A Topic Modeling based Approach to Novel Document Automatic Summarization

Zongda Wu[a], Li Lei[a,*], Guiling Li[b], Chengren Zheng[a], Enhong Chen[c], Guandong Xu[d]

[a]*Oujiang College, Wenzhou University, Wenzhou, Zhejiang, China*
[b]*College of Computer Science, China University of Geosciences, Wuhan, China*
[c]*School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, China*
[d]*Faculty of Engineering and IT, University of Technology, Sydney, Australia*

Abstract

Most of existing text automatic summarization algorithms are targeted for multi-documents of relatively short length, thus difficult to be applied immediately to novel documents of structure freedom and long length. In this paper, aiming at novel documents, we propose a topic modeling based approach to extractive automatic summarization, so as to achieve a good balance among compression ratio, summarization quality and machine readability. First, based on topic modeling, we extract the candidate sentences associated with topic words from a preprocessed novel document. Second, with the goals of compression ratio and topic diversity, we design an importance evaluation function to select the most important sentences from the candidate sentences and thus generate an initial novel summary. Finally, we smooth the initial summary to overcome the semantic confusion caused by ambiguous or synonymous words, so as to improve the summary readability. We evaluate experimentally our proposed approach on a real novel dataset. The experiment results show that compared to those from other candidate algorithms, each automatic summary generated by our approach has not only a higher compression ratio, but also better summarization quality.

Keywords: Novel summarization, topic modeling, topic diversity, compression ratio, readability

1. Introduction

The exponential growth of online text documents on the World Wide Web leads to that the amount of text information people presently can access is much more than the sum of history text information, consequently, making it become more and more important and urgent to compress and summarize text documents. However, for such a huge amount of text information, a traditional manual method is obviously incompetent (Gambhir and Gupta, 2016). To this end, a new technique called automatic summarization was proposed, which, by using computers to automatically summarizing text documents, makes it much more efficient for the large amount of text information to be transferred and browsed on the World Wide Web. In text automatic summarization, extractive summarization is a common and mature technique, whose basic idea is to extract important sentences from text documents, and then recombine them to generate a summary of the text documents (Gambhir and Gupta, 2016; Das and Martins, 2007). The evaluation criteria for the quality of an extractive automatic summary can be summarized as how to not only reduce the redundancy rate of the summary, but also reflect the topic diversity of the source documents (Gambhir and Gupta, 2016). However, it is challenging for extractive summarization to achieve a good balance

*Corresponding Author
Email addresses: zongda1983@163.com (Zongda Wu), lilei4ac@gmail.com (Li Lei), freay@163.com (Guiling Li), zcr1072357579@gmail.com (Chengren Zheng), cheneh@ustc.edu.cn (Enhong Chen), guandong.xu@uts.edu.au (Guandong Xu)

between the two goals. At present, the extractive summarization has been widely applied into the field of multi-documents (i.e., the clustering of related documents of short length) (Ceylan, 2011; Chi et al., 2014).

A novel is a kind of common textual document. According to the explanation from Wikipedia[1], a novel refers to a narrative text document of structure freedom and long length (more than 45,000 words). However, most of existing automatic summarization algorithms are targeted for multi-documents with relatively short length, thus difficult to be applied immediately to summarize novel documents (Ceylan, 2011; Ceylan and Rada, 2007). Specifically, the existing automatic summarization algorithms may have the following problems, consequently, limiting their application in novel document automatic summarization. (1) A text document is generally of relatively short length. Most of the existing algorithms mainly focus on online review (Xiong and Litman, 2014), text page (Wang et al., 2007), text news (Lloret and Palomar, 2013) and so on. Obviously, the length of these text documents is much shorter than that of a novel document. For example, the length of a news article is shorter than that of a novel chapter (about 641 words versus 4,973 words) (Ceylan, 2011). Hence, it is difficult for the existing algorithms to meet the higher compression ratio requirement for summarizing a long novel document (about 10% versus 0.2%). (2) The short length of the documents also results in the limited space of sentence extraction and less context topics. However, for a novel document, its sentence selection space is large and its context topics are complicated. Therefore, it is more challenging to extract important sentences from a novel document, so as to generate an automatic summary with diverse topics under the precondition of a high text compression ratio. (3) Due to seldom considering the problem of efficiency, the existing algorithms generally have worse computational overhead. However, due to its long length, the summarization of a novel document has a much higher requirement on efficiency.

To overcome the above problems, in this paper, based on topic modeling, we propose an extractive summarization approach for a novel document (i.e., a single long text document). Note that a novel is generally organized according to some plot lines. Hence, the approach is developed based on "topic word association", i.e., we use topic modeling to obtain the topic words for a novel, and then expand the topic words to construct a machine summary for the novel. Specifically, based on topic modeling, we first extract the candidate sentences associated with the topic words from a novel document. Secondly, under the precondition of a high compression ratio, we design an importance evaluation function to select the candidate sentences with the most diverse topics to generate an initial summary. Finally, we smooth the initial summary to improve the readability. In addition, we experimentally compare the generated automatic summaries with the manual summaries to demonstrate the effectiveness of our approach. The main contributions of this paper are as follows.

- **Study object**. This paper is targeted for novel documents, which, compared to other documents, have longer length (each novel in our dataset contains about 20,0000 words), higher compression ratio (less than 0.2%) and more complex context (i.e., more diverse topics), leading to a greater challenge to automatic summarization. At present, there are few studies on novel summarization.
- **Topic modeling**. This paper uses topic modeling to capture topic words associated with a novel document, enabling the generated summary to reflect the novel context better than other extractive summarization algorithms, and thus improving the quality of the novel automatic summary.
- **Heuristic selection**. In view of the style particularity of a novel, by combining stylistic features, with the goals of topic diversity and redundancy rate, this paper presents a candidate sentence importance evaluation function and then an efficient algorithm for extractive automatic summarization.
- **Information fusion**. Based on external resources such as SemCor and synonym thesaurus, we smooth each automatic summary to overcome the semantic confusion problem caused by polysemy and synonymy, so as to improve the machine readability of the automatic summary.

The rest of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 formulates topic diversity and compression ratio, and then the problem of extractive novel summarization.

[1]https://en.wikipedia.org/wiki/novella

Section 4 proposes our approach to automatic novel summarization, which first presents a sentence importance evaluation function, and then describes how to conduct heuristic sentence selection and summary smoothing operation. Section 5 presents the experimental evaluation results. Finally, we summarize this paper and discuss the future work in Section 6.

2. Related work

Presently, there have been a number of studies related to extractive automatic summarization, but there are few studies related to novel summarization. In this section, we briefly review the work higher relevant to the study of this paper, including: single document summarization, multi-document summarization and topic modeling summarization.

2.1. Single document summarization

Single document summarization is the process of generating a summary for a single text document, which is the focus of earlier studies on automatic summarization. However, in existing studies, the targeted single documents are generally regular and of short length, e.g., a technological article (Das and Martins, 2007). As pointed out in (Ceylan and Rada, 2007), the approaches proposed in the existing studies are often difficult to be applied to summarize a single document with structure freedom. In (Kazantseva and Szpakowicz, 2010), the authors noted that it is a challenging task to automatically summarize short story documents. In order to summarize the main characters and locations in a story document, by using a machine learning technique combined with manual rules, the authors proposed a summarization approach which can achieve an average compression ratio about 6%. Although achieving good results, it is still an unsolved problem for the approach how to further improve the compression ratio, so as to make it capable of summarizing single documents with longer length. In (Ceylan and Rada, 2007), to overcome the disadvantage of traditional text summarization techniques difficult to be applied in long documents, the authors proposed a summarization approach for long single documents, where the average length of each text document is up to $90,000$ words and the summary compression ratio is about 10%. However, the summary compression ratio of the approach is still high, limiting its practical availability. In (Bamman and Smith, 2013), based on the observation that it is difficult to align each source text sentence with its corresponding summary sentence, two new sentence alignment methods are proposed, which can greatly improve the quality of the generated summaries. In addition, the work also improved the summary compression ratio, reaching to about 1%.

In summary, the earlier methods on single document summarization are usually difficult to be applied to summarize the literature documents of freedom structure. However, existing novel document summarization methods are either not designed for novel documents (i.e., the document length does not meet the novel requirement), or difficult to meet the practical requirement on the compression ratio (an ideal compression ratio of a novel should be less than about 0.2%). Therefore, it is still an unsolved problem how to improve the quality of the novel summarization under the precondition of a high compression ratio.

2.2. Multi-document summarization

Multi-document summarization refers to extracting the important sentences from a cluster of relevant documents, and combining them to form a descriptive summary of the documents (Das and Martins, 2007). The earliest studies on multi-document summarization mainly focus on news documents, and there have been a number of good research results (Alguliev et al., 2013; Ferreira et al., 2014). A long document (such as novels) can also be divided into several short multi-documents according to the document chapters, so that we can use multi-document summarization techniques to realize the automatic summarization for a single long text document. However, as mentioned above, it still has a big disparity between the text length of a multi-document and the length of a novel chapter, resulting in a large computational overhead. For example, in (Alguliev et al., 2013), the authors used an evolutionary algorithm to carry out multi-document summarization, thereby making the generated machine summaries of low redundancy rate and better content correlation, but leading to a relatively large computational overhead. In addition, because of

the strong context and semantic coherence between novel chapters, and the lack of the narrative coherence between the traditional multi-documents, it is difficult to apply the multi-document summarization techniques directly to summarize novel documents. For example, in (Tran et al., 2015), the authors used a joint graph model to carry out the multi-document summarization on the events which have occurred at different times, and finally obtained a good result. However, the simple event time series cannot deal with the complex plot lines in a novel. According to the topic distribution, the paper (Yang et al., 2014) used topic clustering and topic ranking to conduct multi-document summarization, thereby generating high quality summaries, and effectively controlling the redundancy rate of the summaries. However, due to the special topic distribution of the novel body, this approach has to sacrifice the topic diversity of a novel document to a certain extent.

In summary, it is difficult for a multi-document summarization approach to be directly applied to the automatic summarization of a novel document, because of its short text length and single semantic topics, as well as the high computational overhead.

2.3. Topic modeling summarization

The basic idea of topic modeling summarization is to view the text as a cluster of many topic words (Blei et al., 2003). In view of the global and local distribution characteristics of the novel text, the topic diversity is also an important evaluation metric of the quality of novel summarization (Yang et al., 2015). Therefore, using topic modeling techniques to summarize novel documents should be able to greatly improve the quality of the topic selection. In (Bairi et al., 2015), aiming at $8,000$ Wikipedia ambiguity pages with the same titles but different topics, the authors used a topic modeling technique to extract a set of understandable topic words, so as to realize the simplification of a large-scale data set. In (Riddell, 2013), with the help of a topic modeling method, in accordance with the characteristics of literary style, an approach was proposed to classify 93 classical novel documents of an average length about $75,000$ words. In (Yuan et al., 2015), an approach was proposed to construct a recommendation system by using topic modeling to balance the relevance and diversity of user interests. In summary, we can see that the topic modeling methods are not only suitable for large-scale text documents, but also can effectively explore the topic relationship inherent in a text document, thereby enhancing the topic diversity. Therefore, the topic modeling is suitable for the novel automatic summarization. However, there has been few topic model based automatic summarization methods for novel documents.

3. Problem statement

As mentioned in the introduction section, in automatic summarization, there are two important goals, i.e., how to reflect the topic diversity of source novel text (so as to ensure the summary quality), and how to reduce the redundancy rate of a summary (so as to ensure the compression ratio). In this section, we formulate the problem of extractive novel automatic summarization. Table 1 describes some symbols used in this paper.

Definition 1 (Compression ratio). A novel can be represented as a set of sentences, i.e., $\mathrm{S} = \{\mathrm{s}\}$. An extractive summary of the novel S also can be represented as a set of sentences, i.e., $\mathrm{S}^a = \{\mathrm{s}^a\}$. Obviously, we have that $\mathrm{S}^a \subseteq \mathrm{S}$. Then, the compression ratio of the summary S^a related to the novel document S can be defined as

$$\mathbf{compr}(\mathrm{S}^a) = \left(\sum_{\mathrm{s}^a \in \mathrm{S}^a} \mathbf{size}(\mathrm{s}^a) \Big/ \sum_{\mathrm{s} \in \mathrm{S}} \mathbf{size}(\mathrm{s}) \right) \tag{1}$$

Definition 2 (Topic distribution). Let $\mathrm{T} = \{\mathrm{w}^t\}$ denote the topic space consisting of all the topic words, and $Pr(\mathrm{w}^t|\mathrm{S})$ denote the probability of occurrences of a topic word w^t in a text document S. Then, the topic distribution of the text document S can be described using the following vector

$$\mathbf{T}(\mathrm{S}) = \left(Pr(\mathrm{w}_1^t|\mathrm{S}), Pr(\mathrm{w}_2^t|\mathrm{S}), ..., Pr(\mathrm{w}_n^t|\mathrm{S})\right), \text{where } n = \mathbf{size}(\mathrm{T}), \mathrm{w}_1^t, \mathrm{w}_2^t, ..., \mathrm{w}_n^t \in \mathrm{S}$$

Table 1: Symbols and Their Explanations

Symbol	Explanation
$\mathrm{S} = \{\mathrm{s}\}$	A novel, represented as a set of sentences
$\mathrm{S}^c = \{\mathrm{s}^c\}$	A set of candidate sentences, $\mathrm{S}^c \subseteq \mathrm{S}$
$\mathrm{S}^a = \{\mathrm{s}^a\}$	A summary, represented as a set of summary sentences, $\mathrm{S}^a \subseteq \mathrm{S}^c$
$\mathrm{S} = \{\mathrm{w}\}$	A sentence, represented as a set of words
$\mathrm{T} = \{\mathrm{w}^t\}$	The topic space, consisting of all the topic words
$\mathbf{T}(\mathrm{S})$	The topic distribution vector of a text S
$\mathbf{compr}(\mathrm{S}^a)$	The compression ratio of a summary, whose value range is between 0 and 1
$\mathbf{diver}(\mathrm{S}^a)$	The topic diversity of a summary, whose value range is between 0 and 1

Definition 3 (Topic diversity). Given a novel S and its summary S^a, the summary topic diversity can be measured by the cosine similarity between the topic distribution vectors $\mathbf{T}(\mathrm{S})$ and $\mathbf{T}(\mathrm{S}^a)$, i.e.,

$$\mathbf{diver}(\mathrm{S}^a) = \mathbf{cos}\angle\mathbf{T}(\mathrm{S}^a), \mathbf{T}(\mathrm{S}) = \frac{\mathbf{T}(\mathrm{S}^a) \cdot \mathbf{T}(\mathrm{S})}{\parallel \mathbf{T}(\mathrm{S}^a) \parallel \cdot \parallel \mathbf{T}(\mathrm{S}) \parallel} \tag{2}$$

Based on Definitions 1 and 3, we can further formulate the requirements that an ideal extractive summary should satisfy, i.e., the problem of extractive novel automatic summarization.

Definition 4 (Novel summarization). Given a novel document S, the problem of extractive novel summarization can be defined as how to automatically obtain a set S^a of sentences (i.e., a summary) from the novel S, so as to meet the following two requirements as much as possible.

- High summary compression ratio, i.e., $\min\limits_{\mathrm{S}^a} f(\mathrm{S}^a) = \mathbf{compr}(\mathrm{S}^a)$ **s.t.** $\mathrm{S}^a \subseteq \mathrm{S}$.
- Good topic diversity (summary quality), i.e., $\max\limits_{\mathrm{S}^a} f(\mathrm{S}^a) = \mathbf{diver}(\mathrm{S}^a)$ **s.t.** $\mathrm{S}^a \subseteq \mathrm{S}$.

It can be observed that the two requirements contradict with each other. On the one hand, to obtain high compression ratio, an ideal summary should contain as few sentences as possible. On the other hand, to obtain good quality, an ideal summary should cover as many topics of the source novel text as possible. Hence, we use the following equation to combine the two requirements together, i.e., the problem of extractive novel summarization is redefined as:

$$\max_{\mathrm{S}^a} f(\mathrm{S}^a) = \gamma \cdot \mathbf{diver}(\mathrm{S}^a) + (1-\gamma) \cdot \frac{1}{\mathbf{compr}(\mathrm{S}^a)} \ \mathbf{s.t.}\ \mathrm{S}^a \subseteq \mathrm{S} \tag{3}$$

, wherein, $\gamma \in [0, 1]$ is a parameter used to balance the two goals, i.e., the greater the parameter, the more important the topic diversity, and otherwise the more important the compression ratio.

Now, the goal of this paper is described as how to efficiently search a set of sentences (i.e., an automatic summary) satisfying the above equation from a given novel document (Alguliev et al., 2012).

4. Proposed approach

The extractive novel summarization model used in this paper is shown in Fig. 1, which consists of the following four steps.

- **Data preprocess**: i.e., preprocess the novel text, including word segmentation, removing stopwords, stemming and so on. After preprocessing, the novel text information will be more concentrated.

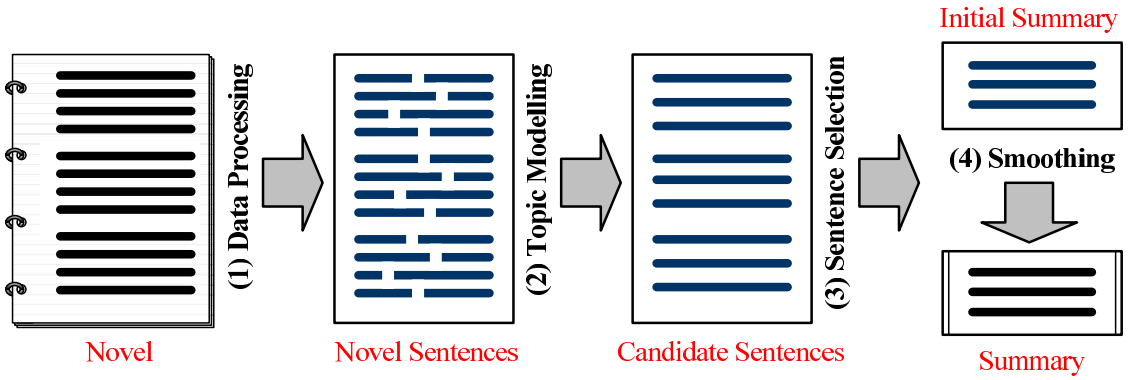


Figure 1: The System Model of Extractive Novel Summarization

- **Topic modeling**: i.e., use a topic model to summarize the sentences in the novel document, so as to obtain the distribution probability of each topic word in the source novel; and then trace back to the sentences associated with the topic words, so as to obtain a set of candidate sentences.
- **Sentence selection**: design an importance evaluation function of candidate sentences, and then according to the desired summary compression ratio, select the sentences with the highest importance scores, so as to obtain an initial machine summary.
- **Summary smoothing**: smooth the initial machine summary, so as to overcome the semantic confusing problem caused by synonymy and polysemy, and thus improve the machine readability of the summary.

4.1. Data preprocess

The reference datasets can be divided into two parts. **(1) A novel dataset**. From Gutenberg Project[2], we choose 63 narrative novels as the novel dataset. The length of each novel is more than 100,000 words and the average length is about 200,000 words. Compared to those used by other studies (Ceylan, 2011; Kazantseva and Szpakowicz, 2010; Bamman and Smith, 2013), each document in the dataset we use has a more uniform length, and can meet the length requirement on a novel. **(2) A summary dataset**. From the Internet, we also gather a number of manual summaries for the novels from Gutenberg Project, used as the reference dataset of subsequent evaluation for automatic summaries. The average length of each manual summary is equal to 500 words. In addition, each manual summary consists of three parts, i.e., the beginning, the body and the ending. Finally, we obtain a manual summary dataset with an average compression ratio about 0.17%. Fig. 2 describes the characteristics of the novel dataset and its summary dataset. Before topic modeling and sentence selection, we need to preprocess the documents in the novel dataset, including chapter segmentation, sentence segmentation, word segmentation, removing stopwords and stemming.

(1) Chapter segmentation. In general, a novel consists of tens of chapters, each of which is assigned by the novel author directly, and the novel chapters are relatively independent of each other. As a result, we can extract the topics for each chapter independently, such that we can use multi-threads to improve the efficiency of the subsequent topic modeling operation, without compromising the effectiveness of topic extraction.

(2) Sentence segmentation. In automatic summarization, the minimal processing unit is a sentence. In our work, we use NLTK, a well-known sentence segmentation tool (Bird et al., 2009), whose basic idea is to scan the text document, and generate a new sentence when encountering a sentence terminator. After sentence segmentation, each novel document can be expressed as a set of sentences, denoted by $S = \{s\}$.

[2]http://www.gutenberg.org

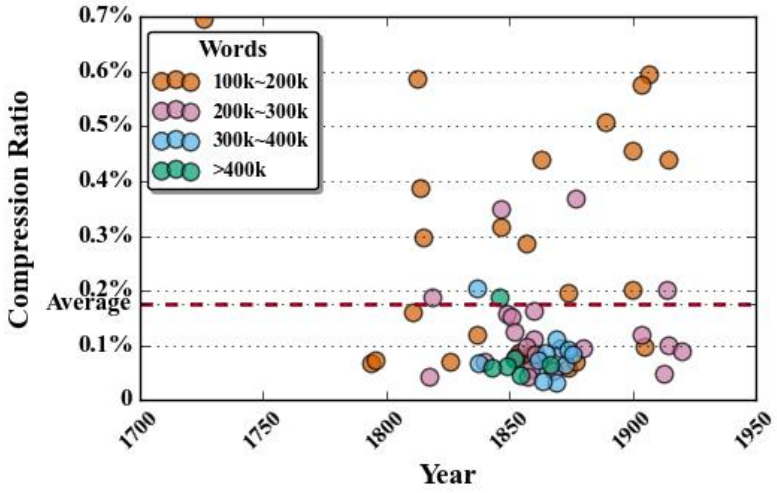


Figure 2: The Characteristics of the Reference Dataset

(3) Word segmentation. It refers to expressing a novel sentence as a set of independent words. Since the English language generally uses space character as a separator, the word segmentation is relatively simple. Now, each sentence $\mathrm{s} \in \mathrm{S}$ is further expressed as a set of words, denoted by $\mathrm{s} = \{\mathrm{w}\}$. In addition, in the word segmentation, we also turn each keyword to lowercase, to facilitate the subsequent processing.

(4) Removing stopwords. Stopwords are the words having no concrete meanings (prepositions, pronouns, articles etc.). These words do not carry any useful information, so we need to remove them in order to avoid interference with our approach. In this paper, we use the stop list given by NLTK to remove stop words for the word set generated by the step of word segmentation.

(5) Stemming. Each word has its stem, so stemming means to change words in different tenses (e.g., past tense, present continuous tense) and different parts of speech (e.g., noun, verb) to their word stems. A stemming operation can centralize the language information, to reduce the calculation scale of follow-up steps. In this paper, we use the famous Snowball tool[3] to carry out stemming.

4.2. Topic modeling

In our approach, the goal of topic modeling is to search the topic words related to a novel document so as to obtain the summary candidate sentences. Topic modeling refers to mining the topics implicitly contained in a text document (Blei et al., 2003). For example, if in an article, there are a number of words such as "earthquake", "survival" and "rescue", then it is very likely that the main topics of this article are related to "earthquake rescue". Here, we use the LDA algorithm (Latent Dirichlet Allocation) for topic modeling and sentence extraction. LDA is a well-known unsupervised learning topic modeling algorithm, which uses the occurrence probability of words to describe the topics of a document. LDA can be described as follows

$$Pr(\mathrm{w}|\mathrm{S}) = \sum_{\mathrm{w}^t \in \mathrm{T}} Pr(\mathrm{w}|\mathrm{w}^t) \cdot Pr(\mathrm{w}^t|\mathrm{S}) \tag{4}$$

, where each symbol is explained as follows:

- $Pr(\mathrm{w}|\mathrm{S})$: the probability of occurrences of a word w in a novel document S, which is a known quantity, whose value is equal to the number of occurrences of w in S divided by the number of all the words in S.

[3]http://snowball.tartarus.org/texts/introduction.html

- $Pr(\mathrm{w}|\mathrm{w}^t)$: the probability of occurrences of a word w under the precondition that the topic w^t is known, which is used to describe the relevance of a word w to a topic corresponding to w^t.
- $Pr(\mathrm{w}^t|\mathrm{S})$: the probability of occurrences of each topic w^t in a novel document S, which is used to describe the relevance of a topic word w^t to a document S.

Given a set of novel documents, using a large number of known quantities $Pr(\mathrm{w}|\mathrm{S})$, the LDA algorithm can train two sets of unknown quantities, $Pr(\mathrm{w}|\mathrm{w}^t)$ and $Pr(\mathrm{w}^t|\mathrm{S})$, so it can be used to calculate and obtain the novel topics from a set of novel documents. In the LDA algorithm, each novel document is represented as a probability distribution of certain topic words, and each topic is a probability distribution of a number of words. Given a novel document S, the LDA algorithm can be described briefly by the following iteration process: (1) from each chapter of the novel S, obtain a topic w^t, according to the topic distribution of the chapter; (2) obtain a word w from the word distribution of the topic w^t; and (3) repeat the above process until not only each word of the chapter but also each chapter of the novel S have been traversed. For the novel summarization model shown in Fig. 1, the topic modeling operation is the most time-consuming among all the steps, which determines the efficiency of summarization. Hence, in the above process, we combine with multi-threads, i.e., by assigning a single thread for each novel chapter of the novel, to improve the topic modeling efficiency. Finally, we obtain a set of topic words, and the distribution probability of each topic word. Then, we trace back to all the sentences (i.e., the topic sentences, or called the candidate sentences) associated with the topic words. As a result, for the novel S, after topic modeling, we can obtain a set of candidate sentences, denoted by $\mathrm{S}^c = \{\mathrm{s}^c\}$ (obviously, $\mathrm{S}^c \subseteq \mathrm{S}$).

In the experiment, we use the tool Gensim (Rehurek and Sojka, 2010) (the version is 0.13.1) to carry out LDA topic modeling, which is an open source third party library developed based on the Python programming language, and has been widely used in LDA topic modeling. It should be noted that except LDA, several other methods can also be used to extract the topics for a text document, such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Explicit Semantic Analysis (ESA) (Evgeniy and Shaul, 2007), Hierarchical Dirichlet Process (HDP) (Teh et al., 2006) and TextRank (Mihalcea and Tarau, 2004). Here, the reason that we choose LDA is because it generally has better overall performances in terms of efficiency (vis-a-vis ESA), simplicity (vis-a-vis HDP and LSA) and effectiveness (vis-a-vis TextRank). In the experiments, we use these topic modeling methods as candidates, and compare them with our approach (see the experiment section for detail).

4.3. Sentence selection

After topic modeling, a novel document S is transformed into a set of candidate sentences, i.e., $\mathrm{S}^c = \{\mathrm{s}^c\}$. Obviously, the sentence set covers all the topics of the novel S, so if it is used directly as an automatic summary of the novel, it can well meet the requirement on topic diversity. However, the size of the sentence set is too large (i.e., the number of candidate sentences is much greater than the length of an ideal summary), making it difficult to achieve the requirement on high compression ratio. To this end, we need to select the most important sentences from the candidate sentence set to generate an ideal summary for the novel S.

From the objective function (i.e., Equation 3) presented in Section 3, we can see that it is very time-consuming if we directly use it to search a summary from the candidate sentence set S^c. The time complexity is equal to $O\left(\binom{|\mathrm{S}^c|}{\theta\cdot|\mathrm{S}|}\right)$ (where θ is a desired compression ratio), and it is also equal to $O\left(\binom{|\mathrm{S}|}{\theta\cdot|\mathrm{S}|}\right)$ (since $|\mathrm{S}^+| \approx |\mathrm{S}|$). However, since the size of S is large, such an exhaustive method is not feasible in practice (it is NP-hard). Hence, we use the following heuristics to carry out sentence selection. First, we think that for the automatic summarization of a novel, high summary compression ratio is the primary goal that has to be satisfied, and thus we can translate the multi-objective optimization problem into a single objective optimization problem, i.e., the problem of novel automatic summarization can be redefined as follows.

$$\max_{\mathrm{S}^a} f(\mathrm{S}^a) = \mathbf{diver}(\mathrm{S}^a) \ \mathbf{s.t.}\, \mathrm{S}^a \subseteq \mathrm{S}, \mathbf{compr}(\mathrm{S}^a) > \theta \tag{5}$$

, wherein θ is an expected compression ratio, and in the subsequent experiment, its value is set to ensure the length of a machine summary not more than 500 words. Then, we define a sentence importance evaluation

function to quantify the important degree of each candidate sentence on the topic diversity. As a result, the optimization search problem in a combination space can be transformed into a greedy search in a linear space. Here, the sentence importance evaluation is based on the performance of each candidate sentence in terms of topic diversity and redundant information overload. Finally, we choose the most important candidate sentences to generate a machine summary for the novel.

Observation 1 (Positive topic diversity). For any sentence in a novel document, the more topics the sentence is related to, the more important the sentence; and the greater the number of occurrences of the related topics in the novel, the more important the sentence.

For example, given two novel sentences s_1 and s_2 in a novel document S, if the sentence s_1 is associated with two topics of higher occurrence frequencies in S, and the other s_2 is only associated with one topic of a lower occurrence frequency. Then, the sentence s_1 can reflect the topic diversity better than s_2, i.e., s_1 is more important.

Definition 5 (Positive topic diversity). Given any candidate sentence $s^c \in S^c$, the positive topic diversity of the sentence can be measured as follows

$$\textbf{posdiver}(s^c) = \textbf{size}\left(\{w|w \in T, w \in s^c\}\right)^{\frac{1}{\theta_1}} \sum_{w \in T, w \in s^c} Pr(w|S) \tag{6}$$

, wherein $\theta_1 \geq 1$ is a parameter.

Example 1. For a sentence in the novel "Jane Eyre" as "I never liked long walks, especially on chilly afternoons: dreadful to me was the coming home in the raw twilight, with nipped fingers and toes, and a heart saddened by the chidings of Bessie, the nurse, and humbled by the consciousness of my physical inferiority to Eliza, John, and Georgiana Reed", where "Reed" and "John" are two topic words, we assume that the occurrence probabilities of them are 0.013 and 0.008, respectively. Then, the positive topic diversity value of the sentence is $0.021 \cdot \sqrt{2}$, which is equal to the sum of 0.013 and 0.008 multiplying by $\sqrt{2}$ ($\theta_1 = 2$).

Observation 2 (Negative topic diversity). Given a temporal summary and a sentence, if any topic related to the sentence does not appear in the summary, then the sentence is important (since the redundancy rate of the sentence related to the summary is small); otherwise, if the greater the number of occurrences of related topics in the summary, then the more unimportant the sentence.

For example, assume that we have obtained a complete novel summary S^a (to simplify the presentation, we assume that the summary contains only one sentence associated with a topic w_1^t). Then, given two novel sentences s_1 and s_2 respectively associated with two equally important topics w_1^t and w_2^t, it is considered by Observation 2 that the sentence s_1 contains topic redundancy (because the topic w_1^t related to s_1 has appeared in the summary S^a), and the other sentence s_2 is more important.

Definition 6 (Negative topic diversity). Given a temporal summary S^a of a novel S, for any candidate sentence $s^c \in S^c$, the negative topic diversity of the sentence can be measured as follows.

$$\textbf{negdiver}(s^c) = 1 + \sum_{w^t \in T, w^t \in s^c} \textbf{num}(w^t, S^a)^{\frac{1}{\theta_2}} \tag{7}$$

, wherein, $\textbf{num}(w^t, S^a)$ denotes the number of occurrences of a topic word w^t in the summary S^a, and $\theta_2 \geq 1$, which is a parameter.

Example 2. For the sentence given in Example 1, we assume that for the topic words "Reed" and "John" related to the sentence, the numbers of occurrences of the two topic words in a temporal summary are 2 and 1, respectively. If $\theta_2 = 1$, then the negative topic diversity of the sentence is $1 + 2 + 1 = 4$ (the smaller the negative topic diversity, the more important the sentence).

Observation 3 (Information redundancy). For any sentence in a novel, the more useless words (e.g., stopwords) it contains, the less important the sentence; otherwise, the less useless words, the more important the sentence.

For example, given two novel sentences s_1 and s_2, if they are related to the same topics, but the number of useless words contained in s_1 is greater than that of s_2, then due to the requirement on a high compression ratio, obviously, it is more appropriate to select the sentence s_2 (i.e., s_2 is more important) than s_1. This is because although both have the same topic diversity, the sentence s_1 has more redundant information.

Definition 7 (Redundancy rate). For any candidate sentence $s^c \in S^c$, let W denote a set of all the useless words. Then, the information redundancy rate of the sentence can be measured as follows

$$\textbf{redun}(s^c) = \frac{1}{\textbf{size}(s^c)} \sum_{w \in W, w \in s^c} \textbf{num}(w, s^c) \tag{8}$$

, wherein, $\textbf{num}(w, s^c)$ denotes the number of occurrences of a word w in the sentence s^c.

From Definition 7, we see that the more useless words a sentence contains, the less information it contains, i.e., the greater the information redundancy rate. For example, for a sentence "What do you do", its information redundancy rate is equal to 1, which indicates that the useful information contained in the sentence is almost equal to 0.

Definition 8 (Sentence importance). Based on Equations 6-8, we obtain a sentence importance evaluation function as follows (the bigger the value, the more important the sentence).

$$\textbf{diver}(s^c) = \frac{\textbf{posdiver}(s^c)}{\textbf{negdiver}(s^c)} \cdot (1 - \textbf{redun}(s^c)) \tag{9}$$

Observation 4 (Sentence position). In general, a narrative novel can be divided into three parts (Leite et al., 2007): the beginning, the body and the ending, and their information quantities are different from each other. Thus, an ideal summary of the novel should contain the corresponding three parts so as to keep up with the topic diversity of the novel text.

Based on Observation 4, we can divide the candidate sentences into three subsets: a beginning set, a body set and an ending set. Next, we select the most important sentences from the three sets, respectively, and then combine them to generate a machine summary of the novel document. The above selection process can be briefly described as follows. First, we determine the proportions of the beginning, body and ending parts in a novel document, denoted by ρ_1, ρ_2 and ρ_3, respectively. According to the general regularity of a narrative novel[4], the proportions of the beginning and ending parts can be both set to 20%, and the proportion of the body part is set to 60%, i.e., $\rho_1 = \rho_3 = 0.2$ and $\rho_2 = 0.6$. Second, according to the candidate sentence set $S^c = \{s_i^c\}_{i=1}^m$ determined by the topic modeling operation (where m denotes the number of all the candidate sentences), we determine the three subsets of candidate sentences as: $S_1^c = \{s_i^c\}_{i=1}^{m_1}$, $S_2^c = \{s_i^c\}_{i=m_1+1}^{m_2}$ and $S_3^c = \{s_i^c\}_{i=m_2+1}^{m}$, where $m_1 = \lceil m\rho_1 \rceil$ and $m_2 = \lceil m(\rho_1 + \rho_2) \rceil$. Finally, we respectively choose the most important sentences from the three subsets to form an automatic summary. Specifically, based on the sentence importance evaluation function, we select the $\theta \cdot |S| \cdot \rho_1$ most important sentences from S_1^c, denoted by S_1^a, the $\theta \cdot |S| \cdot \rho_2$ sentences from S_2^c, denoted by S_2^a, and the $\theta \cdot |S| \cdot \rho_3$ sentences from S_3^c, denoted by S_3^a. As a result, we obtain the final summary $S^a = S_1^a \cup S_2^a \cup S_3^a$. Algorithm 1 details the extractive novel summarization approach.

It can be observed that if we ignore the time overhead from the steps of data preprocessing and topic modelling (i.e., Lines 2-3), the time overhead of Algorithm 1 is mainly dependent on the operation of sentence selection (i.e., Lines 9-15), so the time complexity of Algorithm 1 is equal to $O((\rho_1+\rho_2+\rho_3) \cdot m^2 \cdot \theta)$, i.e., $O(m^2 \cdot \theta)$, where m denotes the number of the candidate sentences from a novel document.

[4]https://en.wikipedia.org/wiki/Wikipedia:How_to_write_a_plot_summary

Algorithm 1: Extractive Novel Automatic Summarization

Input: A novel document $S = \{s\}$.
Output: A novel summary S^a.

begin
  Preprocess the novel document S by chapter segmentation, sentence segmentation, word segmentation, removing stopwords and stemming;
  Leverage the topic modeling algorithm LDA to obtain a set of candidate sentences from the novel S, denoted by $S^c = \{s^c\}$;
  Divide the set S^c into three subsets, i.e., S_1^c, S_2^c and S_3^c;
  foreach $s^c \in S_1^c \cup S_2^c \cup S_3^c$ **do**
    Calculate the positive topic diversity **posdiver**(s^c) of the candidate sentence s^c;
    Calculate the information redundancy **redun**(s^c) of the candidate sentence s^c;
  Set S_1^a, S_2^a and S_3^a to be empty;
  for $k = 1; k \leq 3; k = k + 1$ **do**
    while compr$(S_k^c) < \theta$ **do**
      foreach $s^c \in S_k^c$ **do**
        Based on the current summary $S_1^a \cup S_2^a \cup S_3^a$, calculate the negative topic diversity **negdiver**(s^c) of the candidate sentence s^c;
        Based on **posdiver**(s^c), **redun**(s^c) and **negdiver**(s^c), calculate **diver**(s^c);
      From S_k^c, obtain the most important candidate sentence s^c;
      Add s^c into S_k^a, and remove s^c from S_k^c;
  Return an initial novel summary $S^a = S_1^a \cup S_2^a \cup S_3^a$;

4.4. *Summary smoothing*

In a summary, the existence of polysemous or synonymous words results in a great deal of obstacles to semantic analysis. In order to solve the synonymy problem, we transform some synonymous words in a machine summary S^a into relatively simple words (i.e., basic words)[5], so as to improve the machine readability. To this end, we first need to introduce some external language resources, and build their corresponding internal data structures.

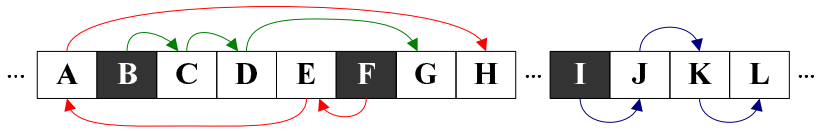


Figure 3: Synonym Net, where the black blocks denote low level words, and the white blocks denote non-low level words

To deal with the synonymy problem in a novel summary, we construct a synonym network. First, from the online version of Roget Thesaurus[6], which is a large dictionary of synonyms (Sinha and Mihalcea, 2009; Jarmasz and Szpakowicz, 2003), we download a set of about $250,000$ synonymous words, where each word corresponds to several synonyms. For example, "good" is a synonym of "great" or "wonderful", so they belong to the same group in the synonym network. Second, we use the "basic word" provided by the Oxford dictionary[7] to group these synonymous words, so as to construct a synonym network. Note that the "basic words" are low-level words extracted by linguists, which can help English learners better understand a text document. Finally, we generate the synonym network shown as Fig. 3, where each end point denotes a "basic word" and each point connecting to an end point denotes a synonymous word of

[5]https://simple.wikipedia.org/wiki/Basic_English
[6]http://www.thesaurus.com/
[7]http://www.oxfordlearnersdictionaries.com/wordlist/english/oxford3000/

the "basic word". In addition, we also sort all the words in the synonym network to improve the efficiency of searching words. With the help of the synonym network, we can convert all the synonymous words in a machine summary S^a to their corresponding basic words, thereby, eliminating the synonym problem and as a result improving the machine readability of automatic summaries.

In addition, there also exists the polysemy problem in a novel summary, e.g., for a polysemous word "Puma", it is difficult for a machine to determine its meaning. In fact, the semantic disambiguation problem can be regarded as a classification task (Navigli, 2009). An effective approach is to use a data set with semantic and part of speech tagging to train a semantic classifier, so that given a target word and its context information, based on the trained classifier, we can obtain the most appropriate semantic meaning of the target word. Here, we use SemCor (Miller et al., 1993) as our training data set. SemCor is a subset of the Brown Corpus, including a total of $360,000$ words and about $234,000$ semantic annotations, which has been widely used for text semantic disambiguation (Fernandez-Amoros and Heradio, 2011).

In short, the above operations of transforming polysemy words and synonymous words into their basic words are called as a basic word translation algorithm. With the help of the basic word translation algorithm, the semantic disambiguation problem in a machine summary generated by Algorithm 1 can be well solved, consequently improving the machine readability of the final machine summary.

5. Evaluation experiment

In this section, we evaluate our approach by experiments from the following two aspects. First, by comparison with other five candidates, we evaluate the effectiveness of our approach on topic diversity. Second, with the help of some evaluation criteria combined with manual summaries, we evaluate the actual quality of the summaries generated by our approach.

5.1. Experimental setup

First of all, we describe our experimental setup, including summary evaluation criteria and candidate algorithms. In addition, since the novel reference dataset and its corresponding manual summary dataset have been described in Section 4.1, we no longer repeat them here.

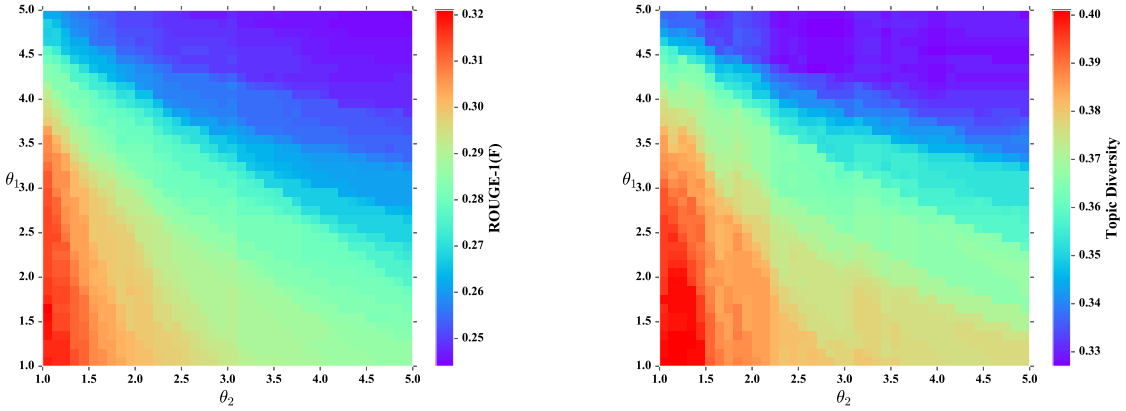
(1) Summary evaluation: We use two methods to evaluate the actual quality of a machine summary, i.e., a manual approach and the ROUGE criteria. First, we invite a group of assessors to score each automatic summary based on the relevance of the summary to its novel document. Second, in view of the subjectivity of manual evaluation, we use ROUGE (Lin, 2004), which is a famous text evaluation tool and regarded as the gold criteria for the evaluation of automatic summarization, to automatically score a machine summary. ROUGE compares the number of overlapping cells (e.g., word, sequence etc.) simultaneously appearing in a machine summary and a manual summary to evaluate the machine summary quality.

(2) Candidate algorithms: In the experiments, we used the following six algorithm candidates.

- **PSO**. It is a dynamic programming algorithm (Poli et al., 2007; Aliguliyev, 2010) developed based on the ROUGE criteria together with manual summaries. It can obtain the optimal machine summary in theory for a text document, so it is used as the upper limit of the summary evaluation. However, it has an obvious shortcoming, i.e., a manual summary of each novel has to be provided in advance. In the experiment, the parameter values are from the recommendation of (Aliguliyev, 2010).

- **TextRank**. It is a graph model based algorithm (Mihalcea and Tarau, 2004), where a topic is scored and recommended by analyzing the relation between texts, so each topic can be recommended by its adjacent topics, i.e., the score of each topic is calculated by the repeated iteration of its adjacent topics. In the experiment, we set the related parameters according to the recommendation from (Leite et al., 2007).

- **LSA**. Its basic idea to use Singular Value Decomposition (SVD) to mine the implication relation between sentences and terms (Deerwester et al., 1990), so as to extract the topics contained in a text document. In previous studies, LSA is generally used to deal with the text documents with short



(a) summary quality (b) topic diversity

Figure 4: Evaluation Result for the Optimal Parameters

length (Kireyev, 2008; Ozsoy et al., 2011). In the experiment, the parameter on the topic number is set to 10 so as to be consistent with LDA.

- **HDP**. It is a nonparametric topic modeling algorithm, i.e., compared to LSA and LDA, it does not require to estimate the number of topics in a text document in advance (Teh et al., 2006). In previous studies, it is also mainly used to deal with the text documents (such as news) with short length (Li et al., 2012; Li and Li, 2013). We set the related parameters according to the recommendation from (Wang et al., 2011).
- **Random**. It randomly chooses novel sentences to form a machine summary, where the length of each random summary is set to be equal to its corresponding manual summary. In our experiment, it is used as the lower limit of the summary evaluation.
- **Our Algorithm**, i.e., the algorithm proposed in this paper. From Formulas 6-9, we know that our algorithm contains two parameters θ_1 and θ_2. To determine the optimal values for θ_1 and θ_2, we performed grid search over the range $(1, 5) \times (1, 5)$, by using the real novel dataset given in Section 4.1 as input, and the summary quality (ROUGE-1) and topic diversity as evaluation indicators. The results are shown in Fig. 4, which show that when $\theta_1 = 2$ and $\theta_2 = 1$ (after rounded), the summary quality and topic diversity indicators both have the best performance.

Note that ESA is a also well-known approach (Evgeniy and Shaul, 2007) that can be used to topic modeling. In ESA, each text document is represented as a vector in a high-dimensional space of concepts derived from Wikipedia. However, the immense concept space leads to the worse efficiency of the approach, thereby making it too time-consuming to run over the novel dataset. Besides, although there are a number of multi-document summarization algorithms (such as (Oskar et al., 2014) and (Baralis et al., 2015)) that can also be extended to summarize novel texts, most of the algorithms are not open source, thereby, making it difficult to compare them with our approach by experiments. Finally, for the novel dataset, the machine summaries generated by all the algorithm candidates have been published to the Google network disk[8].

5.2. Topic diversity evaluation

In the first group of experiments, we aim to evaluate the effectiveness of automatic novel summaries generated by our approach in terms of topic diversity. The topic diversity is an important metric that reflects the quality of the generated machine summaries, and the higher the metric value, the better the quality

[8]https://drive.google.com/file/d/0B26lw2I2tnxCeW5mSzlnalVTa1E/view

Table 2: Topic diversity paired t-test results for statistically significant testing

	Ours vs. Rondom	Ours vs. PSO	Ours vs. TextRank	Ours vs. LSA	Ours vs. HDP
P-value	4.32×10^{-10}	3.39×10^{-2}	3.59×10^{-3}	7.58×10^{-8}	8.77×10^{-12}

[1] Null hypothesis (H0): There is no difference between the two models.
[2] Alternative hypothesis (H1): The first model outperforms the second model.

of the summaries (Alguliev et al., 2012). Based on Definition 3, we define topic distribution similarity to measure the topic diversity.

Metric 1 (Topic distribution similarity). For a candidate algorithm A and a novels set $\mathbb{S}$, let $\mathbb{S}^a$ denote an automatic summary set determined by the algorithm A for the novel set $\mathbb{S}$. Then, the topic distribution similarity of the automatic summaries generated by A for $\mathbb{S}$ can be measured as follows:

$$\text{TMAX}(A,\mathbb{S}) = \max_{\mathrm{S}^a \in \mathbb{S}^a} \textbf{diver}(\mathrm{S}^a);\ \text{TAVE}(A,\mathbb{S}) \quad = \frac{1}{|\mathbb{S}^a|} \sum_{\mathrm{S}^a \in \mathbb{S}^a} \textbf{diver}(\mathrm{S}^a);\ \text{TMIN}(A,\mathbb{S}) = \min_{\mathrm{S}^a \in \mathbb{S}^a} \textbf{diver}(\mathrm{S}^a)$$

In the experiment, the length of the automatic summary of each novel is set to 500, i.e., the compression ratio of each summary is set to about $0.1\% - 0.2\%$. The experimental results are shown in Fig. 5. From the experimental results, we have the following several observations. First, the Random algorithm as the baseline has the worst topic distribution similarity, whose maximum, average and minimum values are equal to 0.57, 0.27 and 0.08, respectively, all lower than those from the other five candidate algorithms. Second, our proposed approach has the best topic distribution similarity: the maximum, minimum and average topic distribution similarity values are equal to 0.24, 0.42 and 0.63, respectively, which are obviously better than those from TextRank, HDP, LSA and Random, and slightly better than those from the PSO algorithm. Third, based on the comparison among the maximum, minimum and average values of topic distribution similarity, it can be seen that our approach has better stability, i.e., for the three topic distribution similarity measures, their values are not much different from each other (compared to other five candidates).

In addition, the paired t-tests for statistical significance (Wu et al., 2012) are performed to verify whether the improvements on topic diversity of our proposed approach over other five candidates are statistically significant or not. The results are shown in Table 2, where "P-value" denotes the percentage value of our approach versus another candidate (Random, PSO, TextRank, LSA or HDP). From the results, we can see that the topic diversity improvements of our approach over other candidates are statistically significant (with a confidence le

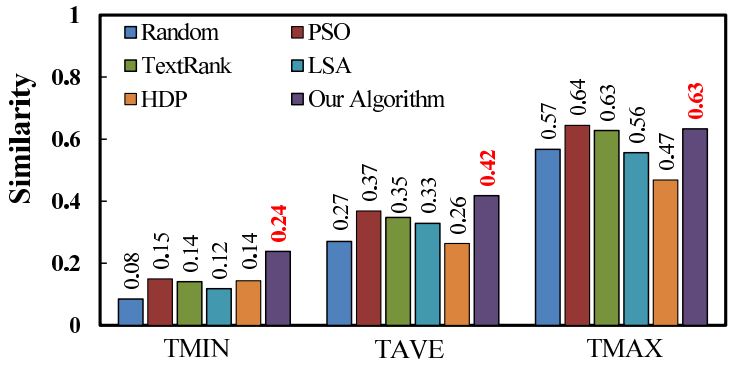


Figure 5: Evaluation Result on Summary Topic Diversity

From the above experiments, we conclude that under the precondition of ensuring a high compression ratio (about 0.1% to 0.2%), our proposed approach can effectively ensure the topic diversity of the generated machine summaries, and hence the quality of the generated machine summaries.

5.3. Actual quality evaluation

In the second group of experiments, we aim to evaluate the actual quality of the machine summaries generated by our approach. First, we use the ROUGE evaluation criteria combined with the manual summaries to conduct the evaluation. Here, we use three evaluation factors commonly used in information retrieval, i.e., Recall, Precision and F-score.

Metric 2 (ROUGE quality). For a candidate algorithm A, and a novels set $\mathbb{S}$ and its corresponding manual summary set $\mathbb{S}^m$, let $\mathbb{S}^a$ denote an automatic summary set generated by the algorithm A for the novel set $\mathbb{S}$, and let $\mathrm{S}_k^a \in \mathbb{S}^a$ and $\mathrm{S}_k^m \in \mathbb{S}^m$ respectively denote the machine summary and manual summary corresponding to a novel document $\mathrm{S}_k \in \mathbb{S}$. Then, the practical quality of the automatic summaries generated by A for $\mathbb{S}$ can be measured as follows:

$$\mathrm{Precision}(A,\mathbb{S}) = \frac{1}{|\mathbb{S}|}\sum_{\mathrm{S}_k\in\mathbb{S}}\frac{|\mathrm{S}_k^a \cap \mathrm{S}_k^m|}{|\mathrm{S}_k^a|}$$

$$\mathrm{Recall}(A,\mathbb{S}) = \frac{1}{|\mathbb{S}|}\sum_{\mathrm{S}_k\in\mathbb{S}}\frac{|\mathrm{S}_k^a \cap \mathrm{S}_k^m|}{|\mathrm{S}_k^m|}$$

$$\mathrm{FScore}(A,\mathbb{S}) = 2\frac{\mathrm{Precision}(A,\mathbb{S})\cdot\mathrm{Recall}(A,\mathbb{S})}{\mathrm{Precision}(A,\mathbb{S})+\mathrm{Recall}(A,\mathbb{S})}$$

Obviously, the greater the values of the three factors, the better the actual quality of the machine summaries, where due to the comprehensive consideration of Precision and Recall, FScore is considered as the most important factor. Here, we adopt three commonly used ROUGE evaluation standards, i.e., ROUGE-1, ROUGE-2 and ROUGE-SU4.

In addition, we also evaluate the quality of the automatic summaries by using a manual approach. Specifically, we invite a group of undergraduate students, each of whom had sufficient judgment ability to conduct the evaluation, to act as assessors to score the summaries based on the relevance of each summary to its novel text. Each summary is first scored by assessors independently (a score between 0 and 1), and then we average the scores given by six assessors for each summary to determine the final score of the summary. In the experiment, in order to reduce the workload of the assessors, we only choose six novels from the novel dataset.

Metric 3 (Manual quality). Given a candidate algorithm A and a manual summary set $\mathbb{S}^m$, let $\mathbb{S}^a$ denote an automatic summary set corresponding to $\mathbb{S}^m$, generated by A, let $\mathrm{S}_k^a \in \mathbb{S}^a$ and $\mathrm{S}_k^m \in \mathbb{S}^m$ respectively denote a machine summary and its corresponding manual summary, and let $\mathbf{score}(\mathrm{S}^a)$ denote a manually determined score for S^a. Then, the quality of the automatic summaries generated by A can be measured as follows:

$$\mathrm{EMAX}(A,\mathbb{S}) = \max_{\mathrm{S}^a\in\mathbb{S}^a}\frac{\mathbf{score}(\mathrm{S}^a)}{\mathbf{score}(\mathrm{S}^m)};\ \mathrm{EAVE}(A,\mathbb{S}) = \frac{1}{|\mathrm{S}^a|}\sum_{\mathrm{S}^a\in\mathbb{S}^a}\frac{\mathbf{score}(\mathrm{S}^a)}{\mathbf{score}(\mathrm{S}^m)};\ \mathrm{EMIN}(A,\mathbb{S}) = \min_{\mathrm{S}^a\in\mathbb{S}^a}\frac{\mathbf{score}(\mathrm{S}^a)}{\mathbf{score}(\mathrm{S}^m)}$$

The experimental results are shown in Fig. 6, where the subfigures (a), (c), (e) and (g) are the evaluation results before summary smoothing, and the subfigures (b), (d), (f) and (h) are the evaluation results after summary smoothing. From the experimental results in Fig. 6, we have the following several observations. First, the Random algorithm as the baseline has the worst performance, i.e., its precision, recall rate, F-score and manual score are all lower than those from other five candidates, before or after the summary smoothing operation. Second, compared to Random, LSA, HDP and TextRank, our approach can greatly improve the actual effectiveness of automatic summarization. Specifically, for the evaluation standards ROUGE-1, ROUGE-2 and ROUGE-SU4, as well as the manual standard, the machine summaries generated by our approach are all significantly better than those from the Random algorithm, slightly better than those from TextRank, LSA and HDP. Third, compared to the PSO algorithm as the upper limit of automatic summarization effectiveness, the machine summaries generated by our approach have similar quality, where the recall rate and manual score are slightly worse, the precision is slightly better, and the overall F-score is

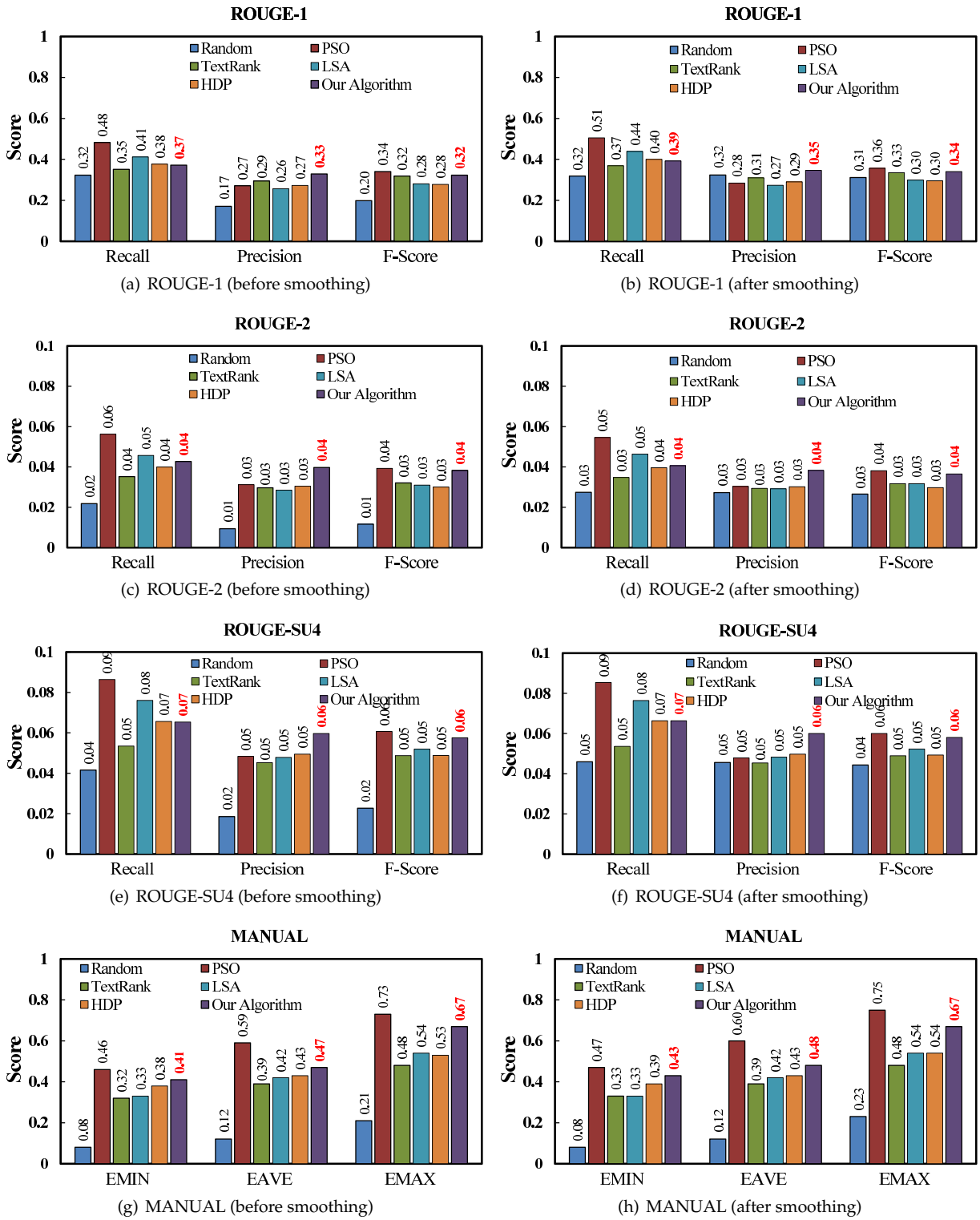


Figure 6: Evaluation Result on Summary Quality

Table 3: Effectiveness paired t-test results for statistically significant testing

	Ours vs. Rondom	PSO vs. Ours	Ours vs. TextRank	Ours vs. LSA	Ours vs. HDP
P-value (ROUGE-1)	7.0×10^{-3}	7.3×10^{-2}	4.42×10^{-2}	1.26×10^{-8}	1.92×10^{-6}
P-value (ROUGE-2)	3.64×10^{-7}	3.63×10^{-2}	2.83×10^{-3}	1.93×10^{-3}	1.43×10^{-4}
P-value (ROUGE-SU4)	2.57×10^{-6}	8.82×10^{-2}	1.97×10^{-3}	5.46×10^{-4}	2.12×10^{-4}
P-value (MANUAL)	2.72×10^{-9}	2.42×10^{-8}	2.64×10^{-6}	8.98×10^{-6}	2.01×10^{-4}

[1] Null hypothesis (H0): There is no difference between the two models.
[2] Alternative hypothesis (H1): The first model outperforms the second model.

basically similar. Fourth, by comparing the subfigures (a), (c), (e) and (g) with the other subfigures (b), (d), (f) and (g), we observe that the summary smoothing operation can improve the quality of the machine summaries generated by the candidates.

In addition, we also perform the paired t-tests for statistical significance to verify whether the improvements on the summary quality of our proposed approach over other candidates are statistically significant or not. The testing results are shown in Table 3. From the results, we can see that: on the one hand, the summary quality improvements of our approach over Random, TextRank, LSA and HDP are statistically significant (with a confidence level of greater than 95%); and on the other hand, although PSO (as the upper limit of the summary evaluation) can obtain the optimal machine summary for each novel in theory, its summary quality improvements over our approach is not statistically significant.

From all the above experiment results, we conclude that under the precondition of a high compression ratio (about 0.1% to 0.2%), our approach can generate approximately the optimal machine summaries (compared to the PSO algorithm), and thus ensure the actual quality of the machine summaries, i.e., ensuring the effectiveness of automatic summarization. In summary, compared to existing approaches, our approach is designed specifically for novel documents of structure freedom and long length, which uses the LDA algorithm to capture the topic words associated with a novel, enabling the generated summary to better reflect the complex context of a novel; and then uses some heuristic rules that are developed based on the stylistic feature, topic diversity and redundancy rate of a novel sentence, to select the most important candidate sentences, enabling the generated summary to obtain a higher compression ratio.

6. Conclusion and future work

In this paper, we proposed an extractive summarization approach for novel documents. The approach was developed based on the LDA topic modeling algorithm, where under the requirements of high compression ratio and topic diversity, the importance evaluation function of candidate sentences was designed to extract a machine summary for a novel document. In addition, the approach also smoothed each machine summary so as to improve the summary readability. Finally, we conducted experiments on a real dataset to evaluate the effectiveness of the approach. The experimental results show that our approach can ensure the topic diversity of a machine summary, under the precondition of a high compression ratio (0.1% to 0.2%).

As the future work, we will try to further study the following problems, i.e., (1) how to extract semantic entities (such as novel characters) and then redesign the sentence importance evaluation function based on the novel context; (2) how to improve the sentence fusion by syntactic and contextual relationships, so as to reduce the sentence overlap information, and thus improve the summary readability; and (3) how to further improve topic modeling by the narrative study and stylistic aspects of knowledge.

Acknowledgment

The work of this paper is supported by the Zhejiang Provincial Natural Science Foundation of China (LY15F020020), the Jiangxi Provincial Natural Science Foundation of China (20161BAB202036), the Wenzhou Science and Technology Program (G20160006 and Y20160070) and the National Natural Science Foundation of China (61402337 and 61572367).

References

Alguliev, R., Aliguliyev, R., Isazade, N. R., 2012. Desamc+ docsum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization. Knowledge-Based Systems 36, 21–38.

Alguliev, R., Aliguliyev, R. M., Isazade, N. R., 2013. Multiple documents summarization based on evolutionary optimization algorithm. Expert Systems with Applications 40 (5), 1675–1689.

Aliguliyev, R. M., 2010. Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization. Computational Intelligence 26 (4), 420–448.

Bairi, R., Iyer, R., Ramakrishnan, G., Bilmes, J., 2015. Summarization of multi-document topic hierarchies using submodular mixtures. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. pp. 553–563.

Bamman, D., Smith, N. A., 2013. New alignment methods for discriminative book summarization. arXiv preprint arXiv:1305.1319.

Baralis, E., Cagliero, L., Fiori, A., Garza, P., 2015. Mwi-sum: A multilingual summarizer based on frequent weighted itemsets. ACM Transactions on Information Systems 34 (1), 5.

Bird, S., Klein, E., Loper, E., 2009. Natural language processing with Python. O'Reilly Media, Inc.

Blei, D., Ng, A., Jordan, M., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3 (Jan), 993–1022.

Ceylan, H., 2011. Investigating the extractive summarization of literary novels. Ph.D. thesis.

Ceylan, H., Rada, M., 2007. Explorations in automatic book summarization. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Association for Computational Linguistics, pp. 280–389.

Chi, L., Li, B., Zhu, X., 2014. Context-preserving hashing for fast text. In: Proc. of SDM. pp. 100–108.

Das, D., Martins, A. T., 2007. A survey on automatic text summarization. Literature Survey for the Language and Statistics II course at CMU 4, 192–195.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41 (6), 391–407.

Evgeniy, G., Shaul, M., 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proc. of IJCAI. pp. 1606–1611.

Fernandez-Amoros, D., Heradio, R., 2011. Understanding the role of conceptual relations in word sense disambiguation. Expert Systems with Applications 38 (8), 9506–9516.

Ferreira, R., de Souza Cabral, L., Freitas, F., Lins, R. D., Franya Silva, G., Simske, S. J., Favaro, L., 2014. A multi-document summarization system based on statistics and linguistic treatment. Expert Systems with Applications 41 (13), 5780–5787.

Gambhir, M., Gupta, V., 2016. Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 1–66.

Jarmasz, M., Szpakowicz, S., 2003. Roget's thesaurus and semantic similarity. In: Proceedings of the International Conference on Recent Advances in Natural Language Processin(RANLP). pp. 212–219.

Kazantseva, A., Szpakowicz, S., 2010. Summarizing short stories. Computational Linguistics 36 (1), 71–109.

Kireyev, K., 2008. Using latent semantic analysis for extractive summarization. Analysis.

Leite, D. S., Rino, L., Pardo, T., Nunes, M., 2007. Extractive automatic summarization: Does more linguistic knowledge make a difference? In: Proceedings of the TextGraphs-2 HLT/NAACL Workshop. p. 17.

Li, J., Li, S., 2013. Evolutionary hierarchical dirichlet process for timeline summarization. In: Meeting of the Association for Computational Linguistics. pp. 556–560.

Li, J., Li, S., Wang, X., Tian, Y., Chang, B., 2012. Update summarization using a multi-level hierarchical dirichlet process model. In: Proceedings of COLING. pp. 1603–1618.

Lin, C.-Y., 2004. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out: Proceedings of the ACL-04 workshop. Vol. 8. Barcelona, Spain, pp. 74–81.

Lloret, E., Palomar, M., 2013. Towards automatic tweet generation: A comparative study from the text summarization perspective in the journalism genre. Expert Systems with Applications 40 (16), 6624–6630.

Mihalcea, R., Tarau, P., 2004. Textrank: Bringing order into texts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP). Association for Computational Linguistics.

Miller, G. A., Leacock, C., Tengi, E., Bunker, R. T., 1993. A semantic concordance. In: Proceedings of the Workshop on Human Language Technology. Association for Computational Linguistics, pp. 303–308.

Navigli, R., 2009. Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41 (2), 10.

Oskar, G., Antoine, D., Hannu, T., 2014. Document summarization based on word associations. In: Proc. of SIGIR. pp. 1023–1026.

Ozsoy, M. G., Alpaslan, F. N., Cicekli, I., 2011. Text summarization using latent semantic analysis. Journal of Information Science 37 (4), 405–417.

Poli, R., Kennedy, J., Blackwell, T., 2007. Particle swarm optimization: An overview. Swarm intelligence 1 (1), 33–57.

Rehurek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Citeseer, pp. 45–50.

Riddell, A., 2013. Demography of literary form: Probabilistic models for literary history. Ph.D. thesis, Duke University.

Sinha, R., Mihalcea, R., 2009. Combining lexical resources for contextual synonym expansion. In: Proceedings of the International Conference on Recent Advances in Natural Language Processin(RANLP). pp. 404–410.

Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M., 2006. Hierarchical dirichlet processes. Journal of the American Statistical Association 101 (476), 1566–1581.

Tran, G., Herder, E., Markert, K., 2015. Joint graphical models for date selection in timeline summarization. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Vol. 1. Association for Computational Linguistics, pp. 1598–1607.

Wang, C., Jing, F., Zhang, L., Zhang, H., 2007. Learning query-biased web page summarization. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, pp. 555–562.

Wang, C., Paisley, J. W., Blei, D. M., 2011. Online variational inference for the hierarchical dirichlet process. Journal of Machine Learning Research 15, 752–760.

Wu, Z., Xu, G., Zhang, Y., Peter, D., Chenglang, L., 2012. An improved contextual advertising matching approach based on wikipedia knowledge. The Computer Journal 55 (3), 277–293.

Xiong, W., Litman, D., 2014. Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 1985–1995.

Yang, G., Wen, D., Chen, N. S., Sutinen, E., 2015. A novel contextual topic model for multi-document summarization. Expert Systems with Applications 42 (3), 1340–1352.

Yang, L., Cai, X., Zhang, Y., Shi, P., 2014. Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization. Information Sciences 260, 37–50.

Yuan, J., Sivrikaya, F., Hopfgartner, F., Lommatzsch, A., Mu, M., 2015. Context-aware lda: Balancing relevance and diversity in tv content recommenders. In: Proceedings of the 2nd Workshop on Recommendation Systems for Television and Online Video.