Compressive Multi-document Summarization with Sense-level Concepts[*]

Xin Shen[1] , **Wai Lam**[1†] , **Xunying Liu**[1] and **Piji Li**[2]
[1]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong
[2]Tencent AI Lab
{xshen, wlam, xyliu}@se.cuhk.edu.hk, lipiji.pz@gmail.com

Abstract

Most existing document summarization methods make use of information appeared in input documents only. They operate on the word level, and do not make use of the sense-level concepts which usually need external semantic knowledge. In this paper, we investigate a compressive Multi-Document Summarization (MDS) model which integrates the sense-level concepts of words and entities. With sense disambiguation of the text, we propose a novel way of salience estimation and redundancy reduction that better capture the semantic meanings of words and entities in the input text. The experimental results demonstrate the effectiveness of our approach.

1 Introduction

The task of text summarization aims to generate a short summary for the given documents [Goldstein *et al.*, 2000; Mihalcea and Tarau, 2004; Bing *et al.*, 2015]. A good summary should be concise and captures the most important information of the original documents. The solution approaches to this task can be divided into extractive models [Goldstein *et al.*, 2000; Mihalcea and Tarau, 2004] and abstractive models [Nallapati *et al.*, 2016]. Extractive models directly select words, phrases, and sentences from the source documents to the summaries. Compression summarization [Knight and Marcu, 2000; Lin, 2003] can be viewed as a special kind of extractive approach, which selects sentences first and then compresses the selected sentences by deleting words and phrases to get a more concise summary. On the other hand, Multi-Document Summarization (MDS) [Goldstein *et al.*, 2000; Lin, 2003; Bing *et al.*, 2015] is a subclass of text summarization tasks, where the input is a cluster of given documents sharing the same event or topic.

Different from most existing automatic text summarization approaches, when human beings write a summary, they not only consider the text to be summarized, but also utilize their common sense knowledge and reasoning ability. In the reading process, humans usually need to figure out the actual meanings of the words and entities in text, which corresponds to Word Sense Disambiguation (WSD) [Navigli, 2009] and Entity Linking (EL) [Shen *et al.*, 2015], respectively. Word sense information or concepts are usually stored in linguistic resources such as Wordnet [Miller, 1995], and entities can be found in different knowledge bases [Nickel *et al.*, 2016]. When humans write a summary, they would use various relations between word senses or entities for reasoning, which can only be retrieved from external sources or knowledge outside the text to be summarized. The analogy of human writing process provides an insight that the sense-level concepts and common sense knowledge are of great importance for automatic text summarization. However, as far as we know, only very few works such as [Pourvali and Abadeh, 2012; Sankarasubramaniam *et al.*, 2014] use such information.

In fact, the above dilemma is not restricted to text summarization. As a general trend, many current NLP systems operate on the word level, that is, individual words constitute the most fine-grained meaning bearing units of the representation [Pilehvar *et al.*, 2017]. This word level mechanism imposes several restrictions and limitations as follows: (1) It is not effective at handling multi-words representation, such as one entity composed of several words, especially when they appear infrequently. (2) It cannot disambiguate between different senses of one word, e.g. polysemy, or recognize different words share the meaning in certain contexts, e.g. synonym. (3) Because of the inability to capture the intended meanings, it cannot effectively utilize external sources of information such as Wordnet or knowledge base. In [Li and Jurafsky, 2015], a multi-sense embedding model is established and it does improve the performance of tasks such as part-of-speech tagging and semantic relatedness measurement. In [Pilehvar *et al.*, 2017], a pipeline is created to integrate sense-level concepts into the down-stream NLP applications, and its effectiveness is validated in tasks such as multiple topic categorization and polarity detection. Motivated by these works, we investigate a compressive MDS model which exploits the sense-level concepts of words and entities.

The contributions of this paper are as follows. We propose to use a disambiguated representation for the input texts of the summarization task, which is rarely explored by previous summarization models. With this form, we propose a novel

[*]The work described in this paper is substantially supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14204418).

[†]Contact Author

semantically-aware technique for salience estimation and redundancy reduction. The experimental results show that our proposed sense-level concepts is beneficial for the compressive MDS task.

2 Framework

The problem setting of MDS is as follows. Given an event (topic) composed of a set of text documents talking about the same event, the aim of MDS is to generate a single summary with a predefined word limit.

2.1 Baseline model

Our baseline model is derived from [Bing *et al.*, 2015][1]. This model first uses the constituency tree to decompose the sentences in inputs into a set of noun phrases (NPs) and a set of verb phrases (VPs). Then it tries to solve the following Integer Linear Programming (ILP) [Dantzig and Thapa, 2006] Eqn (1). The aim is to select the most salient phrases and remove redundant information at the same time:

$$\begin{aligned}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \{\sum_i \alpha_i S_i^N - \sum_{i<j} \alpha_{ij}(S_i^N + S_j^N)R_{ij}^N \\ & + \sum_i \beta_i S_i^V - \sum_{i<j} \beta_{ij}(S_i^V + S_j^V)R_{ij}^V\}\end{aligned} \qquad (1)$$

Here, the variables α_i and β_i indicate selecting phrase i or not, and α_{ij} and β_{ij} indicate whether to select phrases i and j at the same time. S_i^N are S_i^V are the salience score that measures the importance of noun and verb phrases, respectively. Due to the nature of MDS, the documents of the same event often contain redundant information. In Eqn (1), R_{ij}^N and R_{ij}^V measure the similarity between phrases i and j, and the second and fourth terms in Eqn (1) try to avoid to select similar phrases at the same time. Several constraints are designed for ensuring the compatibility of sentence generation but they are omitted here for simplicity.

2.2 Sense-level concept

To get the sense-level concept, we use the Babelfy tool[2] [Moro *et al.*, 2014b; Moro *et al.*, 2014a] to conduct WSD and EL for the input documents. After that, the original word-level representation in the input are transformed into a set of *synsets* in Babelnet[3] [Navigli and Ponzetto, 2012]: $\{s_1, s_2, \cdots, s_m\}$. In Babelnet, *synset* is the basic concept representation that bears the minimal meaningful unit. After the process above, phrases with multi-words such as *United States* are packed into one unit, which is treated as a synset concept, such concept also captures other words such as *America*, acronyms such as *USA*, and other multi-word representations such as *United States of America*, according to their located context and actual meanings. To make the decision of disambiguation, this process makes use of word-sense level information from Wordnet, entity-sense level information from Wikidata and Wikipedia, which have been integrated into Babelnet.

Since the current performance of WSD and EL methods are far from perfect, we propose the following strategy to minimize the effect of error propagation. Specifically, we try to prevent the inaccurately recognized synsets from degrading the performance of our MDS framework. Since the mechanism in Babelfy returns a confidence score for the recognized synset concepts ranging in $[0, 1]$, we set a threshold to keep the synsets with high confidence, and use the original word form for the candidates with lower scores. [Camacho-Collados *et al.*, 2016] provides a method for learning embeddings for both words and BabelNet synsets in the same vector space[4]. By using these pre-trained embedding, we can represent the original documents of one event in a unified concept representation:

$$\mathcal{S}_{docs} = \{\boldsymbol{e}_{s_1}, \boldsymbol{e}_{s_2}, \cdots, \boldsymbol{e}_{s_m}, \boldsymbol{e}_{w_1}, \boldsymbol{e}_{w_2}, \cdots, \boldsymbol{e}_{w_n}\}, \qquad (2)$$

where $\boldsymbol{e}_{s_i}$ is the embedding for the synset s_i and $\boldsymbol{e}_{w_i}$ is the embedding for the word w_i.

As mentioned in Subsection 2.1, a very important issue is how to compute the salience scores S and the similarity measure R. In the following two subsections, we discuss the solutions to these two issues with the sense-level concept representations obtained above.

2.3 Salience estimation

The model in [Bing *et al.*, 2015] employs a concept based method [Li *et al.*, 2011] for salience estimation. In [Bing *et al.*, 2015], the concept set is designated to be the union set of *unigrams*, *bigrams*, and *named entities* in the documents. It counts the frequency for each concept as its weight. Then, the salience of a phrase is calculated as the summed weights of its concepts.

Here, we propose a novel way of salience estimation for phrases with sense-level concepts derived above. Using the above representation, all the documents in one event can be represented as the linear combination of the contained concepts:

$$\boldsymbol{v}_{docs} = \sum_{\boldsymbol{e}_i \in \mathcal{S}_{docs}} p_i \boldsymbol{e}_i, \qquad (3)$$

where p_i is a position-based weighting score for the concept i [Bing *et al.*, 2015]. The motivation of this weighting mechanism is: the importance of one concept is associated with its position in the document. If we denote the set of concept embeddings of the j-th phrase as $\mathcal{S}_{phr_j}$ and use the summation of embeddings contained in $\mathcal{S}_{phr_j}$ to represent the the vector $\boldsymbol{v}_{phr_j}$, then the inner product of $\boldsymbol{v}_{phr_j}$ and $\boldsymbol{v}_{docs}$ can be stated as:

$$< \boldsymbol{v}_{docs}, \boldsymbol{v}_{phr_j} >= \sum_{\boldsymbol{e}_i \in \mathcal{S}_{docs}} \sum_{\boldsymbol{e}_k \in \mathcal{S}_{phr_j}} p_i < \boldsymbol{e}_i, \boldsymbol{e}_k > . \qquad (4)$$

[1] The original model in [Bing *et al.*, 2015] is abstractive which allows to combine noun phrases and verb phrases coming from different sentences. In our paper, we adapt it to a compressive model where the phrases to form a new sentence are from the same source sentence.

[2] http://babelfy.org

[3] https://babelnet.org

[4] The original version of this semantic representation only contains embeddings for noun words and synsets, we use a slightly different version which contains embeddings for verb and adjective synsets and words.

The above formula provides an intuitive insight. When $\boldsymbol{e}_i$ and $\boldsymbol{e}_k$ are the same embeddings, $\sum p_i < \boldsymbol{e}_i, \boldsymbol{e}_k >$ is the frequency of the concept i weighted with position in the document, which is similar to the frequency of unigrams used in existing models such as [Bing *et al.*, 2015]. Otherwise, $< \boldsymbol{e}_i, \boldsymbol{e}_k >$ is the semantic relatedness between these embeddings. Intuitively, when the $\mathcal{S}_{docs}$ has a high frequency of $\boldsymbol{e}_i$, it conveys a strong semantic meaning in this sense concept. The phrase contains similar sensed embedding should also get a salience score from this part. To sum up, $< \boldsymbol{v}_{docs}, \boldsymbol{v}_{phr_j} >$ is the semantic meaning of documents that is conveyed by the j-th phrase. Naturally, we can use it as the salience score for the this phrase.

2.4 Redundancy reduction

In [Bing *et al.*, 2015], Jaccard Index is employed as the similarity measure between phrases. Jaccard similarity is defined as the size of the intersection of the words in the two phrases compared to the size of the union of the words in the two phrases. Obviously, it only measures the superficial intersection and does not consider the semantic similarity. Therefore, it may result in the cases where phrases are literally different but semantically redundant.

For the computation of the similarity between two texts, while the best results are achieved using dedicated models, solutions based on pre-trained word embeddings are often very competitive [Goldberg, 2017]. Since we obtain a more dedicated sense-level embedding than the generally-purposed word-level embedding, it is natural to use it for measuring the similarity R. Specifically, we can use the following semantic similarity as the similarity measure R_{ij} in Eqn (1):

$$\frac{< \boldsymbol{v}_{phr_i}, \boldsymbol{v}_{phr_j} >}{||\boldsymbol{v}_{phr_i}|| \cdot ||\boldsymbol{v}_{phr_j}||}. \tag{5}$$

3 Experiments

3.1 Dataset and Metrics

We use Reader-Aware Multi-Document Summarization (RA-MDS)[5] mentioned in [Li *et al.*, 2017] as the experimental dataset. It is a MDS dataset containing news documents featured with corresponding user comments. We use 8 events under the *Legal* category and 6 events under the *Attacks* category to conduct our experiments. Each event contains 10 news documents. Furthermore, each event is associated with 4 model summaries written by human and they can be treated as the gold-standard summaries for supporting evaluation. We use ROUGE score as our evaluation metric [Lin, 2004]. F-measures of ROUGE-1, and ROUGE-SU4 are reported.

3.2 Comparative methods

We compare our model with the following methods:

- **Lead** [Wasson, 1998]: It ranks the news sentences chronologically and extracts the leading sentences one by one until the length limit.
- **TextRank** [Mihalcea and Tarau, 2004]: It is a graph-based unsupervised method for sentence salience estimation based on PageRank algorithm.
- **Concept** [Bing *et al.*, 2015]: It is the existing model described in Subsection 2.1.

Although RA-MDS dataset contains user comments, we did not consider comments for all the methods here, because comments are not the gist of this paper. We did not compare with the method in [Li *et al.*, 2017] because their model considers user comments.

3.3 Results

Category		**Legal**		**Attacks**
Rouge	**R-1**	**R-SU4**	**R-1**	**R-SU4**
Lead	0.413	0.172	0.378	0.139
TextRank	0.423	0.178	0.381	0.139
Concept	0.432	0.197	0.391	0.153
Semantic	**0.441**	**0.205**	**0.402**	**0.157**

Table 1: Experimental results. R-1 and R-SU4 refer to ROUGE-1 and ROUGE-SU4, respectively.

The results of our framework named as **Semantic** as well as the comparative methods are depicted in Table 1. From the results, it can be seen that our framework outperforms all the comparative methods.

3.4 Case study

As discussed in Subsection 2.4, the similarity measure is of great importance for the redundancy removal in MDS models. As a case study, we compare Jaccard index and semantic similarity measures on one phrase pair from the event *Akademik Shokalskiy Trapping*. The first phrase is *The Snow Dragon* and the second one is *Three icebreakers*. Since these two phrases have no overlapping, the Jaccard Index score is zero. For our sense-level representation, *Snow Dragon* is disambiguated into one synset concept, and a good synset embedding captures the relation that *Snow Dragon* is the name of an icebreaker, thus it returns a high score of 0.801.

Generated summary from the event: *Bremerton Teen Arrested Murder 6-year-old Girl*
A 17 year old boy has been arrested in the death of a 6 year old girl who vanished from her home last Saturday. The teenager, whose name was not released, will be charged with second degree murder, first degree manslaughter and rape, according to Kitsap County Sheriff 's Lt. Earl Smith. The Washington state crime lab made positive confirmation of the suspect through forensic evidence. Hundreds of people, including officers from 15 law enforcement agencies, went door to door at the mobile home park on the west side of Puget Sound, across from Seattle.

Table 2: Case study of the summary generation

As another case study, we select one event and present the generated summary by our semantic-aware model in Table 2.

[5] http://www.se.cuhk.edu.hk/~textmine/dataset/ra-mds/

We use the bold font to highlight the parts that have an overlapping with the gold-standard summaries. From this case study, it can be seen that our framework well captures the salient parts of the input documents, and the generated summary is of good quality.

4 Conclusions

We propose a novel MDS model incorporated with fine-grained sense-level concepts. Specifically, with sense disambiguation of the text, we propose a new way of salience estimation and redundancy reduction. In the experiments, we compare our new method with the baseline and the existing MDS models. The results show that our new model can better capture semantic information than the previous MDS models.

In the future, we hope to use this sense-level representation for other NLP problems. Furthermore, controlling error propagation of the disambiguation process for the downstream NLP tasks is also an interesting research topic.

References

[Bing *et al.*, 2015] Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca Passonneau. Abstractive multi-document summarization via phrase selection and merging. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1587–1597, 2015.

[Camacho-Collados *et al.*, 2016] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64, 2016.

[Dantzig and Thapa, 2006] George B Dantzig and Mukund N Thapa. *Linear programming 1: introduction*. Springer Science & Business Media, 2006.

[Goldberg, 2017] Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

[Goldstein *et al.*, 2000] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, pages 40–48, 2000.

[Knight and Marcu, 2000] Kevin Knight and Daniel Marcu. Statistics-based summarization-step one: Sentence compression. *AAAI/IAAI*, 2000:703–710, 2000.

[Li and Jurafsky, 2015] Jiwei Li and Dan Jurafsky. Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, 2015.

[Li *et al.*, 2011] Huiying Li, Yue Hu, Zeyuan Li, Xiaojun Wan, and Jianguo Xiao. Pkutm participation in tac2011. *Proceeding RTE*, 7, 2011.

[Li *et al.*, 2017] Piji Li, Lidong Bing, and Wai Lam. Reader-aware multi-document summarization: An enhanced model and the first dataset. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 91–99, 2017.

[Lin, 2003] Chin-Yew Lin. Improving summarization performance by sentence compression–a pilot study. In *Proceedings of the sixth international workshop on Information retrieval with Asian languages*, 2003.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[Mihalcea and Tarau, 2004] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of conference on empirical methods in natural language processing*, 2004.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[Moro *et al.*, 2014a] Andrea Moro, Francesco Cecconi, and Roberto Navigli. Multilingual word sense disambiguation and entity linking for everybody. In *International Semantic Web Conference*, pages 25–28, 2014.

[Moro *et al.*, 2014b] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244, 2014.

[Nallapati *et al.*, 2016] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, 2016.

[Navigli and Ponzetto, 2012] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[Navigli, 2009] Roberto Navigli. Word sense disambiguation: A survey. *ACM computing surveys*, 41(2):10, 2009.

[Nickel *et al.*, 2016] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

[Pilehvar *et al.*, 2017] Mohammad Taher Pilehvar, Jose Camacho-Collados, Roberto Navigli, and Nigel Collier. Towards a seamless integration of word senses into downstream nlp applications. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1857–1869, 2017.

[Pourvali and Abadeh, 2012] Mohsen Pourvali and Mohammad Saniee Abadeh. Automated text summarization base on lexicales chain and graph using of wordnet and wikipedia knowledge base. *International Journal of Computer Science Issues (IJCSI)*, 9(1):343, 2012.

[Sankarasubramaniam *et al.*, 2014] Yogesh Sankarasubramaniam, Krishnan Ramanathan, and Subhankar Ghosh. Text summarization using wikipedia. *Information Processing & Management*, 50(3):443–461, 2014.

[Shen *et al.*, 2015] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2015.

[Wasson, 1998] Mark Wasson. Using leading text for news summaries: evaluation results and implications for commercial summarization applications. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 1364–1368, 1998.