Exploring Text Links for Coherent Multi-Document Summarization

Xun Wang, Masaaki Nishino, Tsutomu Hirao, Katsuhito Sudoh and Masaaki Nagata
NTT Communication Science Labortories
Kyoto, 619-0237, Japan

{wang.xun,nishino.masaaki,hirao.tsutomu
sudoh.katsuhito,nagata.masaaki@lab.ntt.co.jp}

Abstract

Summarization aims to represent source documents by a shortened passage. Existing methods focus on the extraction of key information, but often neglect coherence. Hence the generated summaries suffer from a lack of readability. To address this problem, we have developed a graph-based method by exploring the links between text to produce coherent summaries. Our approach involves finding a sequence of sentences that best represent the key information in a coherent way. In contrast to the previous methods that focus only on salience, the proposed method addresses both coherence and informativeness based on textual linkages. We conduct experiments on the DUC2004 summarization task data set. A performance comparison reveals that the summaries generated by the proposed system achieve comparable results in terms of the ROUGE metric, and show improvements in readability by human evaluation.

1 Introduction

Automatic summarization is extremely useful in this age of information overload. It provides readers with easier access to information without the labour of reading the source text. According to the number of documents dealt with, summarization falls into two categories: single document summarization and multi-document summarization. While they both aim to represent the source text using a shorten passage, the latter deals with a set of documents sharing the same topic. Based on the method adopted, existing approaches to summarization can be divided into two kinds: abstraction based or extraction based. The difference lies in the sentences they use to generate summaries: the former selects sentences (clauses, or other text units, hereafter we refer to all of them as sentences.) from source documents and the latter generates new sentences. Most existing summarization systems are extraction-based because abstraction-based methods require the use of natural language generation technology, which is still a growing field. This paper, without exception, also employs extraction-based methods and we focus on multi-documents summarization.

Currently the extraction-based methods face some major challenges. One is informativeness, which means we need to maintain the important information of source documents in summaries. This is the focus of almost all research on summarization. Another challenge is presentation, namely that the extracted text should be well presented, i.e., it should contain little redundancy and be coherent so as to be readily understandable. Previous work has addressed the problem of redundancy, and some successful solutions like Maximum Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) have been proposed and widely adopted (e.g., (Li and Li, 2013)), but very few try to deal with coherence. Therefore the generated summaries generally suffer as regards readability and are very difficult to use for practical applications. In the report of the TAC 2011 summarization task (Owczarzak and Dang, 2011), it is stated that "in general, automatic summaries are better than baselines[1], except *Readability*." Such a statement suggests, as for summarization, coherence should be treated with the same as salience and redundancy.

[1]The baseline they used is the lead paragraph method and summaries are evaluated by human and ROUGE (Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004)).

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 213–223, Osaka, Japan, December 11-17 2016.

Existing work addresses coherence in summarization from different aspects. One kind of method employs reordering after selecting sentences, and the drawback is evident: coherence is considered after sentence selection. Another kind of widely adopted method takes discourse relations into consideration when selecting sentences, as discourse relations are believed to be essential for maintaining textual coherence. Hirao et al. (2013) formulated single document summarization as to extract a sub tree from the complete discourse tree and thus preserve the relations between extracted document units to form a readable text. Wang et al. (2015) extended it to multi-document summarization by regarding a document set as one document and developed a model which combined discourse parsing and summarization together. Christensen et al. (2013) proposed a graph-based model to bypass the tree constraints. They employed rich textual features to build a discourse relation graph for source documents with the aim of representing the relations between sentences (both inter and intra-document relations). Christensen et al. (2013) reported ROUGE scores lower than some baselines. This is because that, they claim, ROUGE is salience-focused and fails to notice the improvement in coherence. In a further human evaluation, they reported improvements in readability.

These discourse-based methods without exception have discourse analysis as a prerequisite. As we all know, discourse analysis is still under development thus preventing the expected improvement. Furthermore, languages other than English do not enjoy plenty of ready-to-use discourse analysis tools. This also limits the usage of these discourse-based methods.

Is it possible to consider coherence in summarization without discourse analysis? Before answering this question, we need to find out what is the key to coherence in text. According to the centering theory (Grosz et al., 1995; Walker et al., 1998), the coherence of text is to a large extent maintained by entities and the relations between them. This indicates that discourse analysis is not a must to preserve coherence; we can directly take advantages of entities and their relations to generate coherence text.

Based on this point, we design a novel graph-based model for multi-document summarization that eliminates the effort of conducting discourse relation analysis (inter or intra document) and generates informative and readable summaries. We formulate the document set as a graph whose nodes corresponds to sentences. These nodes are connected with each other according to the entities they contains and the relations between their containing entities. Each path in the graph represents a piece of text and is evaluated using a novel scoring function that considers informativeness and coherence. To extract a summary is to find a path in the graph with the highest score. This is a weighted longest path problem. We further present a variant of the proposed model based on local coherence and explore decoding algorithms for both of them.

Experiments are conducted on the Document Understanding Conference (DUC) 2004 multi-document summarization task data set. As ROUGE cannot fully capture our improvement in coherence which is one of the key contributions of this work, we also conduct a human evaluation. Results show that we obtain summaries comparative with state-of-the-art systems in terms of ROUGE metrics and get improvements in readability in human evaluations.

This work provides a method of generating high quality summaries without the effort of discourse analysis. The proposed method can be easily extended to other languages without much efforts. It also provides inspiration as regards other tasks that require computers to generate coherent text. The rest of the papers is organized as follows: Section 2 presents the centering theory and a coherence model based on entities. Section 3 presents our model. Section 4 describes the experiments and results. Section 5 presents some previous work and Section 6 concludes this paper.

2 Centering Theory and Coherence Modelling

The centering theory (Grosz et al., 1995) as a popular theory on discourse analysis, serves as the basis of some coherence evaluation methods (Barzilay and Lapata, 2008; Burstein et al., 2010; Li and Hovy, 2014; Li and Jurafsky, 2016) and enables us to measure the coherence score of any given text without discourse parsing solely based on the reappearance of entities. Entities here refer to noun/pronoun

word/phrases [2].

According to the centering theory, we have the following assumptions:

1. Text that contains successive mentions of the same entities would be more coherent.

2. The main entities that are focused on tend to play an important grammatical role, such as the subject or object of the sentences.

Therefore the key to the coherence of a text lies in what entities it contains and how their roles change. The coherence of a generated text can be evaluated accordingly.

Barzilay and Lapata (2008) presented such a model. The key is to represent text as an *entity grid*. Assume text T contains n sentences $\{S_1, S_2..., S_n\}$ and m entities. r_i^k represents the grammatical role of Entity e_k in Sentence S_i. Four kinds of roles are used, i.e., "subj", "obj", "others" and "absent". "Others" indicates that the entity is present, but is neither the subject nor the object. Then the grammatical roles of e_k in text T can be expressed as a sequence: $\{r_1^k, r_2^k, ..., r_n^k\}$. For each entity in T, such a chain showing how the entity's grammatical roles change in T is extracted. Thus text T can be represented as an $n * m$ matrix $M(T)$ where n is the number of sentences and m is the number of entities in T, and $M(T)_{ij}$ corresponds to the grammatical roles of Entity j in Sentence i. $M(T)$ is referred to as the *Entity Grid* of T (Barzilay and Lapata, 2008).

To calculate the coherence score of T, Barzilay and Lapata (2008) used $M(T)$ as a feature vector. They calculated the transition probability for $|\{s(subj), o(bj), x(others), -(absent)\}^2| = 16$ transition patterns from $M(T)$ without distinguishing between entities, to form a vector $f(T)$ for T, and a weight vector w was then learnt from training data so that $w * f(T)$ can be used as the coherence score for T.

This kind of method has been adopted in many studies (Filippova and Strube, 2007; Barzilay and Lapata, 2008; Burstein et al., 2010). In particular, Filatova and Hatzivassiloglou (2004) extends entity grids to model semantical relations between entities, which provides a possible further improvement for our models.

3 Modeling Summarization

The above model can only be used to measure coherence but summarization is much complex as it involves not only coherence bust also informativeness and redundancy. We design a much more sophisticated models leveraging entities.

Two models are presented below. Both of them are based on entities and consider coherence as well as informativeness. The first one is based on global coherence and the second one local coherence. The global coherence consider the full sequence when evaluating coherence and the local coherence is calculated based on relations between adjacent sentences. Intuitively, global coherence is better than local coherence, but considering the full sequence increases the time complexity. The model based on local coherence, on the other hand, reduces the time complexity and enables us to obtain an exact solution efficiently.

3.1 Problem Set-up

Assume we have K documents with n sentences in total. Note that we are dealing with multi-document summarization, and we do not distinguish between inter-document and intra-document relations. We construct a graph with n nodes, each of which corresponds to one sentence. Weighted directed edges are used to connect these nodes together. To each node, we assign a cost score, which is the number of words the corresponding sentence contains. To each path in the directed graph, we assign a gain score. The gain score is a comprehensive evaluation of the informativeness and coherence of the sequence of sentences represented by the path. The problem of extracting a good summary becomes the problem of extracting the best path. Note that it is an asymmetric graph. Gain scores for $A \rightarrow B \rightarrow C$ and $C \rightarrow B \rightarrow A$ are different. The direction determines the positions of corresponding sentences in the generated text.

[2] In some previous work on summarization (Takamura and Okumura, 2009; Hirao et al., 2013), concepts are used to measure informativeness. Concepts can be used to refer any non functional words, including adjectives, adverbs. All the entities can be regarded as concepts, but some concept words (non-nominal words) are not entities. Entity is a subset of Concept.

One more thing to consider is the redundancy. Instead of formulating redundancy explicitly, we remove edges connecting similar sentences to turn the complete graph into an incomplete graph. This ensures that similar sentences do not occupy adjacent positions in the generated summaries and thus reduce redundancy. The similarities of sentence pairs are based on word overlaps and we keep $d\%$ of all the edges.

Note that for temporal text removing edges can also help us maintain the temporal relations between sentences, though we do not explore this point here.

3.2 Summarization Considering Global [3] Coherence

To extract a summary is to find such a sequence of sentences Seq that maximizes $Score(Seq)$.

$$\begin{aligned} Score(Seq) &= \sum_{k=1}^{m} a_k F_k \\ F_k &= \prod_i p_{e_k}(r_i^k r_{i+1}^k), S_i, S_{i+1} \in Seq \\ s.t. &\sum_{S_i \in Seq} length(S_i) \leq threshold \end{aligned} \tag{1}$$

a_k is the weight of Entity e_k. r_i^k is the state of Entity e_k in Sentence S_i. Here we use four states: "s", "o", "x", "-", which represent "subj", "obj", "present" and "absent" respectively. It is also possible to use more or fewer states.

$p_{e_k}(**)$ is the transition probability between two states for e_k. For each document set, the transition probabilities for each entity is estimated using $p_{e_k}(ab) = \frac{\#e_k(a)e_k(b)}{n-K}$. $\#e_k(a)e_k(b)$ marks the times that Entity e_k presents as grammatical role a in the preceding sentence and as grammatical role b in the following one. $n-K$ denotes the total number of adjacent sentence pairs in a document set with K documents and n sentences. F_k is the coherence score contributed by e_k in the extracted sequence Seq. F_k is based on the transitions of e_k between adjacent sentences in Seq. We use $Score(Seq)$ which considers salience, coherence and redundancy as an index as to how suitable the extracted sentence sequence Seq is as a summary. This model is a weighted longest path problem with a fixed length.

This is an NP-hard problem. Due to the time cost, we adopt the simple randomized algorithm as shown in Algorithm 1 to obtain an approximated solution. Other decoding algorithms like greedy algorithms

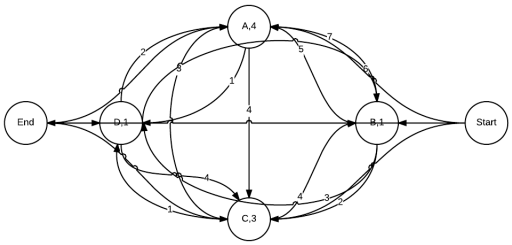
Algorithm 1 A randomized algorithm for the weighted longest path problem

```
Initialization:
Set U ⟵ all the sentences in the current doc set
Set S ⟵ EmptySet
Queue Q ⟵ EmptySet
repeat
  randomly select sentence s ∈ U & s ∉ Q;
  if length(s) + Σ_i length(s_i) <= threshold, s_i ∈ Q then
    push s to the rear of Q
  else
    push Q into S, Queue Q ⟵ EmptySet
  end if
until 10K times
return argmax_Q F(T), Q ∈ S
```

can also be employed. But none of them are capable of obtaining an exact solution. Below we present another model considering *local* coherence.

[3] "Global" means the model considers coherence according to the *whole* text.

Figure 1: A Complete Graph with Dummy Start and End Nodes



3.3 Summarization Considering Local Coherence

The above model considers global coherence which is calculated according to the whole text. The model presented below is directly based on local coherence and enables us to obtain an exact solution. We want to maximize $Score(Seq)$:

$$\begin{aligned} Score(Seq) &= \sum_{S_i \in Seq} (\alpha \sum_{e_k \in S_i} a_k + (1-\alpha) gain_{i,(i+1)}) \\ s.t. \quad & \sum_{S_i \in Seq} length(S_i) \leq threshold \end{aligned} \tag{2}$$

This formulation contains two parts. $\sum_{e_k \in S_i} a_k$ implies the weight of Sentence S_i, which is the sum of its containing entities' weights. $gain_{i,(i+1)}$ is the gain score for $Edge(S_i, S_{i+1})$. α manipulates the impacts of the two parts.

$$gain(S_i, S_{i+1}) = \sum_{e_k \in S_i \bigcup S_{i+1}} p_{e_k}(r_i^k r_{i+1}^k) \tag{3}$$

As is stated, r_i^k is the state of Entity e_k in Sentence S_i.

For the convenience of decoding, we turn the above model to an integer linear programming (ILP) problem. We add two dummy nodes, called $Start$ and End Node. All paths start from $Start$ and end with End. The costs of both $Start$ and End are 0. The gains of edges connected with $Start$ or End are 0. Note that although here we present a full connected graph for simplicity, in reality we deleted several edges to reduce redundancy. Following such a setting, an arbitrary path in the old graph (the one without dummy Start and End nodes) can be represented as a path from $Start$ to End. We write the $Start$ node as Node 0 and the End node as Node t. Then we formulate the problem of the weighted longest path as follows:

$$maximize \alpha \sum_i (\sum_{e_k \in S_i} a_k) x_i + (1-\alpha) \sum_{i,j} gain_{i,j} y_{ij}$$

subject to

$$\begin{cases} 1) \sum_i cost_i x_i \leq threshold \\ 2) \sum_i y_{0i} = 1 \\ 3) \sum_i y_{it} = 1 \\ 4) \sum_i y_{ij} + y_{0j} - (\sum_i y_{ji} + y_{jt}) = 0, \forall j \\ 5) \sum_i y_{ij} + y_{0j} - x_j = 0, \forall j \\ 6) x_i \in \{0,1\}, \forall i \\ 7) y_{ij} \in \{0,1\}, \forall i,j \end{cases} \tag{4}$$

Equations 2 and 3 are used to ensure we have only one start and one end node. Equation 4 ensures that the in degree equals the out degree for all nodes. Equation 5 ensures that the in degree is either 0 or 1 and equals x_a for all nodes. $x_i = 1$ indicates that S_i is selected for the summary. $x_i = 0$ means S_i is

not contained in the summary. $y_{ij} = 1$ means S_i and S_j are selected and placed as adjacent sentences in the summary. $cost_i$ is the number of words in S_i (length of S_i).

We resolve this ILP problem using the dual simplex method provided by IBM CPLEX optimizer [4] which is a powerful optimization software package. CPLEX provides both a primal simplex method and a dual simplex method for ILP problems. Here we adopt the latter.

4 Experiments & Analysis

4.1 Experiment

Experiments are conducted on the data set of the DUC2004 Summarization Task, which is a multi-document summarization task. 50 document clusters, each of which consists of 10 documents, are given. One summary is to be generated for each cluster. The target length is up to 100 words. Weights of entities are learnt by logistic regression as is adopted by Takamura and Okumura (2009) [5]. For entities that are not contained in DUC2003, we assign tf-based weights to them as Barzilay and Lapata (2008) did.

For the evaluation we firstly use the generally acknowledged metric for summarization: ROUGE metric. It essentially calculates n-gram overlaps between automatically generated summaries and human written (the gold standard) summaries. A high level of overlap indicates a high level of shared information between the two summaries. Among others, we focus on ROUGE-1 in the discussion of the result, because ROUGE-1 has proved to have a strong correlation with human annotation (Lin, 2004).

Some necessary preprocessing includes stemming, removing stop-words and simple simplification. In previous work, there is usually no co-reference resolution and different words are regarded as different entities. Here we use Stanford CoreNLP toolkit (Manning et al., 2014) to deal with the co-reference problem. The Stanford CoreNLP toolkit contains a ready-to-use entity identification tool and a co-reference resolution tool. The co-reference resolution is not a must, though preferred if reliable tools are available.

After the co-reference resolution, different forms of the same entities are replaced by their unified forms. For each document set, we need to estimate the transition probabilities for each entity according to the documents contained in the cluster as stated above.

Parameters are tuned using the DUC2003 dataset. d is the threshold of redundancy. We keep d percent of all edges and d varies from 10 to 100 with an interval of 10. We tune the parameter using the randomized algorithm and evaluate the results using ROUGE-1 Recall. In the following experiments, we set $d = 80$, which means we keep 80% of the sentences.

As for the model presented in Section 3.3, we need to tune α. Using the same data, we try α from 0 to 1 with an interval of 0.1 and eventually choose $\alpha = 0.4$.

4.2 Evaluation & Discussion

We compare our models with state-of-the-art multi-document summarization systems using ROUGE and human evaluation. The former aims to evaluate informativeness and the latter targets readability.

ROUGE Evaluation MCKP is the maximum coverage methods proposed by Takamura and Okumura (2009). Lin is a model that uses a class of submodular functions (Lin and Bilmes, 2011). Christ is a graph based model proposed by Christensen et al. (2013). DPP is the determinantal point processes model Borodin (2009) and ICSI is another model based on maximum coverage Gillick et al. (2008). The results of DPP and ICSI comes from the repository presented in Hong et al. (2014). M1 is our model described in Section 3.1. M2 is the model described in Section 3.3, which is resolved using an ILP method. MEAD Radev et al. (2004a) is a baseline that employs ranking algorithms to generate multi-document summaries.

The results are shown in Table 1. As we can see, our system (M1 and M2) produces comparable results to the state-of-the-art systems. With the MCKP method, all content words are used as concepts. But in

[4] http://www-03.ibm.com/software/products/en/ibmilogcpleoptistud/

[5] This method was first proposed by Yih et al. (2007) and then improved by Takamura and Okumura (2009). Here we follow the same steps with Takamura and Okumura (2009).

our systems, only nouns and pronouns are regarded as entities. There are fewer nouns and pronouns than content words. This has a negative impact on the evaluation of information coverage. But according to the experiment results, our approach still obtain satisfying results based on these entities. It proves that even with much simpler feature settings of just nouns and pronouns, the proposed model generates summaries with good coverage of the important information in source documents. We have addressed that ROUGE is merely an index of informativeness and cannot evaluate our improvements in readability as has been proved by *Christ*, another coherence-focused model (Christensen et al., 2013). So we also conduct a human evaluation.

Human Evaluation As some of the systems mentioned in Table 1 are not accessible, in this work we compare summaries produced by some typical systems: M2 (the best proposed system evaluated by ROUGE), MCKP (one of the state-of-the-art salience-focused methods) and humans (the gold standard).

We asked four professional annotators (who are not the authors of this paper and have rich experience in annotating various NLP tasks and are fluent in English) to assign a score to each summary regarding its readability. We randomly selected 48 summaries (16+16+16) from the three systems, and asked them to assign a readability score to each document without reading the source documents (summarization is useful because we do not need to read source documents). The score is an integer between 1 (very poor) and 5 (very good).

The average scores for the 3 systems are $Human = 4.3; M2 = 3.5; MCKP = 3.1$. Significance testing (significance level $\alpha = 0.05$) shows that the summaries generated by the proposed method show improvements in readability compared with previous salience-focused work.

Type	SysName	ROUGE-1(R)
Simple Ranking	MEAD	.339
Maximum Coverage	MCKP	.385
	ICSI	.384
Point Process	DPP	.398
Sub Modular	Lin	.394
Discourse-based	Christ	.373
	M1	.383
	M2	.390

Table 1: ROUGE Results on DUC2004

In our model, we assume the states of entities can be formulated as Markov chains. Although sophisticated models can be employed, such assumptions help simplify the model and they are proved to be of use. Also we can use more or fewer grammatical roles for entities. We tried using just two kinds of roles: presence and absence, and the performance we obtained was unsatisfying.

5 Related Work

A summary is much shorter than the original documents but still needs to provide readers with sufficient information. Hence the summarization systems need to identify important information and keep as much of it as possible. Most existing research follows such a guideline and takes salience as its sole focus.

Salience-focused systems cannot guarantee the readability of the generated text as they fail to take coherence into consideration. Sentence reordering, as a post processing task has began to develop. Apparently, it cannot make up for the flaws of salience-focused systems because it is simply a reorganization of sentences. Besides, it also faces problems when dealing with temporal text (Yan et al., 2011; Ge et al., 2015). A better solution is to consider coherence when selecting sentences. Such comprehensive models have been proposed. Most of them are discourse driven and sacrifice informativeness for coherence. In this sense, our model is novel in dealing with coherence without discourse analysis.

5.1 Salience-Focused Method

As stated, the summarization systems need to identify the important information and keep as much of it in the generated summaries as possible. One straightforward method is Maximum Marginal Relevance (Carbonell and Goldstein, 1998) (MMR). It is a greedy method, and is proposed to select sentences that are most relevant but not too similar to the already selected ones. It tries to keep a balance between relevance and redundancy. MMR is also widely employed to avoid redundancy in summarization systems. Among existing research, one popular kind is the ranking method (e.g., Textrank (Mihalcea and Tarau, 2004), Lexrank (Erkan and Radev, 2004) and its variants (Wan et al., 2007; Wang et al., 2012)), which construct a graph between text units and use ranking algorithms to select top sentences to build summaries. Another kind is the optimization method. Our work is one of this kind. It formulates summarization as finding a subset that optimizes certain objective functions without violating certain constraints. To find such an optimal subset is a combinatorial optimization problem, which is an NP hard problem and hence cannot be solved in linear time (McDonald, 2007).

Recently, maximum coverage methods have been proposed and yield good results (Gillick et al., 2009; Gillick and Favre, 2009; Takamura and Okumura, 2009). Maximum coverage methods formulate summarization as a maximum knapsack problem (MKMC). In MKMC methods, the meanings of sentences are believed to be made up by concepts, which usually refer to content words. And summarization involves extracting a subset of sentences that covers as many important concepts as possible without violating the length constraint. It is usually formulated as an integer linear problem. And some algorithms are proposed for obtaining approximated solutions (Takamura and Okumura, 2009; Gillick et al., 2009). Lin and Bilmes (2011) design a class of submodular functions for document summarization. The functions they use combine two parts, encouraging the summary to be representative of the corpus, and rewarding diversity separately. Other methods that have been applied to summarization include centroid-based methods (Radev et al., 2004b; Saggion and Gaizauskas, 2004), and minimum dominating set methods (Shen and Li, 2010). All these methods suffer in coherence.

5.2 Coherence-Focused Method

Sentence reordering methods are developed to correct the salience-focused models. Sentence reordering tries to generate a more coherent text by reordering its contents. Rich semantic and syntactic features are used to find a better permutation for input sentences (Barzilay et al., 2001; Bollegala et al., 2010; Okazaki et al., 2004).

The drawback to sentence reordering is obvious. The preceding sentence selection focuses solely on informativeness and totally neglects coherence. Thus it prevents the improvements expected from permutation. This is confirmed by the fact that the above methods all reports limited improvement. A consideration of coherence during sentence selection leads to new methods, and these are mainly discourse driven models. Some of the summarization methods encode discourse analysis results in feature presentations together with other frequency based features for sentence selection/compression. The problem is that these discourse based features usually play secondary roles, because the models all try to improve information coverage, which are evaluated by ROUGE. And ROUGE, as is commonly known, is not sensitive to coherence.

Some others work directly on discourse analysis results, and they usually try to derive a passage from a given parse tree. The problem of summarization is regarded as finding a text T so that $T = \arg\max F(T|Tr)$ for a given tree Tr. Here F is the objective function. Early representative work of this kind includes that of Marcu (1998) and that of Daumé III and Marcu (2002). Recently, Hirao et al. (2013) has viewed summarization as a knapsack problem on trees, and uses an integer linear problem (ILP) to formulate it. A sub tree that maximizes some objective function and obeys some given constraints is extracted from the original parse tree as the summary.

Discourse tree based methods cannot be extended to multi-document summarization. Christensen et al. (2013) propose a graph model that bypasses the tree constraints. They build a graph to represent discourse relations between sentences and then extract summaries accordingly.

Recently the neural network based discourse analysis (Li et al., 2014; Ji and Eisenstein, 2014) provides

us with an alternative way of conducting discourse analysis without traditional feature engineering. It can be used in our future work of modelling coherence using semantic relations.

6 Conclusion

Previous summarization methods have usually focused on salience and neglected coherence. This work proposed a novel summarization system that combines coherence with salience. By taking entities and links between them into consideration, our weighted longest path model successfully improves the quality of summaries. The proposed model does not require discourse analysis and hence can be applied to languages which do not enjoy plenty of ready-to-use discourse analysis tools.

In this paper only syntactic linkages are used for modelling coherence. In the future, we can take advantage of the semantic relations between entities to evaluate coherence and to further improve our system.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Regina Barzilay, Noemie Elhadad, and Kathleen R McKeown. 2001. Sentence ordering in multidocument summarization. In *Proceedings of the 1st International Conference on Human Language Technology Research*, pages 1–7. Association for Computational Linguistics.

Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. 2010. A bottom-up approach to sentence ordering for multi-document summarization. *Information Processing & Management*, 46(1):89–109.

Alexei Borodin. 2009. Determinantal point processes. *arXiv preprint arXiv:0911.1153*.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Proceeding of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics ? Human Language Technologies*, pages 681–684. Association for Computational Linguistics.

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

Janara Christensen, Stephen Soderland Mausam, and Oren Etzioni. 2013. Towards coherent multi-document summarization. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 1163–1173.

Hal Daumé III and Daniel Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting of Computational Linguistics*, pages 449–456. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based extractive summarization. In *Proceedings of the 42nd Annual Meeting of Computational Linguistics Workshop on Summarization*, volume 111.

Katja Filippova and Michael Strube. 2007. Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the 11th European Workshop on Natural Language Generation*, pages 139–142. Association for Computational Linguistics.

Tao Ge, Wenzhe Pei, Heng Ji, Sujian Li, Baobao Chang, and Zhifang Sui. 2015. Bring you to the past: Automatic generation of topically relevant event chronicles. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL2015)*.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18. Association for Computational Linguistics.

Dan Gillick, Benoit Favre, and Dilek Hakkani-Tur. 2008. The icsi summarization system at tac 2008. In *Proceedings of the Text Understanding Conference*.

Daniel Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2009. A global optimization framework for meeting summarization. In *Acoustics, Speech and Signal Processing, 2009. IEEE International Conference on*, pages 4769–4772. IEEE.

Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520. Association for Computational Linguistics.

Kai Hong, John M Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *LREC*, pages 1608–1616.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of Computational Linguistics*, pages 13–24.

Jiwei Li and Eduard H Hovy. 2014. A model of coherence based on distributed sentence representation. In *EMNLP*, pages 2039–2048.

Jiwei Li and Dan Jurafsky. 2016. Neural net models for open-domain discourse coherence. *arXiv preprint arXiv:1606.01545*.

Jiwei Li and Sujian Li. 2013. Evolutionary hierarchical dirichlet process for timeline summarization. In *ACL (2)*, pages 556–560. Citeseer.

Jiwei Li, Rumeng Li, and Eduard H Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2061–2069.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies*, pages 510–520. Association for Computational Linguistics.

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

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Computational Linguistics: System Demonstrations*, pages 55–60.

Daniel Marcu. 1998. Improving summarization through rhetorical parsing tuning. In *The 6th Workshop on Very Large Corpora*, pages 206–215.

Ryan McDonald. 2007. *A study of global inference algorithms in multi-document summarization*. Springer.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, volume 4, page 275. Barcelona, Spain.

Naoaki Okazaki, Yutaka Matsuo, and Mitsuru Ishizuka. 2004. Improving chronological sentence ordering by precedence relation. In *Proceedings of the 20th International Conference on Computational Linguistics*, page 750. Association for Computational Linguistics.

Karolina Owczarzak and Hoa Trang Dang. 2011. Overview of the tac 2011 summarization track: Guided task and aesop task. In *Proceedings of the 2011 Text Analysis Conference*.

Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, et al. 2004a. Mead-a platform for multidocument multilingual text summarization. In *Proceedings of the 4th Language Resources and Evaluation Conference*. Language Resources and Evaluation Conference.

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

Horacio Saggion and Robert Gaizauskas. 2004. Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the Document Understanding Conference*, pages 6–7.

Chao Shen and Tao Li. 2010. Multi-document summarization via the minimum dominating set. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 984–992. Association for Computational Linguistics.

Hiroya Takamura and Manabu Okumura. 2009. Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 781–789. Association for Computational Linguistics.

Marilyn A Walker, Aravind Krishna Joshi, and Ellen Friedman Prince. 1998. *Centering theory in discourse*. Oxford University Press.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Annual Meeting-Association for Computational Linguistics*, volume 45, page 552.

Xun Wang, Lei Wang, Jiwei Li, and Sujian Li. 2012. Exploring simultaneous keyword and key sentence extraction: improve graph-based ranking using wikipedia. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 2619–2622. ACM.

Xun Wang, Yasuhisa Yoshida, Tsutomu Hirao, Katsuhito Sudoh, and Masaaki Nagata. 2015. Summarization based on task-oriented discourse parsing. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23:1358–1367.

Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. 2011. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 745–754. ACM.

Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. 2007. Multi-document summarization by maximizing informative content-words. In *IJCAI*, volume 7, pages 1776–1782.