The Role of Discourse Units in Near-Extractive Summarization

Junyi Jessy Li
University of Pennsylvania
ljunyi@seas.upenn.edu

Kapil Thadani, Amanda Stent
Yahoo Research
{thadani,stent}@yahoo-inc.com

Abstract

Although human-written summaries of documents tend to involve significant edits to the source text, most automated summarizers are extractive and select sentences verbatim. In this work we examine how *elementary discourse units* (EDUs) from Rhetorical Structure Theory can be used to extend extractive summarizers to produce a wider range of human-like summaries. Our analysis demonstrates that EDU segmentation is effective in preserving human-labeled summarization concepts within sentences and also aligns with near-extractive summaries constructed by news editors. Finally, we show that using EDUs as units of content selection instead of sentences leads to stronger summarization performance in near-extractive scenarios, especially under tight budgets.

1 Introduction

Document summarization has a wide variety of practical applications and is consequently a focus of much NLP research. When a human summarizes a document, they often edit its constituent sentences in order to succinctly capture the document's meaning. For instance, Jing and McKeown (2000) observed that summary authors trimmed extraneous content, combined sentences, replaced phrases or clauses with more general or specific variants, etc. These *abstractive* summaries thus involve sentences which deviate from those of the source document in structure or content.

In contrast, automated approaches to summarization generally produce *extractive* summaries by selecting complete sentences from the source document (Nenkova and McKeown, 2011) in order to ensure that the output is grammatical. Extractive summarization techniques, which are widely used in practical applications, therefore address a substantially simpler problem than human summarization.

This leads to a natural question: can extractive summarization techniques be used to produce more human-like summaries? We hypothesize that automated methods can generate a wider range of summaries by extracting over sub-sentential units of meaning from the source documents rather than whole sentences. Specifically, in this paper we investigate whether *elementary discourse units* (EDUs) from Rhetorical Structure Theory (Mann and Thompson, 1988) comprise viable textual units for summarization. Our focus is on recovering salient summary content under ROUGE (Lin, 2004) while the composition of EDUs into fluent output sentences is left to future work.

We investigate this hypothesis in two complementary ways: by studying the compatibility of EDUs with human-labeled summarization units from pyramid evaluations (Nenkova et al., 2007) and by assessing their utility in reconstructing real-world document previews chosen by news editors in the New York Times corpus (Sandhaus, 2008). The contributions of this work include:

- A demonstration that EDU segmentation preserves human-identified conceptual units in the context of document summarization.
- New, large datasets proposed for research into extractive and compressive summarization of news articles.
- A study of the lexical omissions made by news editors in real-world compressive summarization.
- A comparative analysis of supervised single-document summarization over full sentences and over a range of budgets in extractive and near-extractive scenarios.

Proceedings of the SIGDIAL 2016 Conference, pages 137–147,
Los Angeles, USA, 13-15 September 2016.

2 Background and related work

Discourse structure in summarization Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) represents the discourse in a document in the form of a tree (Figure 1). The leaf nodes of RST trees are *elementary discourse units* (EDUs) which are a segmentation of sentences into independent clauses, including dependencies such as clausal subjects and complements. The more central units to each RST relation are *nuclei* while the more peripheral are *satellites*. Prior work in document compression (Daumé and Marcu, 2002) and single-document summarization (Marcu, 1999; Louis et al., 2010; Hirao et al., 2013; Kikuchi et al., 2014; Yoshida et al., 2014) has shown that the structure of discourse trees, especially the *nuclearity* of non-terminal discourse relations in the tree, is valuable for content selection in summarization.

The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) on the other hand is theory-neutral and does not define a recursive structure for the entire document like RST. Discourse relations are lexically bound to explicit discourse connectives within a sentence or exist between adjacent sentences if there is no connective. Each relation is realized in two text arguments, which are similar to EDUs. However, unlike EDUs, PDTB relation arguments have flexibility in size, ordering and arrangement and do not form a complete *segmentation* of the text. They are therefore not easily interpretable as textual units that can be combined to form sentences and summaries.

In this paper, we focus on EDUs and explore their viability as basic units for summarization. We did not use PDTB-style arguments to make sure each part of a document belongs to a textual unit and that the units are strictly adjacent to each other. EDU segmentation, typically addressed as a tagging problem early in discourse parsing systems, has seen accuracy and speed improvements in recent years (Hernault et al., 2010; Joty et al., 2015). It is now practical to segment document sentences into EDUs at scale as a preprocessing step for automated summarization.

Textual units in summarization. In extractive summarization, sentences are typically chosen as units to assemble output summaries because of their presumed grammaticality (Nenkova and McKeown, 2011). Finer-grained units such as

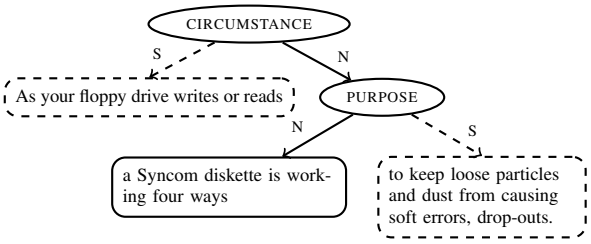

Figure 1: A RST discourse tree with EDUs as leaf nodes (example from Mann and Thompson (1988)).

n-grams are frequently used for quantifying content salience and redundancy prior to summarization over sentences (Filatova and Hatzivassiloglou, 2004; Thadani and McKeown, 2008; Gillick and Favre, 2009; Lin and Bilmes, 2011; Cao et al., 2015). In contrast, when the task at hand is more abstractive, the units are more fine-grained, e.g., n-grams and phrases in abstractive summarization (Kikuchi et al., 2014; Liu et al., 2015; Bing et al., 2015), n-grams and human-annotated concept units in summarization evaluation (Lin, 2004; Hovy et al., 2006). Recently, subject-verb-object triplets were used to automatically identify concept units (Yang et al., 2016) and in abstractive summarization (Li, 2015); however, this requires semantic processing while EDU segmentation is presently more accurate and scalable.

Here, we explore EDUs as a middle ground between fine-grained lexical units and full sentences. While EDUs have been used in prior work to directly assemble output summaries (Marcu, 1999; Hirao et al., 2013; Yoshida et al., 2014), the focus was on using discourse structure as features for sentence ranking, while our work is the first to examine the utility of EDUs themselves.

Datasets. In this work, we address *single-document* summarization. Standard datasets for the task were created for the Document Understanding Conference (DUC) in 2001 and 2002. The datasets for each year were composed of about 600 documents accompanied by 100-word abstractive summaries. In addition, the RST Discourse Treebank (Carlson et al., 2003) contains abstractive summaries for 30 documents, which have been used for evaluation in RST-driven summarization (Hirao et al., 2013; Kikuchi et al., 2014; Yoshida et al., 2014).

In contrast, we propose the use of datasets de-

[The European Airbus A380 flew its maiden test flight from France] [10 years after design development started.]

Figure 2: An EDU-segmented sentence with three human-labeled concepts (SCU contributors).

rived from the New York Times (NYT) corpus[1] that are orders of magnitude larger than the DUC dataset, featuring thousands of article summaries with varying degrees of extractiveness. Although the summaries in this dataset typically contain fewer than 100 words and are sometimes intended to serve as a teaser for the article rather than a distillation of its content, they were nevertheless created by professional editors for a highly-trafficked news website. Prior work has also demonstrated the utility of this corpus for summarization (Hong et al., 2015; Nye and Nenkova, 2015). This dataset therefore enables the study of summarization in a realistic setting.

Compressive summarization. To explore the utility of EDUs in summarization, we examine *near*-extractive summaries in the NYT corpus which are drawn from sentences in the document but omit at least one word or phrase from them. This setting is also explored in the summarization literature for techniques which combine extractive sentence selection with sentence compression (Clarke and Lapata, 2007; Berg-Kirkpatrick et al., 2011; Woodsend and Lapata, 2012; Almeida and Martins, 2013; Kikuchi et al., 2014). These approaches are typically evaluated against abstractive summaries and have not been studied with a natural compressive dataset such as the ones proposed here. We do not address techniques to generate compressive summaries in this work but instead attempt to quantify how the omitted content in a summary relates to its EDU segmentation.

3 EDUs as Concept Units in Summaries

We first investigate whether EDUs from an RST parse of the document can serve as a middle ground between abstract units of information and the sentences in which they are realized. Specifically, given a dataset containing human-labeled concepts in each article, we examine their correspondence with the EDUs extracted automatically from the article in terms of both lexical coverage and content salience.

[1]Available from the LDC at `https://catalog.ldc.upenn.edu/LDC2008T19`

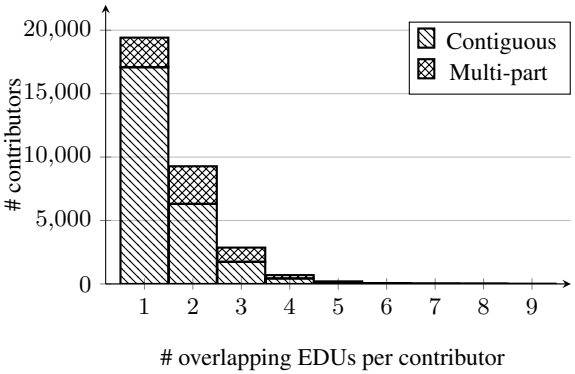

Figure 3: Number of EDUs which overlap with each SCU contributor (single or multi-part) in the DUC/TAC reference summary datasets.

3.1 Data and settings

In the DUC 2005–2007 and TAC 2008–2011 shared tasks on multi-document summarization, evaluations are conducted under the *pyramid* method—a technique which quantifies the semantic content of reference summaries and uses it as the basis of comparison for system-generated summaries (Nenkova et al., 2007). For this, human annotators must identify *summary content units* (SCUs) across reference summaries for a single topic. Each SCU has one or more *contributors* from different reference summaries which express the concept in text. Of the 32,535 contributors in the DUC and TAC data, 79% form contiguous text spans while the rest involve two or more non-contiguous parts within a sentence.

Our primary goal in this section is to investigate the degree to which EDUs correspond to SCUs. For this purpose, we treat each reference summary as an independent article and its SCU contributors as concept annotations. We parse the summaries using the RST parser of Feng and Hirst (2014a) to recover an EDU segmentation, specifically version 2.01 of the parser which shows superior EDU segmentation performance to other discourse parsers (Feng and Hirst, 2014b). An example of an EDU-segmented sentence with its human-labeled concepts is shown in Figure 2.

[The American Bookseller Association represents private bookstore owners] [and sponsors Book Expo, an annual convention.]

[Napster claimed protection under the Millenium Copyright Act] [because they had no control over users' actions.]

Figure 4: Examples of sentences in which human-labeled concepts (indicated by connected lines) span EDUs (in square brackets).

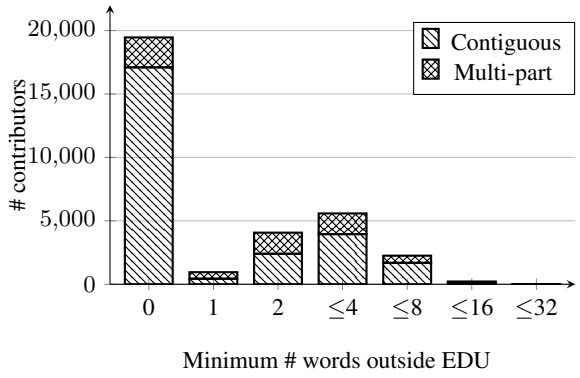

Figure 5: Number of words in SCU contributors which remain uncovered by a single EDU in the DUC/TAC reference summary datasets.

3.2 Concept coverage

Figure 3 indicates the number of EDUs that overlap by one or more tokens with each SCU contributor in the data. Most concepts (62%) are covered by a single EDU. This is more pronounced for concepts which are realized in a contiguous text span (69%), while multi-part concepts are unsurprisingly more likely to overlap with two EDUs. On average, concepts overlap with 1.56 EDUs while EDUs overlap with 1.77 concepts, significantly fewer than the average number of concepts contained in whole sentences (2.18).

Because we consider an overlap of one token to be sufficient to associate an EDU with an SCU contributor, we also examine in Figure 5 the number of non-punctuation contributor words that would need to be deleted for each concept to be covered by a single EDU. The vast majority of SCU contributors are covered by a single EDU, while the remainder typically have 2–4 words uncovered. Fewer than 8% of concepts were observed to have more than 4 words outside their corresponding EDU.

In Figure 4 we show typical examples of sentences with concepts which cross EDU boundaries. A major source for breached boundaries lies within heads of clauses. For instance, the first example contains two verb phrases in separate EDUs which each mark a concept, but their shared head "American Bookseller Association" can be in only one EDU. Errors are also often caused by overly broad SCUs which contain too much content. In the second example, the second EDU holds a causal relation with the first EDU and is thus a a satellite to the discourse relation, whereas the whole relation is combined into a single SCU contributor. These cases can potentially be resolved by taking into account the discourse relation and nuclearity status of the involved EDUs.

3.3 Salience via discourse structure

In addition to coverage of SCU contributors, we would like to see the extent to which EDUs are *meaningful* with respect to summarization concepts. One of the most intriguing aspects of EDUs is that they are not merely textual units but rather units in a discourse tree from which relative concept importance can be derived. In pyramid evaluations, the salience of an SCU is determined by the number of distinct contributors it has across all reference summaries for a topic, and thus each SCU in our dataset has an implicit weight indicating its importance. We therefore investigate the relationship between *inter*-document concept salience using these SCU weights and an *intra*-document counterpart from the EDUs in the discourse tree.

To calculate salience over EDUs, we use the scoring mechanism in Marcu (1999). Intuitively, each EDU which is a nucleus of a discourse relation (as opposed to a satellite) can be promoted one level up in the discourse tree. The score weights each EDU according to the depth that it can be promoted up to: the closer to the root, the more important the EDU is. For this analysis, we impute the discourse salience of a contributor by averaging the Marcu (1999) scores (normalized by tree depth) of the EDUs it overlaps with.

Table 1 shows the mean of these scores over all contributors with a particular SCU weight. In each group with weight w, the average EDU-derived

SCU weight	1	2	3	≥4
Proportion of SCUs (%)	54.3	21.6	13.0	11.2
Mean Marcu (1999) score	0.64	0.66	0.68	0.72

Table 1: Average salience scores of EDUs overlapping with SCU contributors, stratified by SCU weight. Differences between scores for each group are statistically significant under the Wilcoxon rank-sum test ($p < 0.05$).

salience score is significantly higher ($p < 0.05$) compared to the group with weight $w - 1$. That is, the more important a SCU is *across* these documents, the more important its corresponding EDUs are *within* the discourse of each document. We infer that the human authors of these summaries make structural decisions to highlight important concepts, and that these choices are reflected in the derived discourse structure.

With a large fraction of concepts observed to be contained within EDUs, we find compelling evidence to support the notion of EDUs as operational units of summarization. Moreover, we find evidence that the RST discourse structure which typically accompanies EDU segmentation also provides a strong signal of salience, though further experimentation along these lines is left to future work. We now investigate the utility of EDUs in a practical news summarization task using a large dataset.

4 Near-extractive summarization

In order to investigate the viability of discourse units in a practical setting, we use the New York Times Annotated Corpus (Sandhaus, 2008) which contains over 1.8 million articles published between 1987 and 2007 as well as their metadata. We mine this corpus to recover *near-extractive* summaries of articles which reveal how human editors selectively omit information from article sentences in order to preview the article for potential readers. This presents a middle ground between purely extractive and fully abstractive summarization which is useful to study the role of subsentential units in content selection.

4.1 Datasets

The NYT dataset contains editor-produced *online lead paragraphs*[2] which accompany 284,980 articles featured prominently on the NYT homepage from 2001 onwards. They are explicitly intended for presentation to readers and usually consist of one or more complete sentences which serve as a brief summary or teaser for the full article.[3]

We ensure that these online lead paragraphs—henceforth *online summaries*—are composed of complete sentences by filtering out cases which contain no verbs, omit sentence-terminating punctuation or are all-uppercase, respectively indicating summaries which are caption-like, truncated or merely topic/location descriptors. We also exclude articles with frequently repeated titles, first sentences and summaries which we observe to be template-like and thus not indicative of editorial input. Finally, we preprocess the remaining 244,267 summaries by stripping HTML artifacts and structured prefixes (e.g., bureau locations), normalizing Unicode symbols and fixing whitespace inserted within or deleted between tokens. We have released our data preparation code[4] to facilitate future research on the NYT corpus.

Three mutually exclusive datasets[5] are drawn from the processed document collection:

- EX-SENT: 38,921 fully extractive instances in which each summary sentence is drawn whole from the article when ignoring case, punctuation and whitespace.
- NX-SPAN: 15,646 near-extractive instances where one or more summary sentences form a contiguous span of tokens within an article sentence, and the remaining fit the definition above.
- NX-SUBSEQ: 25,381 near-extractive instances where one or more summary sentences form a non-contiguous token subsequence within an article sentence, and the remaining fit either of the definitions above.

The remaining 164,319 instances contain fully abstractive summaries with sentences that cannot be unambiguously mapped to those in the articles; these are not considered in the remainder of this

[2]Despite the name, these are typically not the same as the leading sentence or paragraph of the article.

[3]Note that this differs from the *abstracts* used in prior summarization research (Yang and Nenkova, 2014; Hong et al., 2015; Nye and Nenkova, 2015). We observe that abstracts appear to serve more as high-level structured descriptions of articles (e.g., referring to type of the article and NYT sections, using present-tense and collapsed sentences) rather than narrative summaries intended for presentation to readers.

[4]https://github.com/grimpil/nyt-summ

[5]The NYT document IDs for these datasets are available at http://www.cs.columbia.edu/~kapil/datasets/docids_nytsumm.tgz

NX-SPAN (contiguous)	**Summary:** Now that their season is over, the New York Yankees are likely to shop for new players over the winter. What they really should look for are new fans. **Doc EDUs:** [Now that their season is over,] [the New York Yankees are likely to shop for new players over the winter,] [*and may even*] [*seek a new manager*] [*to take over from the estimable Joseph Paul Torre.*] [What they really should look for are new fans.]
NX-SUBSEQ (non-contiguous)	**Summary:** The country's appetite for real estate propelled sales of newly built homes to a record pace in April, adding to concerns that the housing market may be in overdrive. **Doc EDUs:** [The country's avid appetite for real estate propelled sales of newly built homes to a record pace in April,] [*the Commerce Department reported yesterday,*] [*helping to raise prices*] [*and* adding to concerns] [that the housing market may be in overdrive.]

Table 2: Examples of reference summaries from NX-SPAN and NX-SUBSEQ alongside their source sentences from the article, segmented into EDUs. Tokens omitted by the summary are italicized.

paper but left to future work. Examples of summaries from the two near-extractive datasets are presented in Table 2 along with EDU-segmented source sentences from the corresponding articles.

4.2 Summary coverage

In order for our hypothesis that EDUs are good units for summarization to hold, we would expect the omitted text in these summaries to line up closely with the EDU segmentation of the source sentences. In particular, we expect to empirically observe that the number of of token edits required to recover reference summaries from source document EDUs is small.

For each type of unit—sentence and EDU—and every instance in NX-SPAN and NX-SUBSEQ, we align units derived from the original article with corresponding units from the online summary using Jaccard similarity, which is fairly reliable as the summaries are near-extractive. This procedure for deriving the set of input units matching output units is a necessary first step in training supervised summarization systems. Following this, we inspect the number of tokens that need to be deleted or added for each unit from the original article to match its counterpart in the summary. Distributions of the units in NX-SPAN and NX-SUBSEQ with respect to the number of tokens that need to be deleted or added are shown in Figure 6 and the average counts are presented in Table 3.

We observe that the number of deleted tokens as well as the proportion of units requiring token deletions is dramatically smaller when considering EDUs as summarization units. Token deletions are more frequent in summaries from NX-SUBSEQ in which deletions do not have to be continuous. Since EDUs in the summary may be erroneously aligned to different portions of the document, extraneous tokens may also be introduced; however, we observe these are relatively rare (3% for NX-SPAN and 10% for NX-SUBSEQ). No extraneous tokens are observed for sentence units as both datasets are near-extractive.

Dataset	Unit	# deleted	# added
NX-SPAN	Sent	11.47	0.00
	EDU	1.24	0.39
NX-SUBSEQ	Sent	11.95	0.00
	EDU	1.94	0.77

Table 3: Average #tokens deleted and added from each type of unit in NX-SPAN and NX-SUBSEQ.

We further analyze the types of tokens that are involved in the deletion process when using sentences and EDUs as base units. Figure 7 shows for each dataset the average numbers of deleted tokens grouped by their universal part-of-speech tags (Petrov et al., 2012). We observe that the number of deleted content words drops from 6.83–7.33 in the case of sentences to 0.54–0.92 for EDUs, making them easier to convert into reference summaries. For instance, spurious verbs frequently need to be removed from sentences in both datasets but this is relatively rare for EDUs.

5 Using EDUs for summarization

In this section, we compare EDUs with sentences as base units of selection in extractive and near-extractive single-document summarization. Crucially, we consider summarization under varying summary budget constraints in order to analyze whether EDU-based summarization is versatile enough to compete with typical sentence-based summarization when budgets are generous. Because our goal is to focus on the viability of summarization units for content selection, we evaluated system-generated summaries using ROUGE (Lin, 2004). Recovering readable sentences from EDU-based summaries remains a goal for future work.

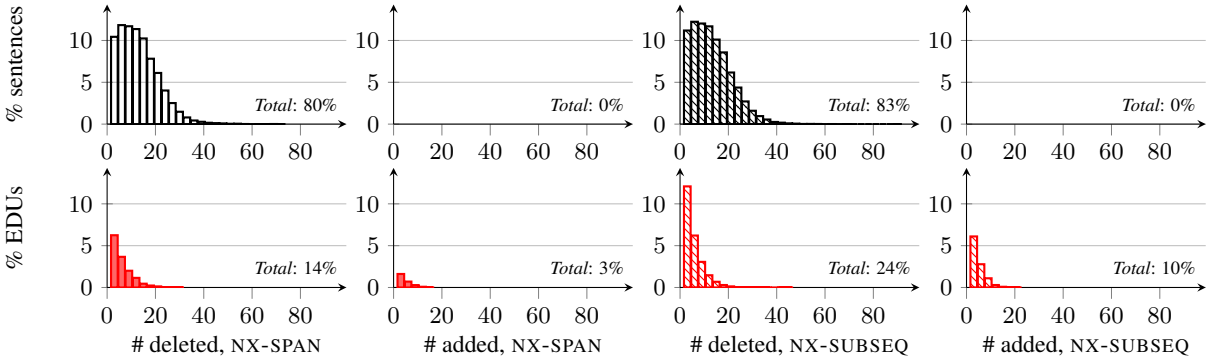

Figure 6: Proportion of source sentences and EDUs with the number of tokens deleted and added to recover summaries from NX-SPAN and NX-SUBSEQ. Cases with zero tokens added/deleted are omitted.

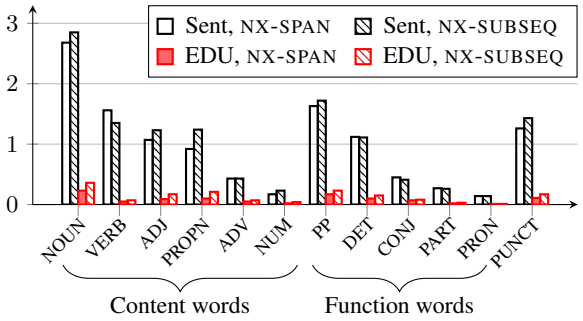

Figure 7: Average number of deleted tokens per instance in NX-SPAN and NX-SUBSEQ.

Summarization framework. We adopt a supervised structured prediction approach to extractive single-document summarization. Summaries are produced through greedy search-based inference with features defined over units in the document as well as over units and partial summaries, resulting in a feature-based generalization of Carbonell and Goldstein (1998).[6] In order to focus on the role of summarization units, we work with a simple standard model using features that are *neutral* to the benefits and/or drawbacks of either sentences or EDUs:[7]

- Position of the unit
- Position of the unit in the paragraph
- Position of the paragraph containing the unit
- TF-IDF-weighted cosine similarity of the summary with the unit added and the document centroid;
- Whether the unit is adjacent to the previous unit added
- Whether the sentence containing the unit is adjacent to the sentence containing the previous unit added

Feature weights are estimated using the structured perceptron (Collins, 2002) with parameter averaging for generalization. As inference is carried out via search, we employ a *max-violation* update policy (Huang and Feyong, 2012) to improve convergence speed and performance.

[6]We also experimented with beam search but did not observe improvements, as was also found in prior work (McDonald, 2007).

[7]For example, we do not use features related to nuclearity, discourse relation labels or discourse tree structure.

	EDU		Sentence	
Dataset	Lead	Greedy	Lead	Greedy
EX-SENT	0.65	**0.67**	0.55	**0.58**
NX-SPAN	0.46	**0.48**	0.32	**0.36**
NX-SUBSEQ	0.54	**0.56**	0.37	**0.40**

Table 4: ROUGE-1 of lead sentences vs. the supervised summarizer under a 200-char budget.

Data and settings. We use the extractive and near-extractive subsets from the NYT corpus described in Section 4.1 to train and evaluate our summarizer. To aid replicability for benchmarking, we partition all datasets by date rather than random sampling. Articles published in 2006–2007 are assigned to a held-out test partition while articles prior to 2005 are used for training, leaving articles from 2005 for a development partition.

The mean and standard deviation of summary lengths (specifically the number of characters) from our three NYT datasets are: EX-SENT 194.0 ± 92.6, NX-SPAN 134.6 ± 31.3, NX-SUBSEQ 143.3 ± 27.9. Summarization budgets are chosen to cover this range and set to 100, 150, 200, 250 and 300 characters. The lower bound (100 characters) is approximately one standard deviation below the mean across all three datasets, while the upper bound (300 characters) is approximately one standard deviation above the mean for EX-SENT, which features the longest summaries.

Comparison with lead. To validate this summarization framework, we first compare trained sum-

Budget	ROUGE-1 EDU	ROUGE-1 Sent	ROUGE-2 EDU	ROUGE-2 Sent	ROUGE-4 EDU	ROUGE-4 Sent
EX-SENT						
300	**0.80**	0.78	0.70	**0.71**	0.59	**0.71**
250	**0.75**	0.69	**0.64**	0.62	0.54	**0.61**
200	**0.67**	0.58	**0.56**	0.49	0.47	**0.48**
150	**0.54**	0.41	**0.43**	0.32	**0.35**	0.31
100	**0.35**	0.21	**0.26**	0.13	**0.20**	0.12
NX-SPAN						
300	**0.61**	0.58	**0.45**	0.44	0.37	**0.42**
250	**0.56**	0.50	**0.41**	0.36	0.33	**0.34**
200	**0.48**	0.36	**0.33**	0.20	**0.27**	0.18
150	**0.38**	0.22	**0.25**	0.08	**0.19**	0.06
100	**0.24**	0.14	**0.13**	0.04	**0.09**	0.03
NX-SUBSEQ						
300	**0.70**	0.69	0.53	**0.55**	0.38	**0.46**
250	**0.66**	0.59	**0.49**	0.44	0.35	**0.37**
200	**0.56**	0.40	**0.40**	0.24	**0.28**	0.20
150	**0.43**	0.22	**0.28**	0.08	**0.19**	0.05
100	**0.29**	0.14	**0.17**	0.04	**0.11**	0.02

Table 5: ROUGE results for EDU- and sentence-based summarization.

marizers against a standard summarization baseline which selects the leading sentence(s) of the document until the budget is exhausted. This evaluation uses a budget of 200 characters, which is about the average length of an extractive summary in our data.[8] ROUGE-1 results are shown in Table 4. Across all datasets and unit settings, the greedy summarizer consistently outperforms the lead baseline, indicating that the datasets involve non-trivial summarization problems.

Results. ROUGE results for all three datasets are shown in Table 5. For all budgets, scores are notably higher for EX-SENT which involves unambiguous alignment of reference units. ROUGE performance is also consistently higher for NX-SUBSEQ over NX-SPAN despite its higher token deletion rates (cf. Table 3), likely owing to a larger training dataset. All scores improve with bigger budgets as ROUGE is a recall-oriented measure.

We observe that EDUs outperform sentences across all datasets and budgets under ROUGE-1, on budgets within 250 characters under ROUGE-2 as well as budgets within 200 characters under ROUGE-4. Interestingly, EDU-based summarization remains competitive even on EX-SENT. The exceptionally strong performance of EDUs under tight budgets confirms our intuition that summarizers are better able to select salient information when working with smaller units. Sentences only hold a material advantage over EDUs when summarization budgets are generous enough to accommodate the more content-dense—and thus longer—source sentences. In our near-extractive datasets, this requires a budget greater than one standard deviation over the average size of reference summaries.

Analysis. Table 6 contains examples of reference summaries along with system-generated summaries produced using EDUs and sentences under a 200-character budget. All examples illustrate a common scenario in which an important source sentence is not selected by the sentence-based summarizer. Yet this is not because the model is unable to capture content salience, as the same features can recover salient EDUs. In each case, the source sentence behind the reference summary is barred from inclusion because of the summarization budget. By breaking these sentences into EDUs, the summarizer has the flexibility to select salient fragments of these sentences.

In addition, we observe a clear correspondence between EDU boundaries and the concepts which human editors selected for inclusion, regardless of whether they appear contiguously (Example B) or not (Example C). The variable length of EDUs is also helpful in keeping interdependent text whole. For instance in Example A, the third segment is 13 tokens long but belongs to a single EDU as it contains only one independent clause. This coherence is likely to be lost when working with smaller sub-sentential units such as n-grams.

6 Discussion and Future Work

In order to compare summarization units fairly, we used a simple model without utilizing the discourse structure of the document. However, the use of discourse trees has yielded promising results in summarization (Hirao et al., 2013; Yoshida et al., 2014). With larger training datasets such as the ones proposed here, an EDU-based summarizer will likely benefit from rich features over discourse relations. For instance, we observed in Section 3.3 that the Marcu (1999) measure can identify EDU importance, and furthermore a consideration of discourse relations across units is likely to encourage coherence in the resulting summary, potentially preventing the inclusion of unimportant and incongruous units.

Our results also highlight a need for future work

[8] We experimented with all other aforementioned budgets with consistent results.

(A) **Ref:** Manager Willie Randolph did not see what the big deal was. All he did before last night's game against San Diego at Shea Stadium was drop Mike Piazza in the batting order to sixth from fifth and promote David Wright to fifth from sixth. But the swap led to a barrage of questions from reporters.
EDU: [Manager Willie Randolph did not see what the big deal was.] [All] [he did before last night 's game against San Diego at Shea Stadium] [was drop Mike Piazza in the batting order to sixth from fifth]
Sent: Manager Willie Randolph did not see what the big deal was. But the swap led to a barrage of questions from reporters. Was Piazza being demoted permanently? How had Piazza and Wright handled the moves?

(B) **Ref:** Big, cheap and somewhere in Manhattan. Those were the starting criteria for Kelli Grant, who was desperate to escape a long bus commute between Midtown and southern New Jersey.
EDU: [Big, cheap] [and somewhere in Manhattan.] [Those were the starting criteria for Kelli Grant,] [and for her boyfriend, James Darling,] [to be with her.]
Sent: Big, cheap and somewhere in Manhattan. At that early, uninformed stage, big meant two bedrooms, they hoped. Cheap meant up to $1,500 a month.

(C) **Ref:** The plan, which rivals the scope of Battery Park City, would rezone a 175-block area of Greenpoint and Williamsburg.
EDU: [The plan,] [which rivals the ambition and scope of the creation of Battery Park City,] [would rezone a 175-block area of Greenpoint and Williamsburg, two neighborhoods]...[and led to intense pressure]
Sent: The plan, which is expected to be approved by the full City Council next week, imposes some novel requirements for developers seeking to build the housing.

Table 6: Examples of NYT reference and system-generated summaries using EDUs and sentences from (A) EX-SENT, (B) NX-SPAN, (C) NX-SUBSEQ. An "..." separates EDUs from different source sentences.

in composing EDUs to form fluent sentences. As suggested by the coverage analysis in Section 3.2, it is very likely that this can be accomplished robustly. For instance, Daumé and Marcu (2002) demonstrated that an EDU-based document compression system can improve over sentence extraction in both grammaticality and coherence.

7 Conclusion

In this work, we explore the potential of elementary discourse units (EDUs) from Rhetorical Structure Theory in extending extractive summarization techniques to produce a wider range of human-like summaries. We first demonstrate that EDU segmentation is effective in preserving concepts extracted from a document. We also analyze summaries in the New York Times corpus whose content is extracted from parts of their original sentences. When recovering the summaries using EDUs, the amount of extraneous information in the form of content words is dramatically reduced compared to their original sentences. Finally, we demonstrate that using EDUs as units of content selection instead of sentences leads to stronger summarization performance on these near-extractive datasets under standard evaluation measures, particularly when summarization budgets are tight.

References

Miguel Almeida and André F. T. Martins. 2013. Fast and robust compressive summarization with dual decomposition and multi-task learning. In *Proceedings of the ACL*.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of ACL-HLT*.

Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca Passonneau. 2015. Abstractive multi-document summarization via phrase selection and merging. In *Proceedings of the ACL*.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of AAAI*.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*, pages 85–112. Springer Netherlands.

James Clarke and Mirella Lapata. 2007. Modelling compression with discourse constraints. In *Proceedings of EMNLP-CoNLL*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*.

Hal Daumé, III and Daniel Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the ACL*.

Vanessa Wei Feng and Graeme Hirst. 2014a. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the ACL*.

Vanessa Wei Feng and Graeme Hirst. 2014b. Two-pass discourse segmentation with pairing and global features. *CoRR*, abs/1407.8215.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. A formal model for information selection in multi-sentence text extraction. In *Proceedings of COLING*.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*.

Hugo Hernault, Helmut Prendinger, David A duVerle, Mitsuru Ishizuka, et al. 2010. HILDA: a discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of EMNLP*.

Kai Hong, Mitchell Marcus, and Ani Nenkova. 2015. System combination for multi-document summarization. In *Proceedings of EMNLP*.

Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of LREC*.

Liang Huang and Suphan Feyong. 2012. Structured perceptron with inexact search. In *Proceedings of NAACL-HLT*.

Hongyan Jing and Kathleen R. McKeown. 2000. Cut and paste based text summarization. In *Proceedings of NAACL*.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2014. Single document summarization based on nested tree structure. In *Proceedings of the ACL*.

Wei Li. 2015. Abstractive multi-document summarization with semantic information extraction. In *Proceedings of EMNLP*.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of ACL*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL 2004 Workshop*.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of NAACL*.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of SIGDIAL*.

William C Mann 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):243–281.

Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, pages 123–136.

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

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3):103–233.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The Pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2), May.

Benjamin Nye and Ani Nenkova. 2015. Identification and characterization of newsworthy verbs in world news. In *Proceedings of NAACL*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Language Resources and Evaluation Conference*.

Evan Sandhaus. 2008. The New York Times Annotated Corpus LDC2008T19. *Linguistic Data Consortium, Philadelphia*.

Kapil Thadani and Kathleen McKeown. 2008. A framework for identifying textual redundancy. In *Proceedings of COLING*.

Kristian Woodsend and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of EMNLP*.

Yinfei Yang and Ani Nenkova. 2014. Detecting information-dense texts in multiple news domains. In *Proceedings of AAAI*.

Qian Yang, Rebecca J Passonneau, and Gerard de Melo. 2016. PEAK: Pyramid evaluation via automated knowledge extraction. In *Proceedings of AAAI*.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. In *Proceedings of EMNLP*.