Multi-Document Concept Graph Construction Based on Open-Domain Extraction: Postprint

Authors: Sheng Yongpan, Fu Xuefeng, Wu Tianxing

Date: 2018-11-29T00:00:00+00:00

Abstract

In the context of information overload, how to mine and organize core concepts and their semantic connections from multiple documents sharing a common theme has become an important challenge in current open information extraction tasks. To this end, we propose a multi-document concept graph construction model based on open-domain extraction. First, topic words are mined based on predetermined themes, and documents are ranked through an improved TF-IDF algorithm; then, through a series of methods including coreference resolution, discourse weight calculation, and open-domain extraction, a large number of triple instances with factual expressive capability are extracted from multiple articles. To remove the noise inherent in open-domain methods and improve the accuracy of information extraction, a fact filtering algorithm is proposed. Through this algorithm, a set of salient factual knowledge with high confidence and good semantic compatibility can be effectively extracted, forming multiple concept subgraphs. Finally, equivalent concepts and relations in different subgraphs are merged to form a connected concept graph with thematic expressive capability. Validated on the Signal Media news dataset, experimental results demonstrate that the proposed model can effectively mine and organize key information related to specific themes across documents, and the constructed concept graph achieves favorable performance on metrics such as thematic concept coverage and compatibility of factual knowledge. Additionally, this model also holds important reference value for applications in automatic document summarization.

Full Text

Preamble

Multi-document Conceptual Graph Construction Research Based on Open Domain Extraction

Sheng Yongpan[1], Fu Xuefeng[2], Wu Tianxing[3]
(1. School of Computer Science & Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China;
2. School of Information Engineering, Nanchang Institute of Technology, Nanchang 330099, China;
3. School of Computer Science & Engineering, Southeast University, Nanjing 211189, China)

Abstract: In the context of information overload, mining and organizing meaningful concepts and their semantic connections from a set of related documents under the same topic has become a significant challenge in open information extraction tasks. This paper proposes a multi-document conceptual graph model based on open-domain information extraction. Firstly, documents are ranked according to the improved TF-IDF weight of extracted topic words under predefined topics. Then, the model relies on a series of methods, including coreference resolution, weight computation, and open-domain information extraction, to extract numerous representative subject-predicate-object triples from multiple documents. To filter out the noise inherent in open-domain information approaches and improve extraction accuracy, this paper presents a fact filtering algorithm to retain only the most salient and compatible facts, forming multiple conceptual subgraphs. Finally, by merging equivalent concepts and relationships across different subgraphs, a fully connected conceptual graph with expressive topic ability is constructed. Experiments on the Signal Media dataset illustrate that the proposed model can discern and effectively group key information corresponding to specific topics within and across documents. The constructed conceptual graph outperforms state-of-the-art algorithms in terms of topic concept coverage rate and compatible facts. Additionally, this model holds significant importance for automatic document summarization applications.

Key words: open-domain extraction; multiple documents; conceptual graph construction

0 Introduction

With the continuous evolution of the big data era, information published through media channels such as newspapers, radio, television, the Internet, Weibo, and WeChat, along with user-generated content, has grown exponentially. Free text data, as a typical representative, has played a crucial role in revealing information essence and semantic relationships. However, obtaining important topic concepts and semantic associations from large-scale, scattered text collections has become a critical challenge in information extraction tasks.

Literature [1] proposed multi-document summarization technology aimed at extracting the main ideas from multiple articles sharing a common topic to form concise, readable summaries that are easy for users to understand. Literature [2] introduced an automated multi-document summarization extraction method based on the LexRank algorithm, which serves as a typical representative of

extractive summarization. Subsequently, Literature [3] utilized neural network models to model central sentences in text, achieving further compression of summary statements. Literature [4] proposed the Latent Dirichlet Allocation (LDA) topic model, which can mine latent topic information from large-scale text datasets but requires manual specification of the number of topics. Literature [5] introduced the Hierarchical Latent Dirichlet Allocation (HLDA) topic model to address this limitation. Literature [6] designed an automatic text summarization extraction system by fully utilizing features such as word frequency, topic sentence position, and topic words. Literature [7] proposed a review-based summarization approach that preserves reviewers' overall opinions while reflecting review diversity. However, these summarization methods generally perform modestly in terms of topic concept coverage (such as integrating scattered topic details) and accuracy. Selecting preferred topic concepts and associated information to generate fluent and meaningful multi-document summaries has long been a key research problem in information extraction.

Literature [8] proposed open-domain extraction methods based on syntactic dependency analysis, which can adapt to open information extraction tasks for unlabeled, unrestricted domain large-scale text. The University of Washington has developed a series of milestone binary OIE (Open Information Extraction) systems in the field of information extraction, including KnowItAll, TextRunner, WOE, ReVerb, and R2A2 [9]. Their primary advantage lies in efficiently extracting binary shallow entity relationships from corpora while considering global contextual information. However, because open-domain methods mainly rely on open templates and syntactic dependency relationships, they cannot effectively identify whether extracted facts accurately express sentence meanings and struggle to effectively connect these factual knowledge across sentences and documents. Literature [10] employed a logical constraint method to achieve cross-document organization of factual knowledge, but since the rules were limited to finite relationship application scenarios, they could not ensure that the connected facts were meaningful or covered important topic information.

Another typical approach in open information extraction is open entity relation extraction, which is based on the assumption that if a specified semantic relationship exists between known entities, all sentences containing these two entities implicitly express this relationship. Open entity relation extraction methods [11] mainly rely on external domain-independent knowledge bases (such as DBPedia, Freebase, YAGO, and Wikipedia text corpora) to map high-quality entity relationships to knowledge base corpora, then use text alignment methods to obtain relationship extraction training data (this process can also be viewed as data annotation), and train models to implement relation extraction tasks. However, these methods have two main problems: (a) training corpora contain considerable noise; (b) annotated entity relationship types are limited. For the former problem, distant supervision extraction methods have received widespread attention from industry experts since their proposal and have achieved good performance. Literature [12] addressed errors and error propagation in feature extraction processes of traditional statistical models, as well as deep learning

limitations, by proposing an entity relation extraction method combining convolutional neural networks with keyword strategies, which proved beneficial for improving extraction accuracy. Literature [13] addressed data annotation errors by using multi-instance learning to extract high-confidence training examples from training sets to train models, which improved algorithm performance to some extent. For the latter problem, more recent entity relation extraction methods [14] attempt to face large-scale open corpora, where the included relationship types will be more comprehensive.

To address the problems of difficulty in organizing factual knowledge information across sentences and documents in open information extraction tasks, and the limited types of annotated entity relationships, this paper proposes a multi-document conceptual graph construction model based on open-domain extraction. This model relies on a series of NLP (natural language processing) methods and tools to represent significant entities, concepts, and their relationships under specific topics in the form of conceptual graphs, achieving the goal of cross-document mining and organizing key topic information, which holds important reference value for further research on topic development trends and automatic document summarization applications.

1 Multi-document Conceptual Graph Construction Model

Constructing a conceptual graph model based on multi-document semantic linking mainly includes four primary tasks: document ranking, concept and relation extraction, fact filtering, and merging equivalent concepts and relations, which are detailed below.

1.1 Document Ranking

Based on predefined topics, the Stanford CoreNLP system [15] is used to mine named entities, gerunds, nouns, and event names from documents as candidate keywords. An improved TF-IDF algorithm calculates their importance to the topic. Compared with traditional TF-IDF algorithms, this improved version not only reduces the probability of misidentifying rare words as topic words but also considers the distribution of keywords across different topics. The calculation formula is:

$$TF\text{-}IDF(w,k|D) = tf(w,k|D) \times idf(w,k|D)$$

where $tf(w,k|D)$ represents term frequency, measuring the importance of keyword w in all documents of topic k; $idf(w,k|D)$ represents inverse document frequency, measuring the general degree of keyword w across all documents, not limited to a specific topic.

$$tf(w,k|D) = \frac{c(w)}{|D_k|} = \frac{c(w)}{\sum_{w \in D_k} c(w)}$$

$$idf(w,k|D) = \log\left(1 + \frac{\sum_{k^* \neq k} f(w,k^*)}{f(w,k)}\right)$$

where: w is a candidate keyword; k is a specific topic; D_k represents the document set under topic k; $|D_k|$ represents the total number of words in D_k; $c(w)$ represents the occurrence count of w in D_k; $f(w,k)$ represents the frequency of w in D_k; k^* represents topics other than k; $f(w,k^*)$ represents the frequency of w in k^*.

The document weight calculation formula is:

$$weight(d) = \sum_{i=1}^{key_n(d)} TF\text{-}IDF(w_i)$$

where: d is a document under topic k; $key_n(d)$ represents the total number of keywords contained in d; $TF\text{-}IDF(w_i)$ represents the TF-IDF value of the i-th keyword in d, which can be calculated according to Equation (1).

1.2 Concept and Relation Extraction

The main task of concept and relation extraction is to extract a large number of triple instances with factual expressive ability from multiple documents under the same topic, mainly including three subtasks: coreference resolution, passage weight calculation, and open-domain extraction.

1) **Coreference Resolution**: Reference types in the same article mainly manifest as personal pronouns, demonstrative pronouns, noun phrase references, and event references. This paper uses the Stanford CoreNLP system, a natural language processing toolkit developed by Stanford University, to replace coreferential pronouns in single documents to improve readability for subsequent open-domain extraction tasks.

2) **Passage Weight Calculation**: The TextRank algorithm [16], derived from Google' s famous PageRank algorithm, segments text into semantic units and builds a graph model to rank important components in text through a voting mechanism, which can be used for keyword extraction and automatic summarization of single documents. This paper adopts the TextRank algorithm to calculate scores of different sentences in documents and uses high-scoring sentences as topic sentences.

3) **Triple Instance Extraction**: Traditional information extraction patterns require limited domains and semantic unit types and cannot be applied to free text corpora without predefined concept relationship types. Therefore, the OLLIE (Open Language Learning for Information Extraction) system [9] developed by the University of Washington is used to extract binary relationships from document topic sentences. Extracted

triple instances can be represented as (subject, predicate, object), where subject and object represent two entities or concepts without nested structures, and predicate represents their relationship, mainly consisting of verbs and verb phrases without nested relationships or modifier phrases. For complex long sentences, the OLLIE system extracts one or more relationship pairs with different confidence levels. Three example sentences are provided for illustration.

Example 1: "82 percent of leaned Democrats say Registered voters' d support a clear Clinton, while 76 percent of leaned Republicans say Registered voters' d back a clear Clinton vs. Trump, were Registered voters the party nominees."

F1: 0.97 (76 percent of leaned Republicans; say; Registered voters' d back a clear Clinton vs. Trump)
F2: 0.94 (82 percent of leaned Democrats; say; Registered voters' d support a clear Clinton)

In Example 1, the OLLIE system extracts triple instance F1 with confidence 0.97 and F2 with confidence 0.94. Higher confidence indicates more accurate factual knowledge expressed by the triple instance.

Example 2: "That compares to a clear Clinton lead among all adults, 51-39 percent, indicating her broad support in groups that are less apt to be registered to vote, such as young adults and racial and ethnic minorities."

F3: 0.45 (That; compares; to a clear Clinton lead among all adults)
F4: 0.90 (51-39 percent; indicating; her broad support in groups)
F5: 0.67 (groups; are; less apt to be registered to vote)

Example 3: "The hypothetical contest compares to a clear Clinton lead among all adults, 51-39 percent, indicating Hillary Clinton broad support in groups that are less apt to be registered to vote, such as young adults and racial and ethnic minorities."

F6: 0.95 (The hypothetical contest; compares; to a clear Clinton lead among all adults)
F7: 0.90 (51-39 percent; indicating; Hillary Clinton broad support in groups)
F8: 0.92 (groups; are; less apt to be registered to vote)

Example 3 shows the result of coreference resolution on Example 2. The underlined "That" is replaced with "The hypothetical contest" and "her" with "Hillary Clinton" . Consequently, the triple instance corresponding to "That" changes from F3 to F6, with confidence increasing from 0.45 to 0.95; the triple instance corresponding to "her" (F4) increases in confidence from 0.67 to 0.92 after coreference resolution.

1.3 Fact Filtering

The OLLIE system is susceptible to dependency parsing errors, generating uninformative or incorrect triple instances. Meanwhile, semantically repetitive

sentences across multiple documents produce a certain proportion of redundant triple instances. To address these issues, this paper proposes a fact filtering algorithm to filter out candidate triple instances unrelated to core topic content and with low confidence, retaining only high-confidence, semantically compatible salient factual knowledge. The algorithm transforms the triple instance filtering problem into an integer programming problem with the objective function and constraints as follows:

Let $T = \{t_1, t_2, ..., t_M\}$ be a set of M triple instances; t_i, t_j represent any two triple instances in the set; x_i is an indicator variable for t_i, where $x_i = 1$ if t_i is retained; y_{ij} is also an indicator variable representing compatibility between t_i and t_j, where if $y_{ij} = 1$, then $x_i = 1$, $x_j = 1$, and both t_i and t_j are retained; c_i represents the confidence of fact expressed by t_i; n_{max} is the number of triple instances in the conceptual graph, which can be set by users, and the generated triple instance set will contain no more than n_{max} instances.

$$k_{sim}(t_i, t_j) = \alpha \cdot s(t_i, t_j) + \beta \cdot l(t_i, t_j)$$

where: $s(t_i, t_j)$ represents semantic relevance between t_i and t_j, mainly calculated through the ADW (Align, Disambiguate and Walk) model [17]; $l(t_i, t_j)$ represents literal similarity between t_i and t_j, mainly calculated through the Levenshtein distance formula; α is a proportional coefficient representing the proportion of $s(t_i, t_j)$ in k_{sim} calculation; β represents the proportion of $l(t_i, t_j)$, with $\alpha + \beta = 1$.

To reduce computational load, the method introduces a sliding window mechanism. As the sliding window moves, only unrepeated triple instances within the window are compared each time, reducing computational complexity from $O(M^2)$ to $O(W \cdot M)$, where W is the window size and Δ is the sliding step size.

1.4 Merging Equivalent Concepts and Relations

The current challenge lies in the diversity of concept mentions and potential noise in concept relationship descriptions. Therefore, this paper proposes the following rules to merge equivalent concepts and relations across multiple conceptual subgraphs.

Rule 1: Synonymous concepts have equivalence. Synonymous concepts have obvious lexical structure features, such as "Billionaire Donald Trump", "Donald Trump", "Donald John Trump", and "Trump" all referring to the same person. For named entities, search engines' powerful entity linking capabilities can be used to check whether they can be accurately linked to the same referent object.

Rule 2: Similar concepts have equivalence. This mainly adopts the ADW model [14], which relies on the WordNet dictionary. By performing random walks, semantic fingerprints corresponding to concepts can be obtained, and similarity between two concept fingerprints can be calculated through Cosine, Weighted Overlap, and Top-k Jaccard methods.

Rule 3: Semantically overlapping concepts have equivalence. This mainly adopts the semantic overlap calculation formula proposed in Literature [18]. This measurement method relies on WordNet's taxonomy structure and converts the path length from two concepts to the root node into information content for calculation.

2 Experimental Analysis

2.1 Experimental Data

The Signal Media news dataset collects news reports published by Reuters during September-October 2015 (http://research.signalmedia.co/newsir16/signal-dataset.htm), totaling 1,000,000 English documents, including 734,488 news articles and 265,512 blog posts, with each document averaging 39 sentences and 1,266 words. In the DUC (Document Understanding Conference) standard corpus [19], the same topic contains approximately 25-40 documents. This experiment randomly selected 10,000 documents as research corpora, including 734 news articles (73.4%) and 266 blog posts (26.6%). The corpora are divided into five topics: Syria refugee crisis, Iran nuclear, Volkswagen scandal, United States presidential election, and Sino-Soviet cooperation. According to Equation (4) in Section 1.1, the top 100 documents are selected for each topic, and documents are randomly chosen for analysis to generate datasets of sizes: 5, 15, 25, 35, 45, 55, 65, and 75 (unit: documents), named $D_{i1}, D_{i2}, D_{i3}, D_{i4}, D_{i5}, D_{i6}, D_{i7}, D_{i8}$ (i represents a specified topic, $15 \leq i \leq N_i + 1$). The specific details of the experimental datasets are shown in Table 1 .

2.2 Evaluation Metrics

The following evaluation metrics are used for comprehensive analysis of the proposed conceptual graph model from five aspects: topic concept coverage, conceptual graph connectivity, conceptual graph readability, model running time, and comparative algorithms.

a) Topic Concept Coverage Rate: Represents the percentage of correctly extracted topic concepts in the conceptual graph, calculated as:

$$C_{theme} = \frac{n_{concept}}{n_{theme}}$$

where $n_{concept}$ is the total number of concepts in the conceptual graph; n_{theme} is the number of topic concepts calculated through the random forest algorithm.

b) Kappa Coefficient: Used for consistency testing of annotation results, calculated as:

$$K = \frac{P_0 - P_e}{1 - P_e}$$

where P_0 and P_e are the observed agreement rate and chance agreement rate of different annotation results, respectively; $1 - P_e$ represents the non-chance agreement rate.

c) ROUGE Evaluation Standards: A recall-based similarity measurement method, mainly including ROUGE-N, ROUGE-L, and ROUGE-S metrics.

ROUGE-N represents N-gram co-occurrence statistics, with precision $P_{ROUGE\text{-}N}$, recall $R_{ROUGE\text{-}N}$, and F-value $F_{ROUGE\text{-}N}$ calculated as:

$$R_{ROUGE\text{-}N} = \frac{\sum_{n\text{-}gram \in S_{ref}} Count_{match}(n\text{-}gram)}{\sum_{n\text{-}gram \in S_{ref}} Count(n\text{-}gram)}$$

$$P_{ROUGE\text{-}N} = \frac{\sum_{n\text{-}gram \in S_{sys}} Count_{match}(n\text{-}gram)}{\sum_{n\text{-}gram \in S_{sys}} Count(n\text{-}gram)}$$

$$F_{ROUGE\text{-}N} = \frac{2 \times P_{ROUGE\text{-}N} \times R_{ROUGE\text{-}N}}{P_{ROUGE\text{-}N} + R_{ROUGE\text{-}N}}$$

where $Count_{match}(n\text{-}gram)$ represents the number of co-occurring N-grams; S_{sys} represents N-grams appearing in generated summaries; S_{ref} represents N-grams appearing in reference summaries.

ROUGE-L represents co-occurrence rate statistics based on Longest Common Subsequence (LCS), with precision P_{lcs}, recall R_{lcs}, and F-value F_{lcs} calculated as:

$$R_{lcs} = \frac{\sum_{i=1}^{m} LCS(r_i, S_{sys})}{\sum_{i=1}^{m} n_i}$$

$$P_{lcs} = \frac{\sum_{j=1}^{n} LCS(s_j, S_{ref})}{\sum_{j=1}^{n} m_j}$$

$$F_{lcs} = \frac{(1 + \beta^2) \times R_{lcs} \times P_{lcs}}{R_{lcs} + \beta^2 \times P_{lcs}}$$

where $LCS(r_i, S_{sys})$ represents the union of LCS between reference summary and system summary; β is a balance coefficient measuring the importance between P_{lcs} and R_{lcs}; m and n are the numbers of sentences in system summary and reference summary, respectively.

ROUGE-S represents co-occurrence rate statistics based on skip-bigram sequences, with precision P_{pair} and recall R_{pair} calculated as:

$$R_{pair} = \frac{\sum_{(x,y)\in S_{ref}} Count_{match}(x,y)}{C(m,2)}$$

$$P_{pair} = \frac{\sum_{(x,y)\in S_{sys}} Count_{match}(x,y)}{C(n,2)}$$

$$F_{pair} = \frac{(1+\beta^2) \times R_{pair} \times P_{pair}}{R_{pair} + \beta^2 \times P_{pair}}$$

where $Count_{match}(x,y)$ represents the number of co-occurring word pairs; β is the balance coefficient measuring importance between P_{pair} and R_{pair}; $C(n,2)$ and $C(m,2)$ represent the numbers of word pair combinations in system summary and reference summary, respectively.

2.3 Topic Concept Coverage Analysis

2.3.1 Parameter Values for γ, η, W_{step}, and n_{max} in Fact Filtering Algorithm Theoretically, larger sliding window value W and sliding step size W_{step} provide more contextual information for triple instances. However, the proposed model mainly focuses on semantic compatibility features of triple instances within the sliding window. Therefore, if these parameters are set too large, they will reduce the overall semantic compatibility of the triple instance set and affect model efficiency, causing resource waste. If set too small, they may not capture enough informative instances. The values of γ and η determine the proportions of semantic relevance and literal similarity factors in measuring semantic compatibility between two triple instances, with $\gamma + \eta = 1$.

Through statistical analysis of all results in the fact filtering algorithm, the parameter values in Table 2 enable the extracted triple instance set to achieve optimal semantic compatibility, where n_{max} represents the number of triple instances in the conceptual graph specified by users; W_{step} depends on the sliding window value W.

2.3.2 Topic Concept Coverage According to Equation (1) in Section 1.1, the influence degree of different candidate topic words on topics is calculated for the five experimental topics. The top 200 concepts with high TF-IDF values are selected as topic concepts for each topic. Using the features described in Table 3 , a binary classifier is trained through the random forest algorithm to calculate the Gini coefficient for each concept in the conceptual graph constructed by the proposed model. When a concept' s Gini coefficient is greater than δ, it is identified as a topic concept; otherwise, it is identified as a non-topic concept. Through training, the model achieves 92.3% accuracy. The algorithm parameters are set as follows: threshold $\delta = 0.5$, total number of triple instances in the conceptual graph $n_{max} = 20$, number of subtrees $n\_estimators = 50$,

maximum features for splitting $max\_features = 6$, maximum depth of decision trees $max\_depth = 3$, minimum samples required for internal node splitting $min\_samples\_split = 4$, and minimum samples required for leaf nodes $min\_samples\_leaf = 2$.

The variation of topic concept coverage C_{theme} across different document data scales N ($N \in \{D_{i1}, D_{i2}, D_{i3}, D_{i4}, D_{i5}, D_{i6}, D_{i7}, D_{i8}\}$) under the five experimental topics is shown in Figure 1 [Figure 1: see original paper].

As shown in Figure 1, topic concept coverage rates show a decreasing trend with increasing document data scale across different topics, mainly because larger document quantities increase the dispersion of topic information distribution, and OLLIE system' s extraction accuracy also decreases. Under DUC standard corpus scale $D_{i3}-D_{i4}$, the model can retain 84% of topic information, indicating good precision and generalization ability. For different topics, due to varying document sizes and candidate topic word granularities, performance differs. For example, under the "United States presidential election" topic, which has the largest single document size and most candidate topic words, C_{theme} reaches 92% at D_{i3} scale and 80% at maximum dataset scale D_{i8}. In contrast, the "Sino-Soviet cooperation" topic has the smallest single document size and fewest candidate topic words, with C_{theme} reaching 84% at D_{i3} scale and decreasing to 68% at D_{i8} scale. The "Iran nuclear" and "Volkswagen scandal" topics contain most similar document information, and the variation of C_{theme} with scale N is also very similar, demonstrating that the proposed model has strong correlation with these factors. Under DUC standard corpus scale with sufficient candidate topic words, the model achieves optimal performance with the current classifier.

2.4 Conceptual Graph Connectivity Analysis

For the five experimental topics, only 47% of triple instances obtained through the fact filtering algorithm are easily connectable on average, meaning at least one of their head or tail concepts has the same form. Therefore, the determination and annotation of equivalent concepts and relations directly affect the connectivity of the final generated conceptual graph. Using DUC standard corpus scale as reference, conceptual graphs are generated with D_{i4} scale data ($15 \leq i \leq 19$) according to the proposed model, with fact filtering algorithm parameters set as: $n_{max} = 20$, and optimal values for γ, η, W_{step}, and W. The analysis results are shown in Table 4 . Meanwhile, Tarjan' s algorithm proposed by Robert Tarjan is used to check the strong connectivity of generated conceptual graphs, with results shown in Table 5.

The analysis reveals that each topic has more than three conceptual subgraphs requiring merging, and most of these subgraphs are strongly connected. Expert annotators show good consistency when synthesizing relations (Kappa = 0.89), indicating the proposed model performs well in ensuring overall conceptual graph connectivity. The maximum strongly connected component confidence ratio refers to the proportion of the sum of confidence values (based on

OLLIE system extraction results) of triple instances in the maximum strongly connected component to all triple instances in the synthesized conceptual graph. A higher ratio indicates stronger factual expressive ability of the strongly connected component and better overall connectivity of the synthesized conceptual graph. For example, although the "United States presidential election"and "Volkswagen scandal" topics have the same number of strongly connected components, the latter' s maximum strongly connected component confidence ratio is only 10% (meaning fewer interconnected triple instances with strong factual expressive ability), indicating poorer overall connectivity. The topic concept coverage in the maximum strongly connected component represents the distribution of topic concepts in the conceptual graph' s connected components. Compared with other topics, the "United States presidential election" topic achieves the best connectivity, with topic concept coverage in the maximum strongly connected component reaching 84%, basically approaching the conceptual graph' s topic concept coverage at the current data scale.

2.5 Conceptual Graph Readability Analysis

An effective conceptual graph needs to cover sufficient topic information and possess good readability. To verify the semantic compatibility between topic concepts and overall information readability of conceptual graphs obtained by the proposed model, this experiment uses ROUGE metrics as evaluation standards. Based on analysis results from Section 2.3.2, the "United States presidential election" topic with highest coverage is selected to generate conceptual graphs with D_{i4} scale document data, with fact filtering algorithm parameters set as: $n_{max} = 20$, and optimal values for γ, η, W_{step}, and W.

To make the generated conceptual graph meet summary format requirements, two concept annotators adjust the order of factual information in the conceptual graph (Kappa = 0.89) to form summaries. For document datasets of this scale, relying on domain experts to analyze and generate extractive summaries is unrealistic. Therefore, all documents in the experiment first undergo coreference resolution to further improve corpus readability, then a classic extractive summarization algorithm proposed in Literature [20] is used to generate reference summaries. Finally, generated summaries are compared with reference summaries, with evaluation results shown in Table 6 .

The comparison shows that ROUGE-2 achieves the highest precision, indicating that factual information extracted by the proposed model has good text coverage and demonstrates certain sequential features, meeting basic readability requirements. However, performance on ROUGE-L and ROUGE-S is relatively modest, mainly due to limitations in document data scale and inherent noise in the OLLIE extraction system, making it difficult for the model to reflect sentence-level factual sequential features.

2.6 Model Running Time Analysis

Using DUC standard corpus scale as reference and based on analysis results from Section 2.3.2, document datasets of different scales N ($N \in \{D_{i1}, D_{i2}, D_{i3}, D_{i4}, D_{i5}, D_{i6}, D_{i7}, D_{i8}\}$) are selected from the "United States presidential election" topic with highest coverage to generate conceptual graphs, with fact filtering algorithm parameters set as: $n_{max} = 20$, and optimal values for γ, η, W_{step}, and W. The four main tasks of the model—document ranking, concept and relation extraction, fact filtering, and merging equivalent concepts and relations—are denoted as Tasks 1, 2, 3, and 4, respectively. The variation of average running time for different tasks across current document datasets is analyzed in Table 7 .

As shown in Table 7, computational consumption on Task 3 is the highest, averaging 60% of total model running time, mainly because this task requires calculating semantic compatibility of triple instances within the current sliding window using the ADW model [17], which relies on WordNet' s extensive dictionary to obtain semantic fingerprint information for concepts, significantly reducing computational efficiency. Task 1' s running time mainly depends on document data quality. Overall, since the selected document datasets have high topic coverage, running time on this task does not show erratic jumps with increasing N. Task 2 mainly depends on OLLIE system performance, with running time showing stable growth with N. Task 4' s running time remains basically stable.

With increasing document data scale, the proposed model' s running time across tasks grows linearly with computational complexity. When document data scale reaches D_{i8}, running time does not show erratic growth and remains within acceptable user range. In summary, the proposed model runs stably and demonstrates good performance in adapting to data growth.

2.7 Comparative Algorithm Analysis

Using DUC standard corpus scale as reference, D_{i4} scale document data (containing five topic document sets, i.e., $D_{41}, D_{42}, D_{43}, D_{44}, D_{45}$) is selected for testing and analysis. The proposed model is compared with representative methods, providing their average values across six test metrics: topic concept coverage, number of equivalent concept pairs, number of new semantic relation labels, number of strongly connected components, compatibility of factual knowledge, and ROUGE-2 F1 value. The compatibility of extracted factual knowledge can be calculated through Equation (11). Literature [20] can be regarded as a typical extractive summarization method, which establishes objective functions through monotonic submodular functions, transforms the selection of backbone sentences from multiple documents into an optimization problem, and obtains optimal solutions through greedy algorithms, achieving good performance. This experiment uses this method to generate reference summaries.

1) **Literature [21]**: Stanford OpenIE model is a representative method in

OIE (Open Information Extraction).

2) **Literature [22]**: Mainly uses subject-predicate-object syntactic relations to extract triple instances.
3) **Literature [23]**: Based on the locality assumption of clausal dependency relations, mainly relies on dependency relations to extract triple instances.
4) **Literature [24]**: Mainly uses a co-training method based on the Adaboost iterative algorithm to strengthen relation extraction models, alleviating noise and errors in triple instances. Entity marking in text sentences depends on the Stanford CoreNLP system.
5) **Literature [25]**: Mainly uses distant supervision for entity relation extraction, utilizing Freebase knowledge base and Wikipedia text corpora to automatically obtain relation extraction training data and train models. Entity marking depends on Stanford CoreNLP system.
6) **Proposed Method (without coreference resolution)** (denoted as MDCGCV1): Constructs conceptual graphs based on the proposed model framework but does not perform coreference resolution on pronouns in single documents.
7) **Proposed Method (without fact filtering)** (denoted as MDCGCV2): Based on the proposed model framework, directly inputs OLLIE extraction results into the equivalent concept and relation merging task to construct conceptual graphs.

As shown in Table 8 , the proposed model outperforms other comparative methods with 84.0% topic concept coverage, 94.7% factual knowledge compatibility, and 0.474 ROUGE F1 value, demonstrating its effectiveness in constructing high-quality conceptual graphs.

Literature [21] model mainly uses sentence linguistic structure information to extract triple instances. Its topic concept coverage differs from the proposed model by about 10 percentage points, mainly because the OLLIE algorithm used in this paper better addresses long-range dependency problems in sentences with relatively higher precision. Literature [21] model achieves 79.2% factual knowledge compatibility and 0.346 ROUGE-2 F1 value, affected by extraction precision.

Literature [22] model' s limitations mainly include: (a) using only subject-predicate-object syntactic relations for triple instance extraction has problems parsing long sentences; (b) over-reliance on correct word segmentation results. Literature [23] model identifies and analyzes dependency relations in syntactic structures based on Literature [22], showing overall better performance with 81.6% topic concept coverage. Both Literature [22] and [23] models exceed 90% in factual knowledge compatibility, demonstrating the effectiveness of the proposed fact filtering algorithm based on syntactic analysis.

Literature [24] model uses co-training based on Adaboost to strengthen relation extraction models, alleviating triple instance noise to some extent. However, for problems such as overly dispersed knowledge units under the same topic and poor corpus completeness in this experiment's corpora, the model's performance

is still limited, achieving 72.8% concept coverage.

Literature [25] model uses distant supervision for relation extraction, achieving 78.4% topic concept coverage. The reasons include: (a) the method's assumption that all sentences containing entity pairs imply potential relationships between them is not well supported by the incomplete text corpora provided; (b) unspecified relation types [26] in this experiment' s corpora are not well annotated; (c) too few training samples constrain model performance.

MDCGCV1 model does not perform coreference resolution on pronouns in single documents, resulting in: (a) significantly increased unclear or invalid triple instances; (b) poor readability of extracted factual knowledge; (c) "false compatibility" phenomena in triple instances generated through fact filtering due to semantic unit ambiguity. Therefore, this model' s average topic concept coverage is only 64.0%, and factual knowledge compatibility is only 42.3%, lower than the average of all methods.

MDCGCV2 model, due to OLLIE extraction results not undergoing fact filtering, cannot effectively handle uninformative or redundant triple instances affected by the model' s inherent precision and error propagation, achieving 67.2% topic concept coverage and 44.2% factual knowledge compatibility.

3 Conclusion

To address the problem that topic information is distributed across documents, making it difficult for users to mine and organize core concepts and semantic connections, this paper proposes a multi-document conceptual graph construction model based on open-domain extraction. To verify the model' s effectiveness, experiments are conducted on the real news dataset released by Signal Media from five aspects: topic concept coverage, conceptual graph connectivity, conceptual graph readability, model running time, and comparative algorithms. Experimental results demonstrate that the proposed conceptual graph construction model can cross-document mine and organize key information related to specific topics, representing significant entities, concepts, and their relationships under specific topics through conceptual graphs. The conceptual graph achieves good performance in metrics such as topic concept coverage and factual knowledge compatibility, and holds important reference value for automatic document summarization applications.

However, the proposed model still has certain limitations. For instance, concept and relation extraction tasks mainly depend on the open-domain extraction system OLLIE, and extracted triple instances contain considerable noise. Additionally, OLLIE is limited to English text and cannot be applied to Chinese text corpora with complex structures and multi-semantic concepts. Therefore, future work will attempt to introduce semantic dependency analysis into this research to more accurately extract topic word pairs and their semantic relationships from document topic sentences, and further expand the application scope of the proposed model.

References

[1] Novak J D, Cañas A J. Theoretical origins of concept maps, how to construct them, and uses in education [J]. Reflecting Education, 2007, 3 (1): 29-42.

[2] Erkan, Günes, Dragomir R. Radev. LexRank: graph-based lexical centrality as salience in text summarization [J]. Journal of Artificial Intelligence Research, 2004, 22: 457-479.

[3] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C]// Advances in Neural Information Processing Systems. 2014: 3104-3112.

[4] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3 (1): 993-1022.

[5] Celikyilmaz A, Hakkani-Tur D. A hybrid hierarchical model for multi-document summarization [C]// Proc of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 815-824.

[6] Lin C Y, Hovy E H. From single to multi-document summarization [C]// Proc of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002: 457-464.

[7] Fabbrizio G D, Aker A, Gaizauskas R. Summarizing online reviews using aspect rating distributions and language modeling [J]. IEEE Intelligent Systems, 2013, 28 (3): 28-37.

[8] Banko M, Cafarella M J, Soderland S, et al. Open information extraction from the Web [C]// Proc of the 18th International Joint Conference on Artifical Intelligence. Cambridge, MA: AAAI, 2007: 2670-2676.

[9] 杨博, 蔡东风, 杨华. 开放式信息抽取研究进展 [J]. 中文信息学报, 2014, 28 (4): 1-11. (Yang Bo, Cai Dongfeng, Yang Hua. Progress in open information extraction [J]. Journal of Chinese Information Processing, 2014, 28 (4): 1-11.)

[10] Manning C, Surdeanu M, Bauer J, et al. Multi-document relationship fusion via constraints on probabilistic databases [C]// Proc of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. 2007: 332-339.

[11] 刘绍毓, 李弼程, 郭志刚, 等. 实体关系抽取研究综述 [J]. 信息工程大学学报, 2016, 17 (5): 541-547. (Liu Shaoyu, Li Bicheng, Guo Zhigang, et al. Review of entity relation extraction [J]. Journal of Chinese Information Processing, 2016, 17 (5): 541-547.)

[12] 王林玉, 王莉, 郑婷一. 基于卷积神经网络和关键词策略的实体关系抽取方法 [J]. 模式识别与人工智能, 2017, 30 (5): 465-472. (Wang Linyu, Wang Li, Zheng Tingyi. Entity relation extraction based on convolutional neural network and keywords strategy [J]. Pattern Recognition and Artificial Intelligence, 2017, 30 (5): 465-472.)

[13] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 1753-1762.

[14] Kumar S. A survey of deep learning methods for relation extraction [EB/OL]. (2017). https://arxiv.org/abs/1705.03645.

[15] Manning C, Surdeanu M, Bauer J, et al. The stanford CoreNLP natural language processing Toolkit [C]// Proc of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2014: 55-60.

[16] Mihalcea R, Tarau P. TextRank: bringing order into texts [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2004: 404-411.

[17] Pilehvar M T, Jurgens D, Navigli R. Align, disambiguate and walk: a unified approach for measuring semantic similarity [C]// Proc of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2013: 1341-1351.

[18] 王桐, 王磊, 吴吉义, 等. WordNet 中的综合概念语义相似度计算方法 [J]. 北京邮电大学学报, 2013, 36 (2): 98-101. (Wang Tong, Wang Lei, Wu Jiyi, et al. Semantic similarity calculation method of comprehensive concept in WordNet [J]. Journal of Beijing University of Posts and Telecommunications, 2013, 36 (2): 98-101.)

[19] 文本理解会议标准语料库 [EB/OL]. [2014-09-09]. https://duc.nist.gov/. (Text understanding conference standard corpus [EB/OL]. [2014-09-09]. https://duc.nist.gov/.)

[20] Li Jingxuan, Li Lei, Li Tao. Multi-document summarization via submodularity [J]. Applied Intelligence, 37 (3): 420-430.

[21] Angeli G, Premkumar M J J, Manning C D. Leveraging linguistic structure for open domain information extraction [C]// Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 344-354.

[22] Jing Tao, Zuo Wanli, Sun Jigui, et al. Semantic annotation of Chinese Web pages: from sentences to RDF representations [J]. Journal of Computer Research and Development, 2008, 45 (7): 1221-1231.

[23] 荆涛. 面向领域网页的语义标注若干问题研究 [D]. 长春: 吉林大学, 2011. (Jing Tao. Research on semantic annotation of domain oriented Web pages [D]. Changchun: Jilin University, 2011.)

[24] 王旭阳, 姜喜秋. 特定领域概念属性关系抽取方法研究 [J]. 吉林大学学报: 信息科学版, 2017, 35 (4): 430-437. (Wang Xuyang, Jiang Xiqiu. Research on extraction

method of specific domain concept and property [J]. Journal of Jilin University: Information Science Edition, 2017, 35 (4): 430-437.)

[25] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data [C]// Proc of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg: Association for Computational Linguistics, 2009: 1003-1011.

[26] Chan Y S, Roth D. Exploiting background knowledge for relation extraction [C]// Proc of the 23rd International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 152-160.

Note: Figure translations are in progress. See original paper for figures.