder reads a sequence of words/sentences into a context vector, from which a decoder yields a sequence of specific outputs. Nonetheless, compared to scenarios like machine translation with an end-to-end nature, aligning a related work section to its source documents is far more challenging.

To address the summarization alignment, former studies try to apply an attention mechanism to measure the saliency/novelty of each candidate word/sentence (Tan et al., 2017), with the aim of locating the most representative content to retain primary coverage. However, toward summarizing a related work section, authors should be more creative when organizing text streams from the reference collection, where the selected content ought to highlight the topic bias of current work, rather than retell each reference in a compressed but balanced fashion. This motivates us to introduce the contextual relevance and characterize the relationship among scientific publications accurately.

Generally speaking, for a pair of documents, a larger lexical overlap often implies a higher similarity in their research backgrounds. Yet such a hypothesis is not always true when sampling content from multiple relevant topics. Take "DSSM"[1] as an example, from viewpoint of the abstract similarity, those references investigating "Information Retrieval", "Latent Semantic Model" or "Click-through Data Mining" could be of more importance in correlation and should be greatly sampled for the related work section. But in reality, this article spends a bit larger chunk of texts (about 58%) to elaborate "Deep Learning" during the literature review, which is quite difficult for machines to grasp the contextual relevance therein. In addi-

---

[1]Learning deep structured semantic models for web search using clickthrough data (Huang et al., 2013)

tion, other situations like emerging new concepts also suffer from the terminology variation or paraphrasing in varying degrees.

In this study, we utilize a heterogeneous bibliography graph to embody the relationship within a scalable scholarly database. Over the recent past, there is a surge of interest in exploiting diverse relations to analyze bibliometrics, ranging from literature recommendation (Yu et al., 2015) to topic evolvement (Jensen et al., 2016). In a graphical sense, interconnected papers transfer the credit among each other directly/indirectly through various patterns, such as paper citation, author collaboration, keyword association and releasing on series of venues, which constitutes the graphic context for outlining concerned topics. Unfortunately, a variety of edge types may pollute the information inquiry, where a slice of edges are not so important as the others on sampling content. Meanwhile, most existing solutions in mining heterogeneous graphs depend on the human supervision, e.g., *hyperedge* (Bu et al., 2010) and *metapath* (Swami et al., 2017). This is usually not easy to access due to the complexity of graph schemas.

Our contribution is threefold: First, we explore the *edge-type usefulness distribution* (EUD) on a heterogeneous bibliography graph, which enables the relationship discovery (between any pair of papers) for sampling the interested information. Second, we develop a novel seq2seq summarizer for the automatic related work summarization, where a joint context-driven attention mechanism is proposed to measure the contextual relevance within both textual and graphic contexts. Third, we conduct experiments on 8,080 papers with native related work sections, and experimental results show that our approach outperforms a typical seq2seq summarizer and five classical summarization baselines significantly.

2 Related Work

This study touches on several strands of research within automatic related work summarization and seq2seq summarizer as follows.

The idea of creating a related work section automatically is pioneered by Cong and Kan (2010) who design two rule-based strategies to extract sentences for general and detailed topics respectively. Subsequently, Hu and Wan (2014) exploit *probabilistic latent semantic indexing* to split candidate texts into different topic-biased parts, then

Authors	Number of papers
Cong and Kan (2010)	20
Hu and Wan (2014)	1,050
Widyantoro and Amin (2014)	50
Chen and Hai (2016)	3

Table 1: Data scales of previous studies on automatic related work summarization.

apply several regression models to learn the importance of each sentence. Similarly, Widyantoro and Amin (2014) transform the summarization problem into classifying rhetorical categories of sentences, where each sentence is represented as a feature vector containing word frequency, sentence length and etc. Most recently, Chen and Hai (2016) construct a graph of representative keywords, in which a *minimum steiner tree* is figured out to guide the summarization as finding the least number of sentences to cover the discriminated nodes. In general, compared to traditional summaries, the automatic related work summarization receives less concerns over the past. However, these existing solutions cannot work without manual intervention, which limits the application scale to an extremely small size (see Table 1).

The earliest seq2seq summarizer stems from Rush et al. (2015) which utilizes a feed-forward network for compressing sentences, and later is expanded by Chopra et al. (2016) with a *recurrent neural network* (RNN). On this basis, Nallapati et al. (2016a,c) and Chen et al. (2016) both present a set of RNN-based models to address various aspects of abstractive summarization. Typically, Cheng and Lapata (2016) propose a general seq2seq summarizer, where an encoder learns the representation of documents while a decoder generates each word/sentence using an attention mechanism. With further research, Nallapati et al. (2016b) extend the sentence compression by trying a hierarchical attention architecture and a limited vocabulary during the decoding phase. Next, Narayan et al. (2017) leverage the side information as an attention cue to locate focus regions for summaries. Recently, inspired by *PageRank*, Tan et al. (2017) introduce a graph-based attention mechanism to tackle the saliency problem. Nonetheless, these methods all discuss the single-document scenario, which is far from the nature of automatic related work summarization.

In this study, derived from the general seq2seq summarizer of Cheng and Lapata (2016), we propose a joint context-driven attention mechanism to

measure the contextual relevance within full texts and a heterogeneous bibliography graph simultaneously. To our best knowledge, we make the first attempt to develop a neural data-driven solution for the automatic related work summarization, and the practice of using the joint context as an attention cue is also less explored to date. Besides, this study is launched on a dataset with up to 8,080 papers, which is much greater than previous studies and makes our results more convincing.

Since text summarization via word-by-word generation is not mature at present (Cheng and Lapata, 2016; Nallapati et al., 2016b; Tan et al., 2017), we adopt the extractive sentential fashion for our summarizer, where a related work section is created by extracting and linking sentences from a reference collection. Meanwhile, this study follows the mode of Cong and Kan (2010) who assume that the collection is given as part of the input, and do not consider the citation sentences of each reference.

3 Methodology

3.1 Problem Formulation

To adapt the seq2seq paradigm, we formulate the automatic related work summarization into a sequential text generation problem as follows.

Given an unedited paper $\mathtt{t}$ (target document) and its n-size reference collection $\mathtt{R_t} = \{\mathtt{r}^{\mathtt{t}}_{1:n}\}$, we draw up a related work section for $\mathtt{t}$ by selecting sentences from $\mathtt{R_t}$. To be specific, each reference (source document) will be traversed one time sequentially, and without loss of generality, in the descending order of their significance to $\mathtt{t}$. Consequently, all sentences to be selected are concatenated into an m-length sequence $\mathtt{S_t} = \{\mathtt{s}^{\mathtt{t}}_{1:m}\}$ to feed the summarizer. For each candidate sentence $\mathtt{s}^{\mathtt{t}}_j$, once being visited, a label $\mathtt{y}^{\mathtt{t}}_j \in \{0, 1\}$ will be determined synchronously based on whether or not this sentence should be covered into the output. Our objective is to maximize the log-likelihood probability of observed labels $\mathtt{Y_t} = \{\mathtt{y}^{\mathtt{t}}_{1:m}\}$ under $\mathtt{R_t}$, $\mathtt{S_t}$ and summarizer parameters θ, as shown below.

$$\max \sum_{j=1}^{m} \log \Pr(\mathtt{y}^{\mathtt{t}}_j \mid \mathtt{R_t}; \mathtt{S_t}; \theta) \qquad (1)$$
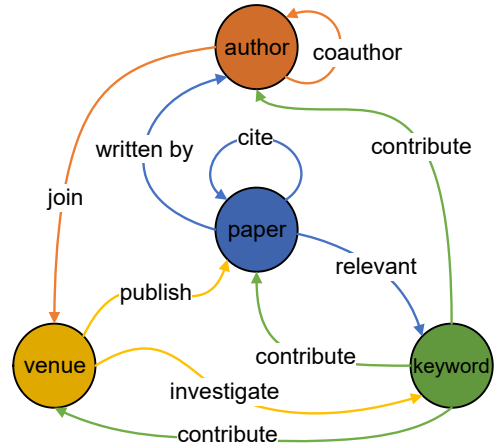

Figure 1: Heterogeneous bibliography graph.

3.2 Random Walk on Heterogeneous Bibliography Graph

Prior works have illustrated that one of the most promising channels for information recommendation is the community network (Guo and Liu, 2015). In this study, we verify this hypothesis toward the content sampling of scientific summarization, by investigating heterogeneous relations among different kinds of objects such as papers, authors, keywords and venues.

For measuring the relationship among scientific publications, we introduce a directed graph $\mathtt{G} = (\mathtt{V}, \mathtt{E})$ to contain various bibliographical connections, as shown in Figure 1, which involves four objects and ten edge types in total. Each edge $\mathtt{e}_{j,i} \in \mathtt{E}$ is assigned a value $\frac{\pi(\mathtt{e}_{j,i})}{z} \in [0, 1]$ to indicate the transition probability between two nodes $\mathtt{v}_j, \mathtt{v}_i \in \mathtt{V}$, where $\pi(\mathtt{e}_{j,i}) \in \mathbb{R}$ returns the unknown edge-type usefulness of $\mathtt{e}_{j,i}$, and $z \in \mathbb{R}$ is a normalizing weight. For most of edge types, we model the weight as one divided by the number of outgoing links of the same kind. But regarding the "contribution" category, the weight modeling is accomplished by *PageRank with Priors* (White and Smyth, 2003). Note that different edge types usually take very uneven importance in one particular task (Yu et al., 2015), and it is quite difficult to enable the classical heterogeneous graph mining without expert defined paths for *random walk* (Bu et al., 2010; Swami et al., 2017).

In this study, we propose an unsupervised approach to capture the connectivity diversity, by introducing an optimal EUD for navigating *random walkers* on the heterogeneous bibliography graph. Given a target document $\mathtt{t}$, the optimized useful-

ness assignment can help those walkers lock a top-n recommendation $\bar{\mathtt{R}}_\mathtt{t}$ to best match the reference collection $\mathtt{R}_\mathtt{t}$, as shown in Eq. 2. On this basis, a well-performing algorithm *node2vec* (Grover and Leskovec, 2016) is adopted to conduct an unsupervised *random walk* to vectorize every node $\forall \mathtt{v}_* \in \mathtt{V}$ into a d-dimensional embedding $\varphi(\mathtt{v}_*) \in \mathbb{R}^d$ so that any edge $\forall \mathtt{e}_* \in \mathtt{E}$ can be calculated therefrom. Specifically, we employ *evolutionary algorithm* (EA) to tune the EUD, which enjoys advantages over conventional gradient methods in both convergence speed and accuracy.

$$\arg\max \sum_\mathtt{t} \sum_{j=1}^n \log \Pr(\mathtt{r}_j^\mathtt{t} \in \bar{\mathtt{R}}_\mathtt{t} \mid \text{EUD}) \quad (2)$$

EA Setup We use an array of real numbers $\mathtt{x}_{1:10}$ to code an individual in the population, where $\mathtt{x}_j \in [0,1]$ denotes the usefulness of j-th edge type. Given an EUD, *PageRank* (Page, 1998) runs on graph to infer the relative importance of each node for each target document, and a fitness function is applied to judge how well this EUD satisfies locating the ground truth references as Eq. 3, in which if $\mathtt{r}_j^\mathtt{t}$ belongs to $\bar{\mathtt{R}}_\mathtt{t}$, then $\alpha(\mathtt{r}_j^\mathtt{t}, \bar{\mathtt{R}}_\mathtt{t}) \in \mathbb{N}$ returns the ranking of $\mathtt{r}_j^\mathtt{t}$ within $\bar{\mathtt{R}}_\mathtt{t}$, and otherwise a big penalty coefficient to prevent irrelevant references to be recommended. Like most other optimizations, this procedure starts with a randomly generated population.

$$\max \frac{1}{\sum_\mathtt{t} \sum_{j=1}^n \left| j - \alpha(\mathtt{r}_j^\mathtt{t}, \bar{\mathtt{R}}_\mathtt{t}) \right|} \quad (3)$$

EA Operator We choose the operator from *differential evolution* (Das and Suganthan, 2011) to generate offsprings for each individual. The basic idea is to utilize the difference between different individuals to disturb each trial object. First, three distinct individuals $\mathtt{x}_{1:10}^{r1}, \mathtt{x}_{1:10}^{r2}, \mathtt{x}_{1:10}^{r3}$ are sampled randomly from current population to create a variant $\mathtt{x}_{1:10}^{\text{var}}$, as shown in Eq. 4, where $f \in \mathbb{R}$ indicates the scaling factor. Next, $\mathtt{x}_{1:10}^{\text{var}}$ is crossed with a trial object $\mathtt{x}_{1:10}^{\text{tri}}$ to build a hybrid one $\mathtt{x}_{1:10}^{\text{hyb}}$ as Eq. 5, in which $c \in [0,1]$ denotes the crossover factor and $u \in [0,1]$ represents an uniform random number. At last, the fitnesses of $\mathtt{x}_{1:10}^{\text{tri}}$ and $\mathtt{x}_{1:10}^{\text{hyb}}$ are compared, and the better one will be saved as the offspring into a new round of evolution.

$$\mathtt{x}_j^{\text{var}} = \mathtt{x}_j^{r1} + f \times (\mathtt{x}_j^{r2} - \mathtt{x}_j^{r3}) \quad (4)$$

$$\mathtt{x}_j^{\text{hyb}} = \begin{cases} \mathtt{x}_j^{\text{var}}, & \text{if} \quad u \le c \\ \mathtt{x}_j^{\text{tri}}, & \text{otherwise} \end{cases} \quad (5)$$

3.3 Neural Extractive Summarization

As Figure 2 shows, we model our seq2seq summarizer with a hierarchical encoder and an attention-based decoder, as described below.

Hierarchical Encoder Our encoder consists of two major layers, namely a *convolutional neural network* (CNN) and a *long-short-term memory* (LSTM)-based RNN. Specifically, the CNN deals with word-level texts to derive sentence-level meanings, which are then taken as inputs to the RNN for handling longer-range dependency within lager units like a paragraph and even a whole paper. This conforms to the nature of document that is composed from words, sentences and higher levels of abstraction (Narayan et al., 2017).

Consider a sentence of p words $\mathtt{s}_j^\mathtt{t} = \{\mathtt{w}_{j,1:p}^\mathtt{t}\}$, where each word $\mathtt{w}_{j,i}^\mathtt{t}$ can be represented by a d-dimensional embedding $\phi(\mathtt{w}_{j,i}^\mathtt{t}) \in \mathbb{R}^d$. Previous studies have illustrated the strength of CNN in presenting sentences, because of its capability to learn compressed expressions and address sentences with variable lengths (Kim, 2014). First, a convolution kernel $k \in \mathbb{R}^{d \times q \times d}$ is applied to each possible window of q words to construct a list of feature maps as:

$$\mathtt{g}_{j,i}^\mathtt{t} = \tanh\left(k \times \phi(\mathtt{w}_{j,i:i+q-1}^\mathtt{t}) + b\right) \quad (6)$$

where $b \in \mathbb{R}^d$ denotes the bias term. Next, *max-over-time pooling* (Collobert et al., 2011) is performed on all generated features to obtain the sentence embedding as:

$$\phi(\mathtt{s}_j^\mathtt{t}) = \max_{1 \le i \le d}\left(\mathtt{g}_{j,1:p-q+1}^\mathtt{t}[i,:]\right) \quad (7)$$

where $[i,:]$ denotes the i-th row of matrix. Given a sequence of sentences $\mathtt{S}_\mathtt{t} = \{\mathtt{s}_{1:m}^\mathtt{t}\}$, we then take the RNN to yield an equal-length array of hidden states, in which LSTM has proved to alleviate the vanishing gradient problem when training long sequences (Hochreiter and Schmidhuber, 1997). Each hidden state can be viewed as a local representation with focusing on current and former sentences together, which is updated as: $\mathtt{h}_j^\mathtt{t} = \text{LSTM}\left(\phi(\mathtt{s}_j^\mathtt{t}), \mathtt{h}_{j-1}^\mathtt{t}\right) \in \mathbb{R}^d$.

In practice, we use multiple kernels with various widths to produce a group of embeddings for each
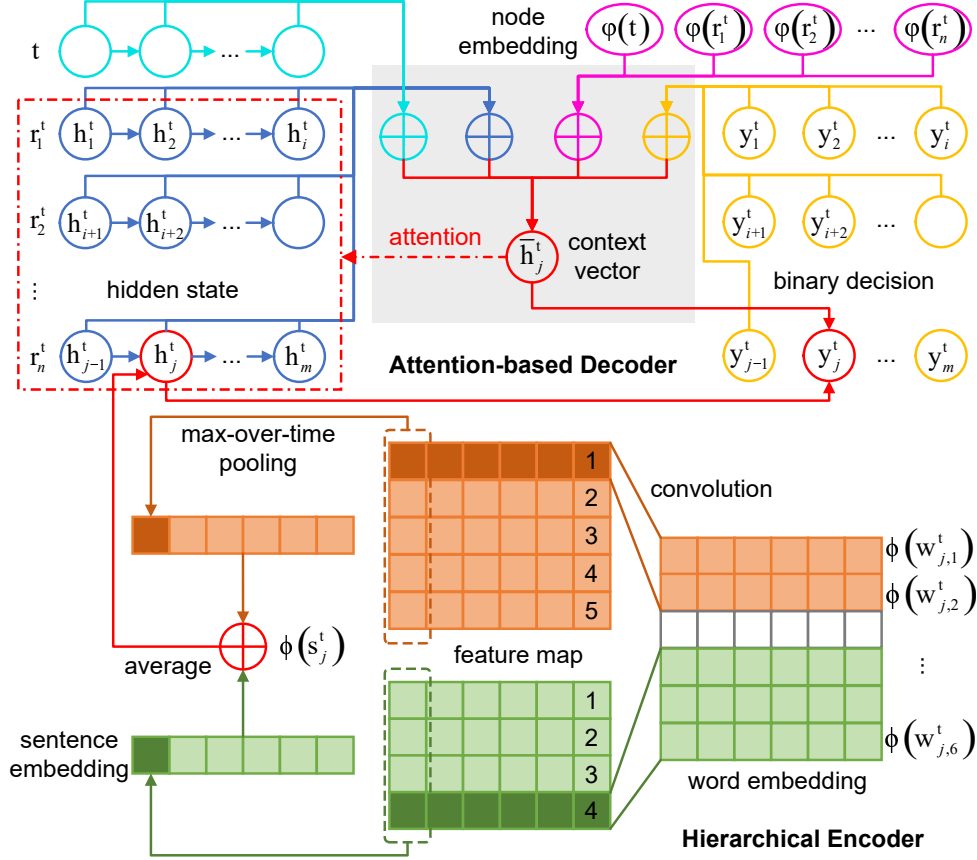
Figure 2: Framework of our seq2seq summarizer.

sentence, and average them to capture the information inside different n-grams. As Figure 2 (bottom) shows, the sentence $\mathbf{s}_j^{\mathtt{t}}$ involves six words, and two kernels of widths two (orange) and three (green) abstract a set of five and four feature maps respectively. Meanwhile, since *rhetorical structure theory* (Mann and Thompson, 2009) points out that association must exist in any two parts of coherent texts, RNN is only applicable to manage the sentence relation within a single document, because we cannot expect the dependency between two sections from different references.

Attention-based Decoder Our decoder labels each sentence $\mathbf{s}_j^{\mathtt{t}}$ as 0/1 sequentially, according to whether it is salient or novel enough, plus if relevant to the target document $\mathtt{t}$ or not. As shown in Figure 2 (top), the binary decision $\mathtt{y}_j^{\mathtt{t}}$ is made by both the hidden state $\mathtt{h}_j^{\mathtt{t}}$ and the context vector $\bar{\mathtt{h}}_j^{\mathtt{t}}$ from an attention mechanism (grey background). In particular, this attention (red dash line) is acted as an intermediate stage to determine which sentences to highlight so as to provide the contextual information for current decision (Bahdanau et al., 2014). Given $\mathtt{H}_{\mathtt{t}} = \{\mathtt{h}_{1:m}^{\mathtt{t}}\}$, this decoder returns

the probability of $\mathtt{y}_j^{\mathtt{t}} = 1$ as below:

$$\mathrm{Pr}(\mathtt{y}_j^{\mathtt{t}} = 1 \mid \mathtt{R}_{\mathtt{t}}; \mathtt{S}_{\mathtt{t}}; \theta) = \mathrm{sigmoid}\big(\delta(\mathtt{h}_j^{\mathtt{t}}, \bar{\mathtt{h}}_j^{\mathtt{t}})\big) \quad (8)$$

$$\bar{\mathtt{h}}_j^{\mathtt{t}} = \sum_{i=1}^{m} \mathtt{a}_{j,i} \mathtt{h}_i^{\mathtt{t}} \quad (9)$$

where $\delta(\mathtt{h}_j^{\mathtt{t}}, \bar{\mathtt{h}}_j^{\mathtt{t}}) \in \mathbb{R}$ denotes a fully connected layer with as input the concatenation of $\mathtt{h}_j^{\mathtt{t}}$ and $\bar{\mathtt{h}}_j^{\mathtt{t}}$, and $\mathtt{a}_{j,i} \in [0, 1]$ is the attention weight indicating how much the supporting sentence $\mathbf{s}_i^{\mathtt{t}}$ contributes to extracting the candidate one $\mathbf{s}_j^{\mathtt{t}}$.

Apart from saliency and novelty two traditional attention factors (Chen et al., 2016; Tan et al., 2017), we focus on the contextual relevance within both textual and graphic contexts to distinguish the relationship from near to far, as shown in Eq. 10 and Eq. 11. To be specific: 1) $\mathtt{h}_j^{\mathtt{t}\mathrm{T}} W_{\mathtt{s}} \mathtt{h}_i^{\mathtt{t}}$ represents the saliency of $\mathbf{s}_i^{\mathtt{t}}$ to $\mathbf{s}_j^{\mathtt{t}}$; 2) $-\mathtt{d}_j^{\mathtt{t}\mathrm{T}} W_{\mathtt{n}} \mathtt{h}_i^{\mathtt{t}}$ indicates the novelty of $\mathbf{s}_i^{\mathtt{t}}$ to the dynamic output $\mathtt{d}_j^{\mathtt{t}}$; 3) $\phi(\mathtt{t})^{\mathrm{T}} W_{\mathtt{t}} \mathtt{h}_i^{\mathtt{t}}$ denotes the relevance of $\mathbf{s}_i^{\mathtt{t}}$ to $\mathtt{t}$ from the textual context; 4) $\varphi(\mathtt{t})^{\mathrm{T}} W_{\mathtt{g}} \varphi(\mathtt{h}_i^{\mathtt{t}})$ refers to the relevance from the graphic context. More

concretely, $W_* \in \mathbb{R}^d$ characterizes the learnable matrix, $\phi(\mathtt{t})$ returns the average of hidden states from $\mathtt{t}$, $\varphi(\mathtt{t})$ and $\varphi(\mathtt{h}_i^{\mathtt{t}})$ return the node embeddings of both $\mathtt{t}$ and the source document that $\mathtt{h}_i^{\mathtt{t}}$ belongs to respectively. Note that $\phi(\cdot)$ and $\varphi(\cdot)$ represent two distinct embedding spaces, where the former reflects the lexical collocations of corpus, and the latter embodies the connectivity patterns of associated graph.

$$
\begin{aligned}
\mathtt{a}_{j,i} = \quad & \mathtt{h}_j^{\mathtt{t}\mathrm{T}} W_{\mathtt{s}} \mathtt{h}_i^{\mathtt{t}} && \text{\# saliency} \\
& -\mathtt{d}_j^{\mathtt{t}\mathrm{T}} W_{\mathtt{n}} \mathtt{h}_i^{\mathtt{t}} && \text{\# novelty} \\
& +\phi(\mathtt{t})^{\mathrm{T}} W_{\mathtt{t}} \mathtt{h}_i^{\mathtt{t}} && \text{\# relevance}_1 \\
& +\varphi(\mathtt{t})^{\mathrm{T}} W_{\mathtt{g}} \varphi(\mathtt{h}_i^{\mathtt{t}}) && \text{\# relevance}_2
\end{aligned}
\tag{10}
$$

$$
\mathtt{d}_j^{\mathtt{t}} = \sum_{i=1}^{j-1} \Pr(\mathtt{y}_j^{\mathtt{t}} = 1 \mid \mathtt{R}_{\mathtt{t}}; \mathtt{S}_{\mathtt{t}}; \theta) \times \mathtt{h}_i^{\mathtt{t}} \tag{11}
$$

The basic idea behind our attention mechanism is as follows: if a supporting sentence more resembles a candidate one, or overlaps less with the dynamic output, or is more relevant to the target document, then it can provide more contextual information to facilitate current decision on being extracted or not, thereby taking a higher weight in the generated context vector. This innovative attention will guide our goal related work section to maximize the representativeness of selected sentences (saliency & novelty), while minimizing the semantic distance to the target document (relevance). This is consistent with the way that scholars consume a reference collection, with the min-max objective in their minds.

4 Experiment

4.1 Experimental Setup

This section presents the experimental setup for assessing our approach, including 1) dataset used for training and testing, 2) implementation details, 3) contrast methods and evaluation metrics.

Dataset We conduct experiments on a dataset[2] created from the ACM digital library, where metadata and full texts are derived from PDF files. To be detailed, this dataset includes 371,891 papers,

---

779,810 authors, 9,204 keywords and 807 venues in total. Note that we ignore the keyword with frequency below a certain threshold, and adopt *greedy matching* of Guo et al. (2013) to generate pseudo keywords for papers lacking topic descriptions. For each target document, the references are traversed by the descending order of the cited number in related work section (primary) and in full paper (secondary) successively. We first apply a series of pre-processings such as lowercasing and stemming to standardize candidate sentences, then remove those which are too short/long (< 7 or > 80 words). On this basis, a total of 8,080 papers are selected to evaluate our approach, each containing more than 15 references found in the dataset and a related work section of at least 500 words. But as for the heterogeneous bibliography graph, all source data have to be imported to ensure the structural integrity of communities. Besides, this graph should be constructed year-by-year to preclude the effect of later publications on earlier ones.

Implementation We use *Tensorflow* for implementation, where both the dimensions of embedding and hidden state are equally 128. For the CNN, *word2vec* (Mikolov et al., 2013) is utilized to initialize the word embeddings, which can be further tuned during the training phase. Meanwhile, we follow the work of Kim (2014) to apply a list of kernels with widths $\{3, 4, 5\}$. As for the RNN, each LSTM module is set to one single layer, and all input documents are padded to the same length, along with a mark to indicate the real number of sentences. Based on these settings, we train our summarizer using *Adam* with the default in Kingma and Ba (2014), and perform *mini-batch cross-entropy* training with a batch of one target document for 20 epochs.

To create training data for our summarizer, each reference needs to be annotated with the ground truth in advance, i.e., candidate sentences are tagged with 0/1 for indicating summary-worthy or not. Specifically, we follow a heuristic practice of Cao et al. (2016) and Nallapati et al. (2016b) to compute ROUGE-2 score (Lin and Hovy, 2003) for each sentence, in terms of the native related work sections (gold standards). Next, those sentences with high scores are chosen as the positive samples, and the rest as the negative ones, such that the total score of selected sentences is maximized with respect to the gold standard. As for

testing, we relax the number of sentences to be selected, and focus on the classification probability from Eq. 8. In this study, *cross validation* is applied to split the dataset into ten parts equally at random, in which nine are used for training and the other one for testing.

Evaluation We adopt the widely used toolkit ROUGE (Lin and Hovy, 2003) to evaluate the summarization performance automatically. In particular, we report ROUGE-1 and ROUGE-2 (unigram and bigram overlapping) as a way to assess the informativeness, and ROUGE-L (the longest common subsequence) as a means to assess the fluency, in terms of fixed bytes of gold standards.

To validate the proposed attention mechanism, we compare our approach (denoted as $P_{\cdot S+N+Rteg+EUD}$) against six variants, including: 1) $P_{\cdot void}$: a plain seq2seq summarizer without attentions; 2) $P_{\cdot S}$: use the saliency as an only attention factor; 3) $P_{\cdot S+N}$: leverage both the saliency and novelty; 4) $P_{\cdot S+N+Rt}$: incorporate the relevance from the textual context; 5) $P_{\cdot S+N+Rtog}$: gain the relevance from the graphic context of a homogeneous citation graph; 6) $P_{\cdot S+N+Rteg}$: utilize the heterogeneous bibliography graph, but with each edge type the same usefulness.

In addition, we also select six representative summarization methods as a benchmark group. The first one is the general seq2seq summarizer by Cheng and Lapata (2016), denoted as PointerNet, which employs an attention mechanism to extract sentences directly after reading them. Following are five classical generic solutions, including: 1) Luhn (Luhn, 1958): a heuristic summarization based on word frequency and distribution; 2) MMR (Carbonell and Goldstein, 1998): a diversity-based re-ranking to produce summaries; 3) LexRank (Erkan et al., 2004): a graph-based summary technique inspired by *PageRank* and *HITS*; 4) SumBasic (Nenkova and Vanderwende, 2005): a frequency-based summarizer with duplication removal; 5) NltkSum (Acanfora et al., 2014): a *natural language tookit* (NLTK)-based implementation for summarization.

For clarity, Luhn, LexRank and SumBasic are analogous to the work of Hu and Wan (2014) which extracts sentences scoring the highest in significance, and they are also contrasted in the latest studies on neural summarizers (Chen et al., 2016; Tan et al., 2017). Meanwhile, MMR often serves as a part/post-processing of existing techniques to avoid the redundancy (Cohan and Goharian, 2017), and we introduce NltkSum to investigate the impact of grammatical/semantic analysis to the automatic related work summarization. Note that former studies specially for this task require extensive human involvements (see Table 1), thus we cannot apply them to such a large dataset of this study.

4.2 Results and Discussion

Table 2 reports the evaluation comparison over ROUGE metrics. From the top half, all scores appear a gradual upward trend with incorporation of saliency, novelty, relevance (from both textual and graphic contexts) and EUD into consideration one after another, which demonstrates the validity of our attention mechanism for summarizing related work sections. To be specific, we further reach the following conclusions:

1) $P_{\cdot void}$ vs. $P_{\cdot S}$ vs. $P_{\cdot S+N}$: Both saliency and novelty are two effective factors to locate the required content for summaries, which is consistent with prior studies.

2) $P_{\cdot S+N}$ vs. $P_{\cdot S+N+Rt}$: Contextual relevance does contribute to address the alignment between a related work section and its source documents.

3) $P_{\cdot S+N+Rt}$ vs. $P_{\cdot S+N+Rtog}$: Textual context alone cannot provide entire evidence to characterize the relationship among scientific publications exactly.

4) $P_{\cdot S+N+Rtog}$ vs. $P_{\cdot S+N+Rteg}$: Heterogeneous bibliography graph involves richer contextual information than a homogeneous citation graph.

5) $P_{\cdot S+N+Rteg}$ vs. $P_{\cdot S+N+Rteg+EUD}$: EUD plays an indispensable role in organizing accurate contextual relevance on a heterogeneous graph.
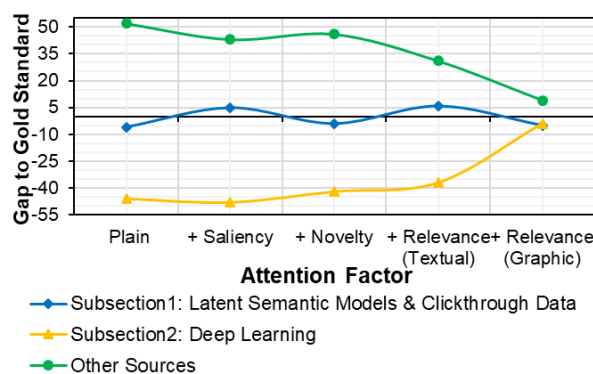

Figure 3: Number of extracted words on each reference cluster under different attention factors.

Continuing the "DSSM", Figure 3 visualizes the number of extracted words on each reference

Methods	ROUGE-1	ROUGE-2	ROUGE-L
$P_{\cdot void}$	26.85*	6.38*	14.22*
$P_{\cdot S}$	26.98*	6.48*	14.36*
$P_{\cdot S+N}$	27.29*	6.65*	14.43*
$P_{\cdot S+N+Rt}$	27.63*	6.72*	14.46*
$P_{\cdot S+N+Rtog}$	27.82*	7.00*	14.55*
$P_{\cdot S+N+Rteg}$	28.56*	7.40	14.70*
$P_{\cdot S+N+Rteg+EUD}$	**29.18**	**7.63**	**14.89**
Luhn	25.76*	5.08*	13.50*
MMR	25.55*	5.14*	13.99*
LexRank	25.07*	5.12*	13.95*
SumBasic	28.01*	5.44*	13.93*
NltkSum	28.07*	6.36*	14.87
PointerNet	27.06*	6.53*	14.41*

* indicates *Wilcoxon signed-rank test* $p < 0.01$, compared with $P_{\cdot S+N+Rteg+EUD}$

Table 2: Rouge evaluation (%) on 8,080 papers from ACM digital library.

cluster[3] under different attention factors. It can be seen that only after adding the relevance especially that from the graphic context into attentions, our summarizer can correctly sample the content from "Deep Learning" (yellow line), and eliminate that originated from "Other Sources" by a big margin (green line). As this example falls into the methodology transferring, a host of its involved word collocations are not idiomatic combinations yet, such as "Deep Neural Network" co-occurs with "Clickthrough Data" that is more frequently related to "Latent Semantic Analysis" at that time, which results in a somewhat biased textual context. By contrast, the graphic context will suffer less from this bias because it characterizes the connectivity patterns (real-time setup) instead of n-gram statistics, thus offering a more robust measure for the contextual relevance.

The bottom half of Table 2 illustrates the superiority of our approach over six representative summarization methods. Above all, Luhn, LexRank and MMR three summarizers that simply exploit shallow text features (word frequency and associated sentence similarity) to measure either significance or redundancy fall far behind the plain variant $P_{\cdot void}$, which partly reflects the strength of seq2seq paradigm in summarizing a related work section. Second, with combination of significance and redundancy, SumBasic achieves a drastic increase on ROUGE-1 and a mild raise on

ROUGE-2 respectively, but it still cannot improve ROUGE-L marginally. This is because simple text statistics cannot present deeper levels of natural language understanding to catch larger-grained units of co-occurrence. Third, NltkSum benefits from a NLTK library so as to access grammatical/semantic supports, thereby having the best informativeness (ROUGE-1 and ROUGE-2) among the five generic baselines, and meanwhile a comparable fluency (ROUGE-L) with our approach. Finally, as a deep learning solution, although PointerNet takes both hidden states and previously labeled sentences into account, at each decoding step it focuses on only current and just one previous sentences, lacking a comprehensive consideration on saliency, novelty and more importantly the contextual relevance ($< P_{\cdot S+N}$).

To better verify the summarization performance, we also conduct a human evaluation on 35 papers containing more than 30 references in the dataset. We assign a number of raters to compare each generated related work section against the gold standard, and judge by three independent aspects as: 1) How compliant is the related work section to the target document? 2) How intuitive is the related work section for readers to grasp the key content? 3) How useful is the related work section for researchers to prepare their final literature reviews? Note that we do not allow any ties during the comparison, and each property is assessed with a 5-point scale of 1 (worst) to 5 (best).

Table 3 displays how often raters rank each summarizer as the 1st, 2nd and so on, in terms of

---

[3]We pack the references cited in the same subsection of the related work section as one reference cluster.

Methods	1st	2nd	3rd	4th	5th	6th	7th	Mean Ranking
Luhn	0.04	0.07	0.09	0.13	0.17	0.23	0.29	5.26
MMR	0.05	0.07	0.11	0.16	0.19	0.22	0.20	4.82
LexRank	0.06	0.09	0.11	0.14	0.17	0.19	0.27	4.93
SumBasic	0.09	0.13	0.18	0.18	0.18	0.15	0.10	4.10
NltkSum	0.21	0.21	0.20	0.15	0.10	0.07	0.04	3.00
PointerNet	0.14	0.20	0.18	0.15	0.13	0.11	0.08	3.54
P.$_{S+N+Rteg+EUD}$	0.40	0.22	0.14	0.09	0.06	0.04	0.02	2.34

Table 3: Human evaluation (proportion) on 35 papers with more than 30 references in the dataset.

best-to-worst. Specifically, our approach comes the 1st on 40% of the time, which is followed by NltkSum that is considered the best on 21% of the time (almost half of ours), and Pointer-Net with quite equal proportions on each ranking. Furthermore, the other four summarizers account for obviously lower ratings in general. To attain the statistical significance, *one-way analysis of variance* (ANOVA) is performed on the obtained ratings, and the results show that our approach is better than all six contrast methods significantly ($p < 0.01$), which means that the conclusion drawn by Table 2 is sustained.

5 Conclusion

In this paper, we highlight the contextual relevance for the automatic related work summarization, and analyze the graphic context to characterize the relationship among scientific publications accurately. We develop a neural data-driven summarizer by leveraging the seq2seq paradigm, where a joint context-driven attention mechanism is proposed to measure the contextual relevance within full texts and a heterogeneous bibliography graph simultaneously. Extensive experiments demonstrate the validity of the proposed attention mechanism, and the superiority of our approach over six representative summarization baselines.

In future work, an appealing direction is to organize the selected sentences in a logical fashion, e.g., by leveraging a topic hierarchy tree to determine the arrangement of the related work section (Cong and Kan, 2010). We also would like to take the citation sentences of each reference into consideration, which is another concise and universal data source for scientific summarization (Chen and Hai, 2016; Cohan and Goharian, 2017). At the end of this paper, we believe that extractive methods are by no means the final solutions for literature review generation due to plagiarism concerns, and we are going to put forward a fully abstractive version in further studies.

References

Joseph Acanfora, Marc Evangelista, David Keimig, and Myron Su. 2014. Natural language processing: generating a summary of flood disasters. *Cell*, 41(2):383–94.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2016. End-to-end attention-based large vocabulary speech recognition. In *Proceedings of the 41st IEEE ICASSP International Conference on Acoustics, Speech and Signal Processing, Shanghai, China*, pages 4945–4949.

Jiajun Bu, Shulong Tan, Chun Chen, Can Wang, Hao Wu, Lijun Zhang, and Xiaofei He. 2010. Music recommendation by unified hypergraph:combining social media information and music content. In *Proceedings of the ACM SIGMM International Conference on Multimedia, Amsterdam, Netherlands*, pages 391–400.

Ziqiang Cao, Wenjie Li, Sujian Li, Furu Wei, and Yanran Li. 2016. Attsum: Joint learning of focusing and summarization with neural attention. *arXiv preprint arXiv:1604.00125*.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA*, pages 335–336.

Jingqiang Chen and Zhuge Hai. 2016. Summarization of related work through citations. In *Proceedings of the 12th IEEE SKG International Conference on Semantics, Knowledge and Grids, Beijing, China*, pages 54–61.

Qian Chen, Xiaodan Zhu, Si Wei, Si Wei, and Hui Jiang. 2016. Distraction-based neural networks for modeling documents. In *Proceedings of the ACM IJCAI International Joint Conference on Artificial Intelligence, New York, USA*, pages 2754–2760.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th ACL Annual Meeting of the Association for Computational Linguistics, Berlin, Germany*.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the NAACL Conference of the North American Chapter of the Association for Computational Linguistics, San Diego, USA*, pages 93–98.

Arman Cohan and Nazli Goharian. 2017. Scientific article summarization using citation-context and article's discourse structure. *arXiv preprint arXiv:1704.06619*, pages 390–400.

Ronan Collobert, Jason Weston, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(1):2493–2537.

Duy Vu Hoang Cong and Min Yen Kan. 2010. Towards automated related work summarization. In *Proceedings of the 23rd ACM COLING International Conference on Computational Linguistics, Beijing, China*, pages 427–435.

Swagatam Das and Ponnuthurai Nagaratnam Suganthan. 2011. Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation*, 15(1):4–31.

Erkan, Radev, and R Dragomir. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. *Journal of Qiqihar Junior Teachers College*, 22:2004.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, Usa*, pages 855–864.

Chun Guo and Xiaozhong Liu. 2015. Automatic feature generation on heterogeneous graph for music recommendation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile*, pages 807–810.

Chun Guo, Jinsong Zhang, and Xiaozhong Liu. 2013. Scientific metadata quality enhancement for scholarly publications. *Ischools*.

Sepp Hochreiter and Jrgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: an optimization approach. In *Proceedings of the ACL EMNLP Conference on Empirical Methods in Natural Language Processing, Doha, Qatar*, pages 1624–1633.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM CIKM international Conference on Information & Knowledge Management, San Francisco, USA*, pages 2333–2338.

Scott Jensen, Xiaozhong Liu, Yingying Yu, and Stasa Milojevic. 2016. Generation of topic evolution trees from heterogeneous bibliographic networks. *Journal of Informetrics*, 10(2):606–621.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Eprint Arxiv*.

Diederik Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. *Computer Science*.

Chin Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the NAACL The Annual Conference of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, USA*, pages 71–78.

H. P. Luhn. 1958. *The automatic creation of literature abstracts*. IBM Corp.

William C. Mann and Sandra A. Thompson. 2009. Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk*, 8(3):243–281.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016a. Sequence-to-sequence rnns for text summarization. In *Proceedings of the International Conference on Learning Representations, Workshop track, San Juan, Puerto Rico*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2016b. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. *arXiv preprint arXiv:1611.04230v1*.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira Dos Santos, Caglar Gulcehre, and Bing Xiang. 2016c. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023v5*.

Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Usa*, pages 542–550.

Shashi Narayan, Nikos Papasarantopoulos, Shay B. Cohen, and Mirella Lapata. 2017. Neural extractive summarization with side information. *arXiv preprint arXiv:1704.04530*.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research*.

L Page. 1998. The pagerank citation ranking : Bringing order to the web, online manuscript. *Stanford Digital Libraries Working Paper*, 9(1):1–14.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the ACL EMNLP Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal*, pages 379–389.

Ananthram Swami, Ananthram Swami, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Canada*, pages 135–144.

Jiwei Tan, Xiaojun Wan, Jianguo Xiao, Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th ACL Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada*, pages 1171–1181.

Scott White and Padhraic Smyth. 2003. Algorithms for estimating relative importance in networks. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, USA*, pages 266–275.

Dwi H Widyantoro and Imaduddin Amin. 2014. Citation sentence identification and classification for related work summarization. In *Proceedings of the ICACSIS International Conference on Advanced Computer Science and Information Systems*, pages 291–296.

Yingying Yu, Xiaozhong Liu, and Zhuoren Jiang. 2015. Random walk and feedback on scholarly network. In *Proceedings of the 1st ACM GSB@SIGIR International Workshop on Graph Search and Beyond, Santiago, Chile*, pages 33–37.