Using Rhetorical Topics for Automatic Summarization

Natalie M. Schrimpf
Department of Linguistics
Yale University
natalie.schrimpf@yale.edu

Abstract

Summarization involves finding the most important information in a text in order to convey the meaning of the document. In this paper, I present a method for using topic information to influence which content is selected for a summary. Texts are divided into topics using rhetorical information that creates a partition of a text into a sequence of non-overlapping topics. To investigate the effect of this topic structure, I compare the output of summarizing an entire text without topics to summarizing individual topics and combining them into a complete summary. The results show that the use of these rhetorical topics improves summarization performance compared to a summarization system that incorporates no topic information, demonstrating the utility of topic structure and rhetorical information for automatic summarization.

1 Introduction

Summarization is the task of creating a shortened version of an input document that retains the important information from the original text but in a more concise form. The goal of summarization is to convey the main concepts of the original document so that a summary user can understand what the document is about without reading the entire text. With large amounts of text available online, it has become increasingly necessary to find ways to allow people to quickly and easily find the information they need. Summarization is useful for this task because it condenses information into a shorter form that can be read instead of a longer text if it provides all the information a user needs or it can be read in order to determine whether the original text contains information relevant to the user's needs, allowing the user to decide which texts would be most useful. In order for summaries to achieve this goal, they must convey the important concepts from the text without including unnecessary information. Most summarization systems perform extractive summarization, which involves creating a summary by extracting complete sentences from the original document (Yih et al., 2007; Conroy et al., 2006; Wong et al., 2008; Christensen et al., 2013). The current research is also focused on extractive summarization.

Different representations of texts and text structure make different assumptions about how texts convey information and how summarization is performed. Much work on summarization does not assume anything about the structure of text. The work in this paper aims to demonstrate that attention to the linguistic structure of a text is useful in performing summarization. The type of linguistic structure and textual organization explored in this work is the notion of topic. In linguistics, there are different notions of what it means to be a topic (Lambrecht, 1996; Gundel, 1988; Blei, 2012; Griffiths et al., 2005; Van Dijk, 1977; Van Kuppevelt, 1995; Asher, 2004). Intuitively, texts are organized into topics or groups of sentences that are more related to each other than they are to sentences in other groups. A summary should include coverage of all these topics. One crucial factor that motivates grouping texts into topics for summarization has to do with summary length. A summary is a condensed form of the original text. One of the challenges of summarization is determining how to convey the same information as the original text in a more limited space.

Proceedings of the Society for Computation in Linguistics (SCiL) 2018, pages 125-135.
Salt Lake City, Utah, January 4-7, 2018

In order to convey the same information, there should be an emphasis on covering the text by including some amount of information about all of the important ideas and by limiting redundancy and in-depth coverage of a particular topic in favor of wider coverage of all topics.

In order to see how useful topics are for summarization, topic information was incorporated into a summarization system. To compare the effects of using topics versus not using topics, summarization was either performed at the level of the whole text or at the level of individual topics. Specifically, the process for incorporating topics into summarization included the following steps: divide a text into topics, summarize the text of each topic, and concatenate the summaries of each topic to create a summary for the whole text. With this method, topics are treated as independent pieces of text that contribute to the overall meaning of the text, and each topic will be represented in the final summary. This agrees with the intuition that texts can be divided into topics and a good summary should contain coverage of all topics that appear in the original text. This is a straightforward way to see how topics affect summarization.

Section 2 describes how texts are separated into topics using rhetorical information. Section 3 describes the techniques used for summarizing texts. Section 4 presents the methods and results of the experiments that were performed. Section 5 summarizes the findings and contributions of this work.

2 RST Topics

2.1 Rhetorical Structure Theory

Rhetorical Structure Theory (RST) is a framework for describing the organization of a text and what a text conveys by identifying hierarchical structures in text (Mann and Thompson, 1988). Pieces of text relate to each other in different ways in order to accomplish the writer's purpose with some pieces more central than others. Beginning with the clause level, relations can hold between successively larger spans of text forming a hierarchical structure of how all spans of the text are related to each other. One intuition behind RST is that the text structure itself conveys information beyond the information explicitly asserted by clauses in the text. Relations connect two non-overlapping pieces of text, and their combination conveys information beyond that of the individual clauses, such as the relational proposition that the information from one clause is evidence for the information in the other clause. Relation types include enablement, circumstance, background, justification, and evidence, among others. An important part of RST relations is the distinction between nuclei and satellites. The nucleus of a relation is one of the spans of text connected by the relation that is more essential to the purpose of the writer and is comprehensible on its own without the satellite. The satellite is the element that generally cannot appear on its own but provides some type of supporting information for the nucleus. The satellite of the relation provides information that will increase the belief in the nucleus of the relation, and the effect is that the reader has an increased belief in the information given in the nucleus. In this example of an evidence relation from Mann and Thompson the information in the satellite about the program producing the correct calculation provides evidence for the nucleus which states that the program works.

> Nucleus: The program as published for calendar year 1980 really works.
>
> Satellite: In only a few minutes, I entered all the figures from my 1980 tax return and got a result which agreed with my hand calculations to the penny.

The evidence relation and relations in general capture information about how different pieces of the text connect to each other and work together to achieve the writer's purpose.

Past work has considered whether rhetorical information is useful for summarization (Marcu, 2000; Chengcheng, 2010; Cardoso et al., 2015; Goyal and Eisenstein, 2016). Marcu (2000) explores how to use RST structures by combining the hierarchical structure of RST with the nucleus/satellite distinction to create an ordering of the units in the text based on importance and salience. Louis et al. (2010) explore the usefulness of different features, including discourse features, for selecting content in extractive summarization. Among the discourse-based features, there are some that score text units based on how high in the discourse tree they are promoted, and others that penalize satellite units relative to nucleus units. Chen et al. (2015) combine topic information and rhetorical structure into a single model, recognizing the importance of both of these types of knowledge for understanding the structure of a document. Specifically, their goal is to model a document's intent structure by assuming that documents contain two types of words: topic words and rhetorical words. In the current work, rhetorical information is used as part of determining the division of texts into sections corresponding to topics.

2.2 Proposal for using RST Topics

As rhetorical and structural information has been shown to be useful for tasks such as summarization, there is motivation for combining this information with the idea of topics. Additionally, instead of using it directly to determine which sentences to choose for a summary, I use this information at a different point in the process. I propose using RST to inform the division of texts into topics. RST relations capture how parts of a text connect to each other to accomplish the writer's purpose. They therefore provide useful information about which sentences in a text are most closely related to each other in a more structural sense than comparing the words they contain. Grouping sentences according to how they are related in a rhetorical structure provides a way to divide texts into topics. Specifically, topic relations that indicate a change in topic provide a natural grouping of sentences.

A few RST relation types are related to topics and topic changes within the text. These types are topic shift and topic drift. Topic shift is a relation that connects large sections of text when there is an abrupt change between topics. On the other hand, topic drift is a relation that connects large sections of text when the change between topics is smooth rather than abrupt, and there is still some similarity between topics. These topic relations provide a way to partition texts into topics. Specifically, these relations can be used as dividing points, with the sections of text connected by these relations considered distinct topics. In this paper, I explore a notion of topic based on this type of structural information and the utility of these topics for summarization.

3 Summarization System

In order to test how the use of topic structure affects summarization, I explored the impact of topics on the performance of previously proposed algorithms for extractive summarization that are implemented in the Sumy Python library.[1] Specifically, several common summarizers including LexRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004), and SumBasic (Nenkova and Vanderwende, 2005) were used. These summarization methods depend on word frequency and co-occurrence rather than using any substantial information about text structure.

Given these summarizers, the following process for summarization was used. First, the complete text was summarized by one of these summarizers. Texts were then divided into topics according to the topics in the RST annotation of the texts. Then each topic was summarized, and the outputs were combined to create a summary of the whole text. In each case, a value of 20% was used for the summarization, meaning the summarizer would return 20% of the original text, where length is measured in sentences. For example, for a text containing 10 sentences, the summarizer would return 2 as the summary. The value of 20% was used when summarizing the entire text or when summarizing an individual topic. Ideally, this will result in similar length summaries whether or not topics are used, because taking 20% of several smaller sections and combining them should be the same as taking 20% of the entire text. In addition to 20%, three other values of the summarization percentage were tested: 10%, 30%, and 40%.

4 Experiments

Experiments were conducted to see how topic structure influences summarization performance. Three conditions were tested. The first condition did not incorporate topic structure. Entire texts were summarized using the summarizers described in the previous section. The second condition used RST topics. Texts were divided into topics as described above. Each of the topics was summarized, and the outputs were combined to create a summary of the entire text. The third condition used random topics. Using the topic sizes from the RST topics, texts were randomly divided into topics of the same size. This condition provided a control to see whether topic divisions informed by RST information resulted in better summaries than random divisions or whether simply dividing a text into smaller sections improves performance.

[1] https://github.com/miso-belica/sumy

4.1 Data

The data for these topic summarization experiments comes from the RST Discourse Treebank (Carlson et al., 2002). This corpus contains 385 Wall Street Journal articles that have been annotated with RST structure. Dividing a text into RST topics depends on the presence of topic relations in the annotated text, specifically topic-shift or topic-drift relations. Not all texts in the corpus include topic relations in their annotations. Therefore, these experiments were limited to texts that do contain topic relations. In the corpus, there are 71 documents with topics.

Another feature of the RST Discourse Treebank is the presence of summaries for some documents. Gold-standard summaries are crucial for evaluating the output of a summarization system. For 150 documents in the corpus, there are 2 manually-created extractive summaries. Two analysts created these extracts by selecting a number of Elementary Discourse Units (EDUs) based on the square root of the total number of EDUs in the text. EDUs are the building blocks of RST structure. They are the lowest level units that are arguments of RST relations. EDUs are typically clauses.

Since gold-standard summaries are required to evaluate system-produced summaries, these experiments were performed on texts that have corresponding summaries. Of the 71 documents in the corpus that have topics, 51 documents also have extractive summaries. These 51 documents are the core dataset for the topic summarization experiments.

4.2 Division into Topics

Texts are divided into topics using RST topic relations. In the most straightforward case, all units in the text are explicitly designated as part of a topic. This case can be seen in Figure 1. The first line indicates that the text contains 31 units, as the Root spans the entire text. The next line shows that units 1-13 are part of a topic-drift relation. Skipping down to the other argument of this relation on the last line shows that the other element of the topic relation includes the rest of the text, units 14-31. In this notation, these text spans are arguments of the same relation because when combined they form a continuous sequence with the second argument starting directly after the first, and visually the two arguments occur at the same indent level. Therefore, this text can easily be divided into two topics. The first topic begins with the first unit of the text and continues to unit 13, and the second topic begins at unit 14 and continues to the end of the text.

```
( Root (span 1 31)
  ( Nucleus (span 1 13) (rel2par Topic-Drift)
    ( Nucleus (span 1 8) (rel2par span)
                    .
                    .
  ( Nucleus (span 14 31) (rel2par Topic-Drift)
```

Figure 1: RST annotation with all units in explicit topics

However, in other texts, not all units within a text are necessarily included as part of an explicit topic relation. In these cases, in order to divide a text into topics the topic relations were used as dividing points. Each occurrence of a topic relation signaled the beginning of a new topic. Anything before that point is grouped together as a topic, and anything after a topic relation is grouped as a topic. In that way, all units in a text are included as part of a topic. This topic division can be seen in Figure 2. The first line shows that the text contains 88 units. In contrast to the previous example, the first relation is not a topic relation. The first explicit topic relation begins with unit 7 and ends with unit 38. The other argument of that relation begins with unit 39 and continues to unit 53. No explicit topic relations include either the beginning or the end of the text. Using the topic relations as dividing points, all units can be placed into a topic. Units 1-6 become a topic, spanning from the beginning of the text to the first topic relation, and units 54-88 become a topic, spanning from the end of a topic relation to the end of the text.

This division method creates a partition of the text into a sequence of non-overlapping topics. Using this method means that topics contain adjacent units. Each unit in the text is contained in exactly one topic. Pseudocode for the topic division process is shown in Figure 3.

```
( Root (span 1 88)
  ( Nucleus (span 1 53) (rel2par span)
    ( Nucleus (span 1 6) (rel2par span)
      …
    ( Satellite (span 7 53) (rel2par background)
      ( Nucleus (span 7 38) (rel2par Topic-Drift)
          …
      ( Nucleus (span 39 53) (rel2par Topic-Drift)
          …
```

Figure 2: RST annotation without all explicit topics

```
input: RST annotation file
Topics = list of topics

# Find explicit topics in RST annotations
for line in annotations:
    relation = find label of relation type
    span = (x, y) where x is start and y is end
    if relation = 'topic-shift' or 'topic-drift':
        add span to Topics

# If no topic starting with the first unit, add one
minimum = lowest value in Topics
if minimum != 1:
    add (1, minimum-1) to Topics

# If no topic ending with the last unit, add one
total_len = total number of units
maximum = highest value in Topics
if maximum < total_len:
    add (maximum+1, total_len) to Topics

# Remove topics with overlapping starting or ending points to ensure sequence of non-overlapping topics
for (x, y) in Topics:
    if y is not smallest value for x:
        remove (x, y) from topics
    if there are multiple values of x for y:
        if x = lowest value:
            remove (x, y) from topics
            add to Topics (lowest value, w-1) where (w, z) in Topics and w-lowest value is smallest

# If unit not included in any topic, add it
for i from 1 to total_len:
    if i is not covered by any topic in Topics:
        add (i, i) to Topics

return Topics
```

Figure 3: Pseudocode for dividing into topics

4.3 Evaluation

Summary evaluation is a difficult task. There can be more than one good summary of a text, and when people are instructed to create summaries, they do not necessarily contain the same sentences. Since there is no single correct answer for what a summary should contain, evaluation typically involves comparing a system-produced summary to a manually-created reference summary. Summary quality is based on some measure of similarity or overlap with a reference summary.

ROUGE (Lin, 2004) is a measure to evaluate performance on the task of automatic summarization. ROUGE is a standard measure used in the field of summarization (Erkan and Radev, 2004; Lin and Hovy, 2003; Xie et al., 2008; Wong et al., 2008; Nallapati et al., 2016; Chopra et al., 2016). Recall-Oriented Understudy for Gisting Evaluation involves comparing a summary produced by a summarization system to reference or gold-standard summaries created by humans. Specifically, ROUGE-N measures n-gram (unigram, bigram, etc.) recall between a system summary and a reference summary. Recall refers to how many of the reference n-grams were included in the system summary. The equation for ROUGE-N is presented below.

$$\frac{\sum_{S\in\{Ref\ Sum\}}\sum_{gram\ n\in S} Count_{match}(gram_n)}{\sum_{S\in\{Ref\ Sum\}}\sum_{gram_n\in S} Count(gram_n)} \tag{1}$$

In this equation, n refers to the size of the n-gram, such as unigram (1) or bigram (2). An n-gram itself is represented by $gram_n$, and $Count_{match}(gram_n)$ refers to the number of times that the n-gram $gram_n$ appears in the system summary. Therefore, the numerator is the number of matching n-grams, and the denominator is the total number of n-grams in the reference summaries.

One downside of ROUGE is that it is entirely recall-based. In general, a summary will be rewarded for including more n-grams without being penalized for containing n-grams that do not appear in the reference summary. In the extreme case, a summary that is the same length as the original text being summarized could achieve perfect recall even though such a summary would clearly not be considered a good summary, since the goal of summarization is to produce a shortened version of the input. In order to avoid this problem, summary length must be controlled. Specifically, since ROUGE-N is a word-based evaluation measure, the summary length in terms of word count must be controlled so that system-produced summaries are similar in length to the reference summaries.

Unit overlap is another evaluation measure for summarization (Steinberger and Ježek, 2012). It finds the similarity between two texts by looking at

	LR	LR-T	TR	TR-T	SB	SB-T
Avg ROUGE-1	0.496	0.588	0.554	**0.607**	0.420	0.463
Avg ROUGE-2	0.330	0.442	0.415	**0.458**	0.214	0.275
Avg Unit Overlap	0.261	**0.317**	0.260	0.289	0.241	0.260
Avg Cosine Similarity	0.668	**0.711**	0.694	0.710	0.619	0.650

Table 1: Results of using the summarizers with and without topics. LR: LexRank, LR-T: LexRank with Topics, TR: TextRank, TR-T: TextRank with Topics, SB: SumBasic, SB-T: SumBasic with Topics. Highest values for each measure are in bold.

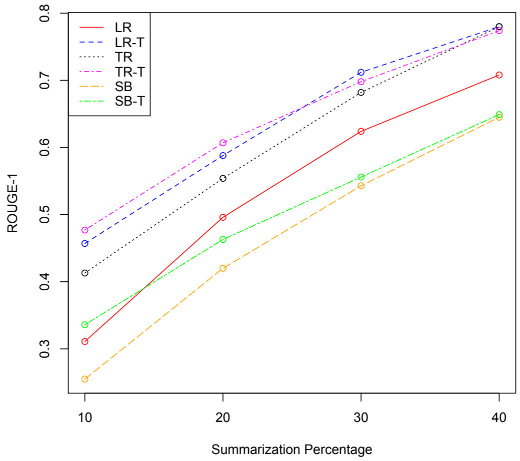

Figure 4: Values of ROUGE-1 as percentage increases

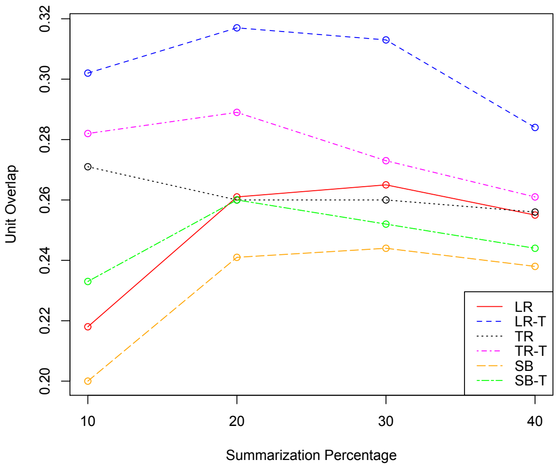

Figure 5: Values of unit overlap as percentage increases

the number of words they have in common compared to the number of non-overlapping words they contain.

$$unit\ overlap(X,Y) = \frac{\|X \cap Y\|}{\|X\| + \|Y\| - \|X \cap Y\|} \quad (2)$$

X and Y are the words in the documents being compared. In contrast to ROUGE, unit overlap penalizes an evaluated text for containing words that do not appear in the gold-standard text. A summary will not be rewarded simply for being longer.

The final evaluation measure used is cosine similarity (Steinberger and Ježek, 2012). It is a measure of similarity between documents using vectors of word frequency. Similar to unit overlap, cosine similarity takes document length into account and prevents texts from being rewarded for being longer.

To evaluate the summaries, each system-produced summary is evaluated against each of the two corresponding gold-standard summaries. Scores are calculated for each document, and the scores from all documents in the corpus are averaged to produce an overall value for each measure.

4.4 Results and Discussion

Table 1 shows the results of using three different summarizers to summarize texts with and without topics. These are the results when a summarization percentage of 20% was used. Each pair of columns shows the result of a different summarizer. The first column in each pair shows the results without using topics, and the second shows the results of using RST topics. The highest value for each measure is in bold. Looking at the results shows several interesting effects. For each measure, the highest value is achieved when using topics. The highest ROUGE values are found with TextRank, and the highest values for unit overlap and cosine similarity are found with LexRank. In these cases, using topics results in improvements in performance of around 5%. While different summarizers perform slightly better on different measures, in this paper I am interested in the fact that regardless of evaluation measure or summarizer, the inclusion of topics improves performance.

Text: [Nissan Motor Co. expects net income to reach 120 billion yen (U.S. $857 million) in its current fiscal year, up from 114.6 billion yen in the previous year, Yutaka Kume, president, said. Mr. Kume made the earnings projection for fiscal 1990, ending next March 31, in an interview with U.S. automotive writers attending the Tokyo Motor Show. The executive said that the anticipated earnings increase is fairly modest because Nissan is spending heavily to bolster its dealership network in Japan and because of currency-exchange fluctuations.]$_{\text{Topic 1}}$

[During the next decade, Mr. Kume said, Nissan plans to boost overseas vehicle production sufficiently to account for a majority of sales outside Japan. Last year, Mr. Kume said, Nissan exported slightly over one million vehicles, and produced 570,000 cars and trucks at its factories in North America, Europe and Australia. But by 1992, he added, Nissan will build one million vehicles a year outside Japan, or sufficient to equal exports. "By the end of the 1990s," he said, "we want to be producing roughly two vehicles overseas for every vehicle that we export from Japan." That will involve a substantial increase in overseas manufacturing capacity, he acknowledged, but didn't provide specific details.]$_{\text{Topic 2}}$

Summary without Topics: But by 1992, he added, Nissan will build one million vehicles a year outside Japan, or sufficient to equal exports. "By the end of the 1990s," he said, "we want to be producing roughly two vehicles overseas for every vehicle that we export from Japan."

Summary with RST Topics: Nissan Motor Co. expects net income to reach 120 billion yen (U.S. $857 million) in its current fiscal year, up from 114.6 billion yen in the previous year, Yutaka Kume, president, said. During the next decade, Mr. Kume said, Nissan plans to boost overseas vehicle production sufficiently to account for a majority of sales outside Japan.

Figure 6: Example text and summaries

Figures 4 and 5 show how the values of ROUGE-1 and unit overlap change as different summarization percentages are used. In general, values of the evaluation measures increase as the percentage increases. However, the increases depend on which measure is considered. ROUGE values increase by the largest margin and with the most consistency. On the other hand, there are smaller increases for the other measures. ROUGE has such large and consistent increases because it is recall-based, so longer summaries will always perform better. This issue will be discussed further below. The results show that topics create more of an improvement in performance when the percentage is lower and the summaries are smaller, suggesting that topics do provide useful information for summarization, and that information is the most useful when space is the most limited.

Figure 6 provides an example text along with the summaries produced when no topics are used and when RST topics are used. When no topic structure is used, all sentences in the summary come from one topic, showing how topic structure is needed to ensure coverage of all ideas in the text.

An important factor to consider when comparing the results of performing summarization with and without topics is summary length. It is possible that summarizing at the topic level could result in summaries of different lengths from the summaries produced by summarizing the entire text. Differences in length could affect these evaluation measures, particularly ROUGE, which is recall-based and therefore benefits from including more words by increasing the chances of having more words in common with the gold-standard.

One way of dealing with this potential problem is to compare RST topics with random topics. Using the topic sizes from the RST topics, texts were randomly divided into topics of the same size. While the RST topics always contain adjacent sentences, the random topics are not constrained in this way. If the topics were contiguous, they could not be both random and equal in size to the RST topics. Therefore, the random topics are equal in size but do not follow the same adjacency restrictions as the RST topics. Since the random topics are the same size as the RST topics, length should not have an effect. The model was run 25 times with random topics. Figure 7 provides a visual illustration of the results. The bars in the graph represent the mean values. The error bars show two standard deviations below and above the mean. The points represent the values when using RST topics. Comparing the mean values to the values with RST topics, the RST values are higher than the random topics for all measures. Looking at the RST values compared to the means + 2 standard deviations, the RST values are greater than or very similar to the random values, indicating that the RST values are significantly different from random.

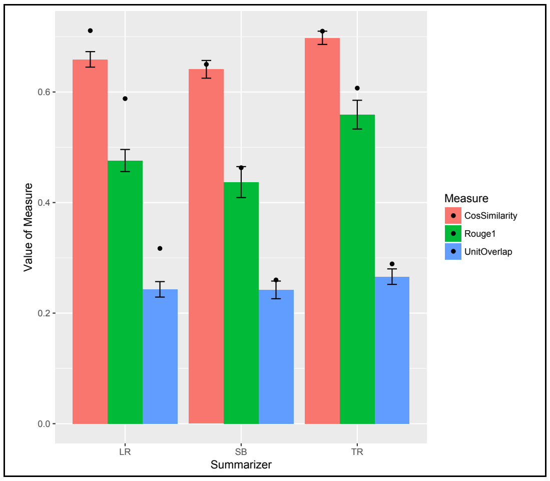

Figure 7: Results with random topics

ROUGE-1		
Factor	Estimate	P-value
Word Count	0.0004995	1.51e-08 ***
Random	-0.01885	0.31
RST Topics	0.08704	4.17e-06 ***
Unit Overlap		
Word Count	-0.0001083	0.0967
Random	-0.01616	0.2508
RST Topics	0.05705	6.32e-05 ***

Table 2: Linear Regression Results for LexRank

Another way to explore the effects of length is to consider whether the use of topics has an effect on performance separate from any effect of summary length. Table 3 shows the results of a linear regression exploring the effects of different factors on ROUGE and unit overlap to see whether length differences are having an effect. The factors considered were word count in the summary and RST Topics/Random Topics/No Topics. These are the results when using LexRank as the summarizer. The results for the other summarizers are similar.[2] As shown in the table, word count and RST topics had a significant effect on ROUGE, while random topics had no significant effect. The significance of RST topics shows that using RST topics improves performance compared to not using topics as well as compared to using random topics. As discussed above, ROUGE is affected by differences in word count, and these results show that word count was a significant factor in predicting ROUGE scores. However, the unit overlap scores are not affected in the same way, and the linear regression confirms that for unit overlap the only significant factor is the use of RST topics.

These results demonstrate the positive impact that the use of topics has on summarization performance. Specifically, dividing texts into topics using topic relations from RST results in summaries that are more similar to manually-created gold-standard summaries than summarizing texts without the inclusion of topic structure.

Given the improved summarization performance seen when using topics based on RST, it is worth considering whether other notions of topic, particularly common topic modeling methods, are useful for this task. Latent Semantic Analysis (LSA) was tested as another way to divide texts into topics, using an implementation from Gensim topic modeling software (Rehurek and Sojka 2010). An important part of LSA is the number of dimensions that are used when reducing the semantic space. Given the size of the training data and common values suggested in previous research, three values were tested for the number of dimensions: 50, 100, and 200. To divide a text into topics, the number of topics, n, to choose was taken from the RST annotation. Then this number was used as k in a k-means clustering algorithm. Clustering was performed over the sentence vectors that represent a document's sentences in the LSA semantic space. These vectors contain values for how related a document is to each of the dimensions in the model. Clustering divides the sentences of a text into k topics based on similarity of the sentence vectors. Clustering was performed using scikit-learn (Pedregosa et al. 2011). The results are presented in Table 3, which shows the mean of each evaluation measure over 10 runs of the model, using LexRank as the summarizer, with 100 dimensions in the LSA model. The values when not using topics and when using RST topics are repeated in the table for comparison.

[2] TextRank P-values, ROUGE: Word Count 2.21e-09 ***, Random 0.83744, RST topics 0.00345 **; Unit Overlap: Word Count 0.0713, Random 0.8942, RST topics 0.0170 *. SumBasic P-values, ROUGE: Word Count 1.02e-10 ***, Random 0.5335, RST topics 0.0121 *; Unit Overlap: Word Count 0.2439, Random 0.7664, RST topics 0.0505.

	R1	R2	UO	CS
Mean	0.4873	0.3105	0.2546	0.6720
No Top	0.496	0.330	0.261	0.668
RST	0.588	0.442	0.317	0.711

Table 3: Results when using LSA topics compared to no topics and RST topics; Mean = mean of LSA runs

The results are similar when using different numbers of dimensions as well as the other summarizers. Looking across all summarizers and numbers of dimensions, RST topics perform better than LSA topics. The difference in performance is evident for all evaluation measures. In general, LSA topics perform similarly to using no topics at all. Overall, the types of topics found by using LSA are not very useful for a summarization system that uses topics. The results suggest that while LSA has been successfully used to classify documents and find documents related to a query (Deerwester et al., 1990; Zelikovitz and Hirsh, 2001), LSA is not sufficient to distinguish between different topics of a single document and does not find topics that improve summarization.

5 Conclusion

By performing summarization at the level of the entire text and at the level of individual topics, I investigated the influence of topic information on summarization performance. I explored a notion of topic that uses information about a text's rhetorical structure in the form of RST relations. The direct comparison of summarization when using topics versus not using topics showed that topic information improves performance. Improvements were found with several evaluation measures, including ROUGE and unit overlap. Performance also improved regardless of which summarizer was used.

The strong performance of the model when using topics has several interesting implications that highlight the contributions of this work. First, the results demonstrate the usefulness of topic structure. Conceptualizing texts as composed of a number of topics not only improves human processing of texts but also increases the quality of summaries produced by automatic systems. In this work, topics were incorporated in a straightforward way, by summarizing a text's topics and combining them to create a complete summary. The results showed that this simple method for including topic information improves performance compared to not using any topics.

Another important finding of this work is the utility of a notion of topic based on rhetorical information. The topics were based on RST relations that connect pieces of a text when the topic has changed between the sections. Using these relations to signal boundaries between topics proved to be a reasonable method to automatically separate a text into its component topics, and specifically a method that is useful for finding topics relevant for summarization. The improvements in performance seen with this notion of topic also demonstrate another way that rhetorical information such as RST can be used as part of the summarization process.

These results demonstrate the benefits of using one notion of topic for summarization and motivate further investigation into the use of topic structure, including comparisons with other methods for dividing a text into topics. This work also motivates the consideration of other ways to use rhetorical information for summarization. An area for future work is to explore how to automatically find RST-type topics without requiring a full RST annotation.

References

Asher, Nicholas. 2004. Discourse topic. *Theoretical Linguistics*, *30*(2-3), pp.163-201.

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM*, *55*(4), pp.77-84.

Cardoso, Paula C.F., Maria L.R.C. Jorge, and Thiago A.S. Pardo. 2015. Exploring the Rhetorical Structure Theory for multi-document summarization. In *Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, XXXI*. Sociedad Española para el Procesamiento del Lenguaje Natural-SEPLN.

Carlson, Lynn, Mary Ellen Okurowski, and Daniel Marcu. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

Chen, Bei, Jun Zhu, Nan Yang, Tian Tian, Ming Zhou, and Bo Zhang. 2016. Jointly Modeling Topics and Intents with Global Order Structure. In *AAAI* (pp. 2711-2717).

Chengcheng, Li. 2010. Automatic text summarization based on rhetorical structure theory. In *International Conference on Computer Application and System Modeling (ICCASM)* (Vol. 13, pp. V13-595). IEEE.

Chopra, Sumit, Michael Auli, Alexander M. Rush, and S.E.A.S. Harvard. 2016. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In *HLT-NAACL* (pp. 93-98).

Christensen, Janara, Stephen Soderland Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards Coherent Multi-Document Summarization. In *HLT-NAACL* (pp. 1163-1173).

Conroy, John M., Judith D. Schlesinger, and Dianne P. O'Leary. 2006. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the COLING/ACL on Main conference poster sessions* (pp. 152-159). Association for Computational Linguistics.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), p.391.

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

Goyal, Naman and Jacob Eisenstein. 2016. A Joint Model of Rhetorical Discourse Structure and Summarization. *EMNLP 2016*, p.25.

Griffiths, Thomas L., Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2005. Integrating topics and syntax. In *Advances in neural information processing systems* (pp. 537-544).

Gundel, Jeanette K. 1988. Universals of topic-comment structure. *Studies in syntactic typology*, *17*, pp.209-239.

Hyona, Jukka, Robert F. Lorch, and Johanna K. Kaakinen. 2002. Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology*, *94*(1), pp.44-55.

Lambrecht, Knud. 1996. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents* (Vol. 71). Cambridge University Press.

Lin, Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop* (Vol. 8).

Lin, Chin-Yew and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 71-78). Association for Computational Linguistics.

Lorch Jr, Robert F., Elizabeth Pugzles Lorch, and Ann M. Mogan. 1987. Task effects and individual differences in on-line processing of the topic structure of a text. *Discourse Processes*, *10*(1), pp.63-80.

Louis, Annie, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 147-156). Association for Computational Linguistics.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, *8*(3), pp.243-281.

Marcu, Daniel. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.

Mihalcea, Rada and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *EMNLP* (Vol. 4, pp. 404-411).

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

Nenkova, Ani and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, *101*.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and Jake Vanderplas. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), pp.2825-2830.

Rehurek, Radim and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Steinberger, Josef and Karel Ježek. 2012. Evaluation measures for text summarization. *Computing and Informatics*, *28*(2), pp.251-275.

Van Dijk, Teun A. 1977. Sentence topic and discourse topic. *Papers in Slavic Philology*, *1*, pp.49-61.

Van Kuppevelt, Jan. 1995. Discourse structure, topicality and questioning. *Journal of linguistics*, *31*(1), pp.109-147.

Wong, Kam-Fai, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1* (pp. 985-992). Association for Computational Linguistics.

Xie, Shasha and Yang Liu. 2008. Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* (pp. 4985-4988). IEEE.

Yih, Wen-tau, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. 2007. Multi-Document Summarization by Maximizing Informative Content-Words. In *IJCAI* (Vol. 7, pp. 1776-1782).

Zelikovitz, Sarah and Haym Hirsh. 2001. Using LSI for text classification in the presence of background text. In *Proceedings of the tenth international conference on Information and knowledge management* (pp. 113-118). ACM.