Sentence Simplification, Compression, and Disaggregation for Summarization of Sophisticated Documents

Catherine Finegan-Dollak
Department of EECS, University of Michigan, 3861 Beyster Building, Ann Arbor, MI 48109. E-mail: cfdollak@umich.edu

Dragomir R. Radev
Department of EECS, University of Michigan, 3917 Beyster Building, Ann Arbor, MI 48109. E-mail: radev@umich.edu

Sophisticated documents like legal cases and biomedical articles can contain unusually long sentences. Extractive summarizers can select such sentences—potentially adding hundreds of unnecessary words to the summary—or exclude them and lose important content. Sentence simplification or compression seems on the surface to be a promising solution. However, compression removes words before the selection algorithm can use them, and simplification generates sentences that may be ambiguous in an extractive summary. We therefore compare the performance of an extractive summarizer selecting from the sentences of the original document with that of the summarizer selecting from sentences shortened in three ways: simplification, compression, and disaggregation, which splits one sentence into several according to rules designed to keep all meaning. We find that on legal cases and biomedical articles, these shortening methods generate ungrammatical output. Human evaluators performed an extrinsic evaluation consisting of comprehension questions about the summaries. Evaluators given compressed, simplified, or disaggregated versions of the summaries answered fewer questions correctly than did those given summaries with unaltered sentences. Error analysis suggests 2 causes: Altered sentences sometimes interact with the sentence selection algorithm, and alterations to sentences sometimes obscure information in the summary. We discuss future work to alleviate these problems.

Received December 20, 2014; revised April 29, 2015; accepted April 30, 2015

 • Published online 23 September 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23576

Introduction

Most automatic summarizers are extractive: They select complete sentences from the original document(s) (Mani & Maybury, 1999; Radev, Hovy, & McKeown, 2002). However, sentences in an original document are often not ideally suited to a summary. One problem is that a goal of summaries is to convey information concisely, but sentences from the original document were not written with this limitation in mind. Thus, an extracted sentence might contain both essential and extraneous information. Several researchers have sought to address this problem through sentence simplification and sentence compression. (Knight & Marcu, 2002; Siddharthan, Nenkova, & McKeown, 2004; Zajic, Dorr, Lin, & Schwartz, 2007).

These efforts have largely focused on deleting portions of sentences deemed unimportant. Even systems that implement forms of simplification and compression other than deletion also use deletion as one of their operations. Although these systems have as their goal the reduction of sentence length with minimal loss of meaning, deletion necessarily entails some loss of meaning.

Naturally, a summary cannot include all of the meaning of the original document. Summarization algorithms use information from the entire document to select important portions of the meaning for inclusion in the summary. It makes little sense, then, to make deletion a separate step from sentence selection, as this takes the decision about what meaning is important away from the summarization algorithm. We might thus expect improved summarization if we divide the sentences into smaller units of meaning that the selection algorithm can act upon.

Our approach to such division is to disaggregate sentences—that is, split one long sentence into two or more

shorter sentences—before running the sentence selection algorithm. As an example, the sentence

> The district court's decision cannot be affirmed on the ground that the petition was untimely, and we must take up the merits. (1)

might be disaggregated to

> The district court's decision cannot be affirmed on the ground that the petition was untimely. We must take up the merits. (2)

Disaggregating sentences thus allows extractive summarizers to act on shorter units. Rather than deciding whether all 23 words of Example (1) should be included in the summary, the system can make separate decisions about the 15-word first sentence and the six-word second sentence of Example (2), perhaps allowing it to include important information in the summary while leaving out less important information. Unlike sentence compression, however, disaggregation only *reorganizes* meaning; it does not *remove* content.

Another problem with prior work on sentence compression for summarization is its focus on newswire articles. While research has begun on summarizing documents such as scientific journal articles, legal cases, and dissertations (Hachey & Grover, 2006; Jha, Coke, & Radev, 2015; Ou, Khoo, & Goh, 2007), research on sentence compression for summarization has largely ignored such documents. These sophisticated documents—written for audiences with years of specialized education, like lawyers, scientists, and doctors—are particularly likely to include long, complex sentences, and thus particularly likely to benefit from a procedure that allows the summarizer to choose from shorter units of meaning.

In addition, preserving as much meaning as possible is especially important in these sophisticated domains. For example, while it might be safe to compress

> Sources say that the House is likely to vote on the issue tomorrow. (3)

in a newswire article into

> The House is likely to vote on the issue tomorrow. (4)

we cannot compress

> Plaintiffs say that we must find Defendant in contempt. (5)

in a legal case into

> We must find Defendant in contempt. (6)

without significantly altering the meaning. Disaggregation thus seems likely to be particularly useful in sophisticated domains.

We evaluate several alternative methods of summarizing two types of sophisticated documents: legal cases and biomedical research articles. We compare the performance of an extractive summarizer when it selects from unaltered sentences with its performance on simplified, compressed, or disaggregated sentences.

The rest of this paper is organized as follows. We first describe related work in summarization and simplifying and compressing sentences. Second, we describe the task and the documents we seek to summarize. Third, we describe the existing sentence compression and sentence simplification systems we test with our summarizer, as well as a sentence disaggregation technique that uses 112 manually written rules to split sentences. Fourth, we describe the extractive summarization algorithm used in all conditions of our experiment. Next, we describe our experiment comparing the four systems. Finally, we discuss possible reasons for the observed results, address future avenues of research, and present our conclusions.

Related Work

Summarization[1]

From the earliest days of automatic text summarization through today, extraction of sentences from the original document has been the preferred approach (see, e.g., Barzilay & Elhadad, 1997; Erkan & Radev, 2004; Haghighi & Vanderwende, 2009; Hovy & Lin, 1998; Luhn, 1958; Mihalcea & Tarau, 2004; Piwowarski, Amini, & Lalmas, 2012; Yang & Wang, 2008).

Evaluation of summaries may be intrinsic or extrinsic. Rath, Resnick, and Savage (1961) began the popular approach of intrinsic evaluation—that is, of evaluating summaries by determining how similar they were to human-written summaries. Many modern automatic summary evaluations use the ROUGE system (Lin, 2004), which measures the similarity of the automatic summary to several human-written summaries. Similarly, for the pyramid method of evaluation, human judges compare the target summary to several human-generated models. (Nenkova, Passonneau, & McKeown, 2007).

An alternative approach is extrinsic (task-based) evaluation. Rather than asking, "How similar is this summary to a human-written one?" we ask, "How well does this summary enable its readers to complete a task?" Morris, Kasper, and Adams (1992) evaluated summaries by having subjects who read a condensed form complete a reading comprehension test to assess how well they understood the content of the original document. Hand (1997) described how well humans who read the summaries could categorize the document and decide whether it was relevant to a query. McKeown, Passonneau, Elson, Nenkova, and Hirschberg (2005) tested the ability of human readers to complete a time-limited, fact-gathering task using summaries. Otterbacher, Radev,

[1]A complete review of prior work in automatic summarization is beyond the scope of this paper; the interested reader is encouraged to consult Mani and Maybury (1999), Mani (2001), and Radev et al. (2002) as a starting point.

DOI: 10.1002/asi

and Kareem (2008) used comprehension questions to evaluate the usefulness of hierarchical summaries to readers using mobile devices. Murray et al. (2009) evaluated summaries of meetings by asking human readers about how and why a decision was made at the meetings. And Rastkar, Murphy, and Murray (2014) used task-based evaluation on summaries of software bug reports.

Sentence Simplification, Compression, and Disaggregation

Background and early work. Work in shortening sentences falls into two main categories: sentence simplification and sentence compression. Early simplification research was intended to benefit readability, parsing, and summarization and to improve accessibility for people with disabilities. It generally involved simplifying both structure and word choice. Carroll, Minnen, Canning, Devlin, and Tait (1998) described a pipeline incorporating analysis, lexical simplification, and syntactic simplification. Their syntactic simplifier was based on handwritten rules, such as replacing passive constructions with active constructions. Chandrasekar and Srinivas (1997) learned simplification rules automatically. In the earliest work on pure sentence compression that we are aware of, Grefenstette (1998) described "telegraphic text reduction" to allow blind people using a reading machine to skim text; a user could choose, for example, to see only proper nouns; or only subjects, head verbs, and objects of the main clause; or only subjects and object nouns including subclauses.

In the early 2000s, simplification work split in two directions. While some researchers focused on simplification for readability, others began to work on sentence compression for summarization. Confusingly, some work on sentence compression was still called simplification. For consistency, we refer to efforts to shorten sentences by discarding some content as compression, efforts to make text easier to read as simplification, and efforts to split sentences as disaggregation.

Sentence compression for summarization. Knight and Marcu (2002) presented two approaches to sentence compression for summarization: a noisy-channel model and a decision-tree approach. In both cases, their goal was to generate a grammatically correct compression that included the most important pieces of information from the original sentence but deleted some subset of the words. Turner and Charniak (2005) created unsupervised and semisupervised models to complement Knight and Marcu's supervised learning approach. Examining Knight and Marcu's hypothesis that sentence compression could improve summarization, Lin (2003) found that an oracle method of compression—reranking candidate compressions using the manual summaries of the same documents—did improve performance on a forerunner of the ROUGE evaluation (Lin, 2004); however, Knight and Marcu's noisy-channel compression model actually worsened performance. Lin noted that even the oracle condition did not improve summaries as much as expected and suggested that sentence compression "might drop some important content."

Siddharthan et al. (2004) compressed sentences for summaries by removing parentheticals. They found that this improved clustering, in that it got rid of background information. They extracted the desired sentences for the summary, then added the parenthetical information back where it was needed—the first time the entity it described was mentioned in the summary.

Some groups permitted the sentence extraction module to choose from more than one possible variation on the same sentence. Zajic et al. (2007) altered a parse-and-trim compression approach so that it produced multiple compressions of a sentence, then used a sentence selector to choose from the pool of candidate sentences based on a linear combination of features. Similarly, Vanderwende, Suzuki, Brockett, and Nenkova (2007) wrote manual rules to remove certain syntactic units such as appositives, then provided their summarizer with both the compressed sentence and the original sentence. They relied upon the sentence selection algorithm's ability to deal with redundancy to ensure that it did not select both versions of a single sentence. On Document Understanding Conference (DUC) 2006 data, over 40% of the sentences the selection component chose to include were the simplified sentences, which resulted in the ability to add one extra sentence to each summary on average.

Using integer linear programming (ILP) for sentence compression (Clarke & Lapata, 2007, 2008) led to improved results and a burst of related research. Clarke and Lapata's model maximized a scoring function while adhering to sentence-level constraints to ensure grammatical output. They later added the further step of considering the context surrounding the sentence to be compressed when choosing a compression. (Clarke & Lapata, 2010). Martins and Smith (2009) improved summarization performance by optimizing an objective function with a single set of constraints that incorporated both compression and extraction. Similarly, Berg-Kirkpatrick, Gillick, and Klein (2011) found that jointly learning a model for extraction and compression outperforms the model that only learns extraction for multidocument summaries. One problem with their model was that solving the ILP for joint extraction and compression was an order of magnitude slower than solving the ILP for extraction only; sometimes the process was prohibitively slow. Chali and Hasan (2012) compared three methods of query-focused multidocument summarization using ILP-based compression and extraction: choosing sentence compressions and then doing sentence extraction; doing extraction and then compressing the sentences; and combining compression and extraction. Combined extraction and compression performed the best, followed by first extracting and then compressing sentences. Li, Liu, Weng, and Liu (2013), noting that joint compression and selection by ILP can be prohibitively expensive, proposed an alternative: use summary-guided compression to get a set of good possible compressions, apply a preselection step, then use the ILP selection framework to select the compressed sentences.

DOI: 10.1002/asi

The most recent work with ILP for sentence compression has been in Joint Structural Inference (Thadani & McKeown, 2013). Whereas earlier ILP compression systems used either a language model or edges in a dependency graph to represent text, this approach uses both, providing performance gains without requiring hand-picked constraints. A follow-up to that paper (Thadani, 2014) addressed the slowness of joint inference by solving the two subproblems separately and generating approximate solutions to the graph-based subproblem.

Simplification for readability. Work on simplification for readability has continued in parallel. Siddharthan (2002, 2006) proposed a manual-rule-based model with analysis, transformation, and regeneration stages. He described handwritten rules to recognize and simplify adjectival or relative clauses, adverbial clauses, coordinate clauses, subordinate clauses, correlated clauses, participial phrases, appositive phrases, and passive voice based on chunking, part-of-speech (POS) identification, and constituency parsing. Siddharthan (2010) incorporated handwritten rules that matched patterns in a typed dependency parse and that could make certain lexico-syntactic substitutions so that the resulting sentences included the correct parts POS tags.

Wubben, Van Den Bosch, and Krahmer (2012) moved away from the handwritten rules model of simplification and instead used a monolingual phrase-based machine translation (PBMT) model trained using Wikipedia and Simple Wikipedia as parallel corpora. Noting that simplification work was becoming divided between manual rules and statistical methods, Siddharthan and Angrosh (2014) introduced a method that used both manual rules and rules learned from parallel English Wikipedia and Simple English Wikipedia texts.

Disaggregation. The only research that we are aware of that disaggregates sentences is Siddharthan (2002, 2006) and Klebanov, Knight, and Marcu (2004).[2] While Siddharthan (2002) described a wide range of syntactic constructs that could be simplified according to rules, the rules actually implemented only split certain constructs into multiple sentences. Jonnalagadda and Gonzalez (2010) successfully applied Siddharthan's sentence splitting system to improve recall without harming precision in information extraction from biomedical papers. Klebanov et al. (2004) introduced the concept of an "Easy Access Sentence" (EAS) for information extraction. An EAS should have one finite verb, make only claims that were present in the original sentence, and include as many entity names as possible. To generate EASs, they used a named entity recognizer and a parser, then constructed a sentence for every verb in the original sentence.

Problem Definition and Algorithms

Task Definition

Our goal is to generate improved summaries of sophisticated documents by selecting from shorter sentences than the original document contained. By sophisticated documents, we mean documents written for a highly educated audience, which are likely to include long or complicated sentences. For this study, we consider legal cases and biomedical articles.

In the United States legal system, when a court decides an issue of law a judge writes a document called an "opinion," "decision," or "case," explaining the circumstances of the case, the issues before the court, and what the court decided and why. Such legal opinions are precedent that guide—and in some situations bind—courts that consider similar issues in the future. Thus, legal professionals must know about opinions. Yet these professionals face information overload, as tens of thousands of opinions are issued in the U.S. federal courts alone each year. Summarization of cases is therefore essential. Unfortunately, legal cases are quite challenging to summarize, in part because of the writing style they often use, where long, complicated sentences are the norm (Tiersma, 1999).

We assembled a corpus of 30 random legal cases from all 13 U.S. Federal Circuit Courts of Appeal (2009–2013). The cases include criminal, civil, and bankruptcy matters; administrative agency appeals; and proceedings originating in the appellate courts. No unpublished cases were included. We also collected manually written summaries of each of these cases from LexisNexis.[3]

Like legal cases, biomedical articles are a domain where there is "a pressing need for distillation and management of complex information presented in vast amounts of text," according to the TAC 2014 Biomedical Summarization Track.[4] We assembled 35 randomly chosen articles from PubMed Central (PMC) (2009–2013).[5] Topics range widely; our corpus includes behavioral genetics, a thyroid cancer case study, evaluation of the safety of a treatment for stroke, and mapping of a gene.

At first glance, sentence lengths in both the biomedical and the legal corpus appear modest. The development (dev) set for legal cases had a mean sentence length of 24.01 words, while the biomedical articles dev set had a mean of 23.53. For comparison, the corpus of newswire articles for Task 2 of DUC 2004 had a mean sentence length of 25.11. However, the legal and biomedical corpora had much higher variance than the DUC example. Table 1 shows the percentage of sentences from each corpus with more than 50 words or more than 75 words. Such long sentences are rare in newswire articles, but common in the sophisticated documents. As the histograms in Figure 1a–c illustrate, the legal cases and biomedical articles include an unusually high

[2]Although termed "simplification" by their authors, these works actually broke long sentences into pieces while preserving the meaning and thus are disaggregation as defined here.

[3]http://www.lexisnexis.com.

[4]http://www.nist.gov/tac/2014/BiomedSumm/.

[5]http://www.ncbi.nlm.nih.gov/pmc/.

DOI: 10.1002/asi

TABLE 1. Percentage of words from the legal and biomedical dev sets and the DUC comparison corpus containing more than 50 or more than 75 words.

	Sentences over length	
Document type	50 words	75 words
Legal	9.4%	2.0%
Biomedical	8.2%	1.4%
DUC	1.8%	0.1%

number of extremely short sentences[6] and a long tail of sentences over 100 words long, with maxima over 250 words. Clearly, including a single 250-word sentence in a summary leaves little room for other sentences.

The desired output for this task is a single-document summary of limited length for each document. For biomedical articles, we limited summary length to 20% of the original document, following Reeve, Han, and Brooks (2007) and Ou et al. (2007). For legal cases, we limited summary length to the length of the human-generated summary for that case, which for this study were on average 26% of the length of the input cases.

Existing Systems

To solve the problem of summarizing documents that include long, complicated sentences, the current work tries to make the sentences shorter before running a summarization algorithm. We compare three such methods: an existing compression method, an existing simplification method, and our disaggregation method.

Compression system. For the compression system, we used Napoles's implementation[7] of the (Clarke & Lapata, 2008) ILP algorithm for sentence compression. A complete explanation of ILP is beyond the scope of this paper. Briefly, though, ILP seeks to identify the values for decision variables that maximize a linear objective function.

In this case, the objective function combines a language model with a significance scoring function. The language model, trained on the English Gigaword corpus (Graff & Cieri, 2003), allows it to estimate the probability that each unigram, bigram, and trigram in a proposed compression would occur in English. Thus, for example, given the sentence

The dogs barked at the hissing cats. (7)

the language model portion of the objective function would tend to prefer the compression

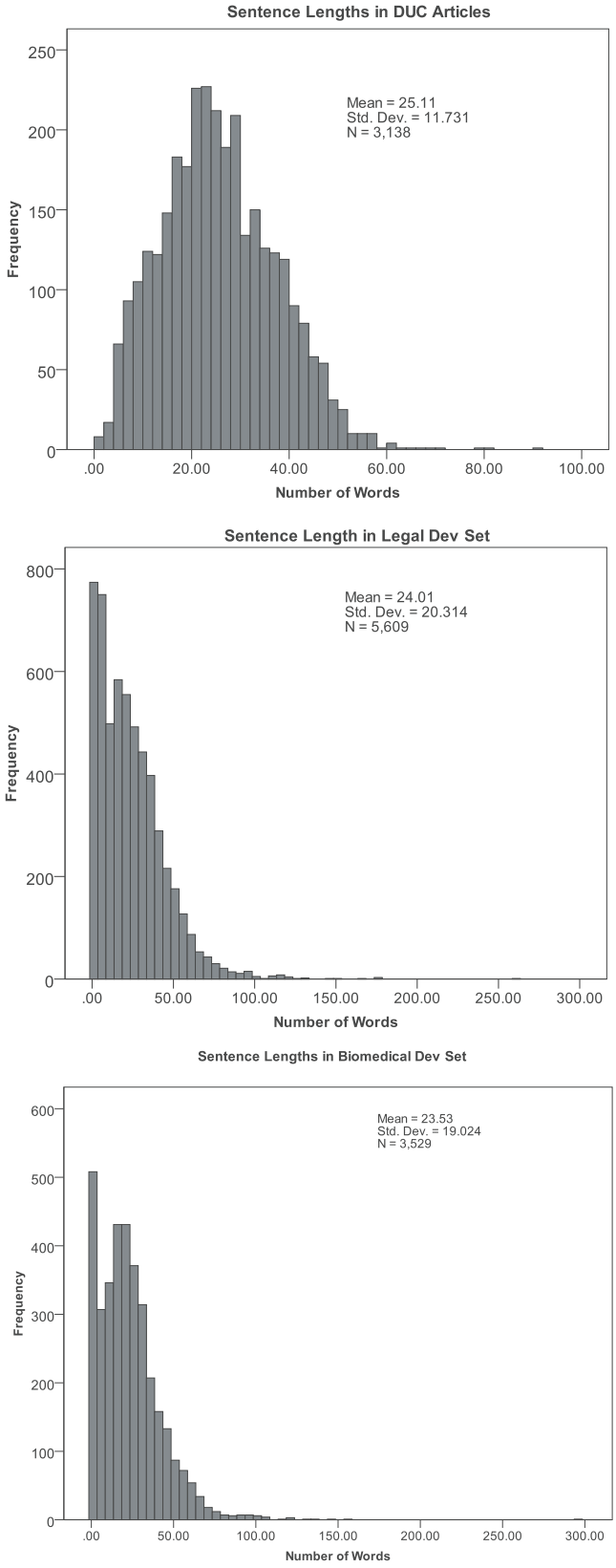

FIG. 1. Distribution of sentence lengths in DUC corpus is nearly normal. Distribution of sentence lengths in dev sets of legal cases and biomedical articles is noticeably skewed, with a large number of very short sentences and a long tail of very long sentences. Maximum sentence lengths for both were over 250 words.

[6]The short sentences include headings, figure and table names, references to journals or legal authorities, and the like.

[7]https://github.com/cnap/sentence-compression.

DOI: 10.1002/asi

Dogs barked at cats. (8)

over

The barked the hissing. (9)

The significance scoring portion of the function assigns words a weight representing their significance, to ensure that the compression contains the most important words of the original sentence.

The system also includes constraints. For example, a modifier will not be included in the compression unless its headword is included. This prevents a compression that includes "hissing" but excludes "cats."

ILP-based compression for summarization generates improved output when the ILP performs both compression and sentence selection jointly (Chali & Hasan, 2012), but here we have taken a pipeline approach of compressing and then selecting sentences. The pipeline is necessary to allow us to compare apples to apples: If we used a joint compression and selection system, we would not be able to determine whether differences in performance between that and the other systems was due to the method of shortening sentences or due to the difference between sentence selection algorithms.

Simplification system. For sentence simplification, we use the hybrid rule-based and learned simplification model of Siddharthan and Angrosh (2014). The system combines handwritten rules for syntactic simplification and rules learned by a translation system trained on aligned English Wikipedia and Simple English Wikipedia texts. The system recognizes patterns in dependency parses of sentences, applies rewrite rules to the parse tree according to which pattern(s) it matched, and generates the simplified sentence text from the altered parse tree. For example, the paper describes a handwritten syntactic simplification rule to replace passive with active voice, shown in Figure 2.

RULE: PASSIVE2ACTIVE
1. DELETE
(a) nsubjpass(?X0, ?X1)
(b) auxpass(?X0, ?X2)
(c) agent(?X0, ?X3)
2. INSERT
(a) nsubj(?X0, ?X3)
(b) dobj(?X0, ?X1)
3. NODE OPERATIONS
(a) AGR-TENSE:?X0←?X2
(b) AGR-NUMBER:?X0←?X3

FIG. 2. A rule from the simplification system for replacing passive voice with active voice.

Given a sentence

The cats were chased by a dog. (10)

(modified from an example in the paper) with a dependency parse as shown in Figure 3a, this rule would generate the new parse tree shown in Figure 3b, which in turn would generate the simplified sentence

A dog chased the cats. (11)

The automatically generated rules work similarly but focus on lexical and lexico-syntactic simplification. Thus, for example, there is a rule that, if X_1 is a word from the set "[extensive, large, massive, sizable, major, powerful, unprecedented, developed, giant]," will delete amod(X_0, X_1) and replace it with amod(X_0, X_2), where X_2 is "big." The rule will also transfer any subtree rooted at X_1 to instead have X_2 as its root.

The simplifier permits five operations:

1. Delete: remove the specified dependency relations from the old sentence.
2. Insert: add the specified dependency relations to the new sentence(s).
3. Order: process the daughters of a node of the dependency graph in the specified order.
4. Move: delete the specified node (as opposed to just the relation) from the graph, and attach its children to a specified other node or, if unspecified, to the parent of the deleted node.
5. Node Operations: make morphological changes to a specified word to ensure its agreement with other words in the sentence.

The explicit goal of this system is to make sentences easier to read, not simplification for summarization. It is nevertheless well suited for comparison with the other systems, not only because it is a state-of-the-art simplification system, but also because it has performed particularly well on tests of meaning preservation, perhaps because it does not delete information through sentence compression. It therefore can be expected to be particularly competitive with the disaggregation system in generating sentences that are shorter than the source sentence but still retain all of the meaning.

det(cats-2, The-1)
nsubjpass(chased-4, cats-2)
auxpass(chased-4, were-3)
root(ROOT-0, chased-4)
det(dog-7, a-6)
agent(chased-4, dog-7)
(a) Original Parse

det(cats-2, The-1)
root(ROOT-0, chased-4)
det(dog-7, a-6)
nsubj(chased-4, dog-7)
dobj(chased-4, cats-2)
(b) Simplified Parse

FIG. 3. The dependency parse of model sentence 10 before and after applying the simplification rule in Figure 2.

DOI: 10.1002/asi

Disaggregation System

We built the sentence disaggregator from a modified version of the (Siddharthan & Angrosh, 2014) sentence simplifier. Our changes fell into two categories: changes to rules and changes to the system that implements the rules.

Modification of the simplification system for disaggregation. In addition to the five operations of the simplification system described earlier, we found a sixth operation, copy, necessary for disaggregation for summarization, due to a difference in the best way to handle duplicate subtrees. Sometimes when a sentence is split, the same subject must be used in both sentences. For example, in the sentence

> The blue and yellow balloons floated above the ground and drifted in the breeze. (12)

"balloons" is the subject of both "floated" and "drifted." If the resulting split sentences are read together, it is redundant to say

> The blue and yellow balloons floated above the ground.
> The blue and yellow balloons drifted in the breeze. (13)

and would instead be more desirable to simplify to

> The blue and yellow balloons floated above the ground.
> These balloons drifted in the breeze. (14)

The simplifier therefore replaces the subtree that depends from this repeated subject with the determiner "this" or "these" as appropriate.

In a summary, though, where one of the resulting sentences might be extracted without the other, this substitution could cause confusion or inaccuracy. For example, the summary could misleadingly say

> The red balloons had not yet been inflated. These balloons drifted in the breeze. (15)

A better disaggregation for extractive summarization requires including the subtree that descends from "balloons" instead of replacing it with a determiner. In other circumstances, though, the flexibility to exclude some or all of the subtree may be desirable. For instance, it is preferable to split

> The court held that community antenna television ("CATV") systems, which retransmitted signals, did not infringe. (16)

into

> The court held that community antenna television ("CATV") systems did not infringe.
> CATV systems retransmitted signals. (17)

Disaggregation for summarization therefore benefits from the option to include the subtree or not. Hence, we modified the simplifier so that rules could specify that a word should be copied without its subtree.

Disaggregation rules. With the modified system in place, we identified disaggregation rules. First, we removed all rules from the original system that involved lexical substitution, as well as rules for converting passive to active voice and rules standardizing quotations. These are pure simplification rules and are not a part of disaggregation. We retained most rules for splitting appositions, relative clauses, subordination, and conjunction into separate sentences. Some of these rules were not well suited for summarization, however. For example, the simplifier included a rule that splits a sentence such as

> The court reduced the term of imprisonment after considering the factors set forth in section 3553(a). (18)

into

> The court reduced the term of imprisonment.
> This happened after considering the factors set forth in section 3553 (a). (19)

If the second sentence is included in a summary without the first, the reader will be left wondering *what* happened after these factors were considered. We therefore disabled 21 such rules, leaving 94 of the original, handwritten syntactic simplification rules in place.

Preliminary testing showed that sentences in legal cases and biomedical articles often could not be split by these rules. We therefore wrote additional disaggregation rules designed for these complex sentence structures. But choosing what rules to write is a subjective task. Therefore, to choose the best disaggregation, one of the authors prepared a preliminary set of disaggregation instructions, and then the authors and a colleague from the same lab individually disaggregated a set of 10 sample sentences from the development set using those instructions. The resulting disaggregations were compared. The researchers discussed the best disaggregation of each sample sentence for the summarization task. When consensus was reached, the result became the gold standard. Manual rules were then written based on the gold standard models and variations on them.

For example, one original sentence in the gold standard was

> Just before the passage of the 1976 Copyright Act, the Supreme Court held in Fortnightly Corp. v. United Artists Television, Inc. and Teleprompter Corp. v. Columbia Broadcasting System, Inc. that community antenna television ("CATV") systems—which captured live television broadcasts with antennas set on hills and retransmitted the signals to viewers unable to receive the original signals—did not infringe the public performance right because they were not "performing" the copyrighted work. (20)

DOI: 10.1002/asi

TABLE 2. Examples of the difference in output of the disaggregation system and the simplifier.

Original sentence	Disaggregated sentence	Simplified sentence
Actinic keratosis, also known as senile keratosis, results from the proliferation of atypical keratinocytes as a consequence of long exposition to ultraviolet radiation and it has been considered a premalignant lesion which may evove[sic] to squamous cell carcinoma.	Actinic keratosis are also known as senile keratosis. Actinic keratosis results from the proliferation of atypical keratinocytes as a consequence of long exposition to ultraviolet radiation. It has been considered a premalignant lesion. This premalignant lesion may evove[sic] to squamous cell carcinoma.	Actinic keratosis, sometimes called as senile keratosis, results from the proliferation of atypical keratinocytes because of long exposition to ultraviolet radiation and it has been thought a premalignant lesion. This lesion may evove[sic] to squamous cell carcinoma.

The gold standard called for the relative clause set off by em-dashes (—) to be split off and become two new sentences:

> CATV systems captured live television broadcasts with antennas set on hills.
> CATV systems retransmitted the signals to viewers unable to receive the original signals. (21)

Although the underlying simplification system included rules to separate relative clauses into their own sentences, such rules required the dependency parser to use the "rcmod" dependency. The parser did not apply this relationship to relative clauses set off by em-dashes rather than commas. We therefore needed a rule to recognize the actual dependency parse of a relative clause without relying on the rcmod relationship to signal it. For this gold standard sentence, the rule would need to make "systems" the subject of the new sentences and keep the "CATV" abbreviation modifying it. Such a rule was therefore specific to sentences with a token in an abbreviation relationship with the noun that would become the new subject. But of course, we wanted the system to recognize relative clauses without an abbreviation, and this required a related rule.

In total, our disaggregation system applied 112 rules—94 from the simplifier, plus 18 that we developed using this approach. The difference in output between our disaggregator and the original simplifier is illustrated in Table 2.

Summarization System

For sentence selection, we use the graph-based LexRank summarizer (Erkan & Radev, 2004). In the LexRank Only condition of our experiment, we run LexRank on the sentences of the original document to sort its sentences in descending order of importance. In the Simplified, Disaggregated, and Compressed conditions, we instead use the simplified, disaggregated, or compressed sentences as input to LexRank.

LexRank represents sentences as nodes of a graph. The graph of the document includes an edge between a pair of sentences if their cosine similarity exceeds a certain threshold. Edges coming into a sentence indicate that it is important to the document, so sentences "vote" for related sentences. However, a measure based purely on the degree of a node would allow a group of outlier sentences that were closely related to each other, but not to the main thrust of the document, to create an illusion of centrality. To avoid this problem, LexRank weights a node's votes using the node's centrality.

Starting with the most important sentence, we select sentences to add to the summary until we encounter one that would cause our summary to exceed the maximum word count. But because we are dealing with sophisticated documents containing long sentences, we may have found a particularly long sentence that could not be split or compressed, and there may in fact still be room in the summary for one or more normal-length sentences. We therefore check at this point to see if one of the next three most important sentences could be added to the summary without exceeding the word count. If so, we add it and continue adding the sentences after it, following the same rule when we encounter another sentence that would make the summary exceed the maximum word count. But if none of the next three most important sentences is short enough to be added, we stop adding sentences to the summary, to avoid adding a collection of unimportant but very short sentences simply because they fit.

Once the system has identified the sentences to include in the summary, it puts them in the order in which they appeared in the original document.

Experimental Evaluation

Evaluation Method

Human-judged evaluation is necessary, since automatic evaluation using ROUGE would not be appropriate for this experiment. ROUGE uses n-gram overlap between the summary and model summaries. That would bias the evaluation against the simplification system, which substitutes simpler words for more complex ones. In addition, automatic evaluations have not been shown to correlate with human evaluations of summaries when the sentences included in the summary are not precise cut-and-paste extractions from the original document. Thus, we use two types of human evaluation.

DOI: 10.1002/asi

TABLE 3. Examples of multiple choice comprehension questions used for extrinsic evaluation.

What part of the lower court's decision does Rodriguez-Ocampo appeal?
a. Denial of his motion to suppress statements made before he was advised to his right to counsel
b. His conviction of two counts of illegal entry for a single offense
c. Allowance of the prosecution's motion to admit portions of his juvenile record
d. A sixteen-level sentencing enhancement
e. I can't answer this question using this summary.
What method did the researchers use to collect data?
a. Review of medical records
b. Pre- and post-operative interviews with patients and their families
c. fMRIs 3, 6, 9, and 12 months post-surgery, and annually thereafter
d. Self-report using a smart-phone app
e. I can't answer this question using this summary.

TABLE 4. Word count in the original document and mean word count across summaries of that document.

Original document	Summaries (mean)
2,011	611
4,110	632
9,013	1,701
7,156	1,416
4,745	843
5,594	1,076

The first is an intrinsic evaluation, where evaluators rate each summary on the DUC quality questions.[8] As noted by Ou et al. (2007), human participants can judge a summary directly, rather than compare it to an "ideal" summary.

The second is an extrinsic evaluation. As Daumé and Marcu (2005) argue, extrinsic evaluation has an advantage over intrinsic evaluation: it does not merely determine how similar the output is to the way a human would do it, but determines how useful the output is for the desired task. If we wish to generate summaries that convey important information from a document, we should test whether someone who has read the summary has absorbed that information. We therefore followed Otterbacher et al. (2008)'s approach, developing comprehension questions and comparing evaluators' ability to answer those questions using summaries output by the different pipelines.

Specifically, we randomly selected three cases and three articles from the test set of the corpus described earlier.[9] For each document, one researcher wrote five comprehension questions.[10] Each question is multiple choice and can be answered based on the full text of the original document with no outside knowledge. Example questions appear in Table 3. Each evaluator was asked to answer the questions regarding the cases and articles.

For each biomedical article, the evaluator was assigned to one of four experimental treatments, which determined what form of summary would be given to them:

1. "LexRank Only": Summary of source document with no simplification, disaggregation, or compression of sentences.
2. "Simplified": Summary of document after sentence simplification.
3. "Disaggregated": Summary of document after sentence disaggregation.
4. "Compressed": Summary of document after sentence compression.

For legal cases, evaluators could be assigned to any of treatments one through four, or a fifth condition:

5. "Human": Human-generated summary.

Fourteen evaluators participated in the experiment. All were college graduates. Three reported at least 1 year of law school. Two reported studying biology or medicine at the college level or above for a year or more.

Each evaluator was randomly assigned to one condition per original document. We then adjusted the assignments slightly so that each condition/document pairing received approximately equal numbers of evaluations (either three or four). Finally, we ensured that no evaluator saw the same document in more than one condition and that no evaluator saw the same condition for more than two documents.

Evaluators were told that they would answer "opinion questions" asking about their impressions of the quality of the summary and "information questions" to help the researchers "understand how well the summary conveys information from the original document." The instructions acknowledged that summaries could contain grammatical errors or be missing information. Since the goal of this extrinsic evaluation was to determine how useful output summaries are, and summaries that fail to convey information—either because they do not contain it or because they are so ungrammatical that they are too difficult to read—are unlikely to be useful, the instructions pointed out the availability of option E for questions the evaluator could not answer. Evaluators were asked not to consult outside sources to answer the questions.

As noted, the word count upper bound for biomedical article summaries was 20% of the word count of the original document, and for legal cases it was the word count in the corresponding human-written summary. Actual word counts for the six test documents are shown in Table 4.

[8]http://duc.nist.gov/duc2004/quality.questions.txt.

[9]Since humans were evaluating the summaries, we began with a smaller number of summaries than we could have evaluated with ROUGE. However, as discussed in the Results and Discussion, *infra*, clear patterns emerged even with only six documents, so it was unnecessary to test using a larger data set.

[10]We limited the number to five to avoid overwhelming our volunteer evaluators. When writing questions, we sought to cover a cross-section of topics, rather than all topics in each document.

DOI: 10.1002/asi

Results

For each summary evaluation, we calculated a single comprehension score, which reflected the number of comprehension questions about that summary that an evaluator answered correctly. Since we asked five comprehension questions for each summary, these scores could range from zero to five. Mean comprehension scores appear in Table 5 and Figure 4.

As expected, scores on comprehension questions were best when evaluators used human-written summaries. Contrary to our expectation, however, the next highest performances were those in the LexRank Only group, followed by the Simplification group, with the Disaggregation and Compression groups lagging behind. A one-way analysis of variance (ANOVA) using Tukey's honest significant difference (HSD) revealed that the differences between LexRank and Compression, LexRank and Disaggregation, and Human and Compression were significant at the $p = .05$ level; the other differences were not significant. A two-way ANOVA to determine whether the document type (legal case vs. biomedical article) interacted with condition revealed no significant effects.

Because the quality questions seek information about different aspects of summary quality, we could not combine the seven quality questions as we could the comprehension questions. Thus, each summary evaluation received seven different quality scores, Q.1 through Q.7. Each score ranged from 4 points for "a," the most positive evaluation, to 0 for "e," the most negative evaluation.

Performance of the different systems on the quality questions is summarized in Figure 5. Although the ratios between the systems differ, the pattern for all questions except for 3 and 5 is quite similar to the pattern seen in

TABLE 5. Mean comprehension scores for each treatment condition.

Condition	Mean
Simplified	3.11
Compressed	2.00
Disaggregated	2.26
LexRank Only	3.50
Human	3.75

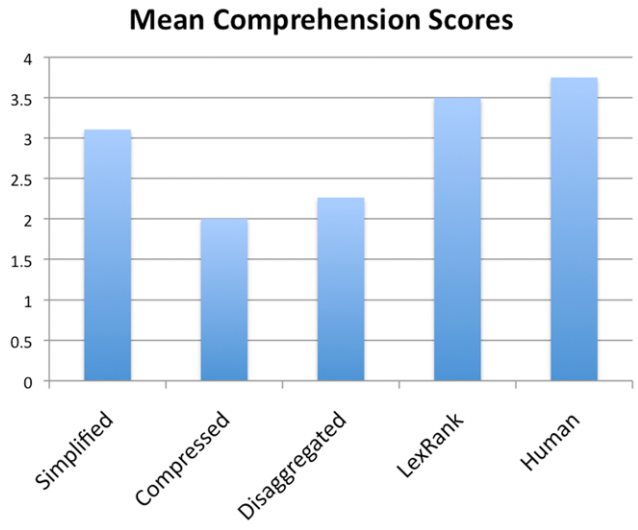

FIG. 4. A comparison of mean comprehension scores achieved by the different systems. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

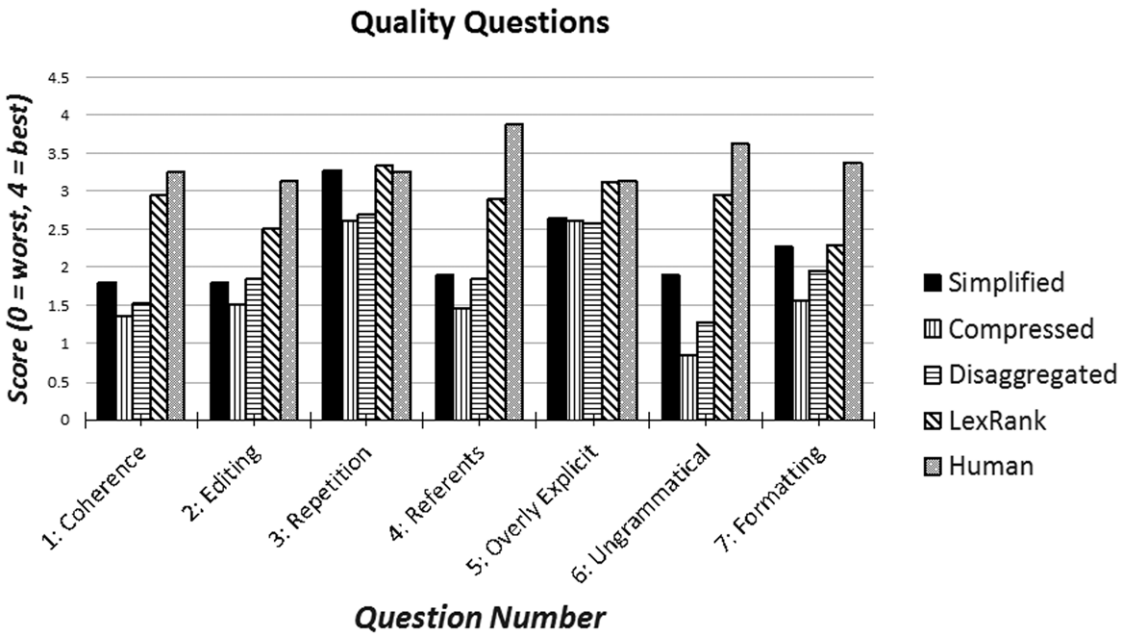

FIG. 5. A comparison of scores of the different systems on the DUC quality questions.

DOI: 10.1002/asi

TABLE 6. Pearson's r correlation coefficients between scores on comprehension questions and quality questions. All are significant at the .01 level or better.

Quality question	1	2	3	4	5	6	7
Pearson's r	0.410	0.438	0.288	0.418	0.330	0.440	0.286

Figure 4. As expected, the human-generated summaries usually scored best on the quality questions. On questions 3 and 5, the LexRank Only summaries performed as well as or better than the human-generated summaries; however, the differences were not significant. Generally, summaries from the Simplified condition performed slightly worse on the quality questions than the LexRank Only summaries, followed by Disaggregated, with Compression ranked last. Significance of differences varied. On questions 3 and 5, there were no significant differences between any of the conditions. LexRank Only and Human were never significantly different. Means and significant differences for all seven quality questions in all five conditions are available online at http://www-personal.umich.edu/~cfdollak/sophisticated_documents_paper/quality_questions_appendix.pdf.

Except for questions 3 and 7, all possible pairings of the quality question scores showed a significant positive correlation using Pearson's r. Scores on information questions were significantly positively correlated with scores on all seven quality questions; these correlations appear in Table 6.

Discussion

Analysis of Results on Quality Questions

With the exception of question 3, the results on the quality questions followed a consistent pattern of Human/LexRank performing best, followed by Simplification, Disaggregation, and Compression, in that order. It is not surprising that the human-written summaries perform well on this portion of the evaluation; if anything, it is surprising that they did not perform even better than they did.

The lack of a significant difference between LexRank and Human conditions on any of the quality evaluations appears impressive, although we must be cautious not to read too much into it. Because the Human condition applied only to the legal cases, not to the biomedical articles, the sample size of the Human group was smaller than those of the other conditions ($N_{Human} = 8$; $N_{LexRank} = 18$). A larger study might find significant results where this one did not. In addition, it is possible that by showing evaluators summaries from the much lower quality Simplified, Compressed, or Disaggregated conditions alongside LexRank and Human summaries, we may have distorted their perception of the quality scale. A truly fair test of LexRank's performance would require a larger number of summaries and only the Human and LexRank conditions.

We can say with confidence, however, that the LexRank summaries outperformed the Compressed and Disaggregated ones on all of the quality questions and the Simplified system on most of them. On some quality questions, this is unsurprising. The extra step of simplifying, compressing, or disaggregating a sentence adds an opportunity for a previously grammatical sentence to become ungrammatical. In addition, removing part of a sentence increases the likelihood of having unclear referents.

However, our hypothesis suggested that on quality question 2, which asks how much useless or confusing material ought to be removed from the summary, we could expect the Simplified, Compressed, and Disaggregated conditions to perform well. After all, their purpose was to automatically remove useless material. Yet all three performed significantly worse on this question than the Human condition, and Compressed also performed significantly worse than LexRank.

One possible explanation is that the compound question has blurred the issue. Perhaps the evaluators felt that they would need to edit a great deal of confusing material from the Simplified, Compressed, and Disaggregated summaries, but were neutral or even positive as to the amount of useless material. Another possibility is that the experimental systems rendered some sentences so confusing as to be useless.

A review of the actual summaries supports this second view. Table 7 contains the first hundred words from the LexRank, Simplified, Compressed, and Disaggregated versions of the same summary.[11] In the Compressed and Disaggregated summaries, and to a lesser extent the Simplified summary, sentences have been altered and taken out of context in ways that make them quite confusing. For instance, in the Simplified summary the heading "Austin Bank—Troup, Texas," which in the original document introduced a recitation of the facts regarding the robbery of that bank, was transformed into the rather bewildering sentence "Austin Bank is Troup, Texas." This type of change explains why, even if some useless material was excluded by the Simplified, Compressed, and Disaggregated conditions, evaluators would still score these conditions as needing to have more useless or confusing material trimmed.

Additionally, we note that question 3, which asks whether the summary is repetitive, is the only quality question that does not follow the pattern of the others, as can be seen in Figure 5. For question 3 only, the Simplification system performed approximately as well as the LexRank Only and Human conditions. Because the differences in question 3 were not statistically significant, this apparent deviation from the pattern may be illusory. But this result could also suggest that the simplifier's use of determiners instead of repeated noun phrases helped avoid overly repetitive summaries.

Another notable result was that Simplified and Disaggregated were not significantly different on any quality measures except for question 6, which asks about ungrammatical sentences. Given the engineering described previously to avoid introducing ambiguous determiners and to repeat modifiers

[11]The summaries in their entirety are available online at http://www-personal.umich.edu/~cfdollak/sophisticated_documents_paper/example_summaries.pdf.

DOI: 10.1002/asi

TABLE 7. Segments of summaries output by (a) LexRank Only, (b) Simplified, (c) Disaggregated, and (d) Compressed conditions.

a: (2010) UNITED STATES of America, Plaintiff-Appellee, v. Paul Edward THOMAS; Derrick Van Hodges, Defendants-Appellants. Each robbery was completed within two minutes. The basis for the warrant was DNA evidence linking Hodges to a glove dropped during a bank robbery in Henderson, Texas. Thomas and Hodges were named in an 18-count indictment charging them with conspiracy, bank robbery, and weapons offenses related to the following bank robberies: 1. DISCUSSION I. Sufficiency of the Evidence Thomas and Hodges argue the government presented insufficient evidence identifying them as the bank robbers. Thomas claims the evidence against Hodges was much stronger and implies that Thomas was found guilty by association.

b: Each robbery was completed within two minutes. Thomas and Hodges were named in an 18-count indictment charging them with conspiracy, bank robbery, and weapons offenses related to the following bank robberies: 1. Austin Bank is Troup, Texas. DISCUSSION I. Sufficiency of the proof Thomas and Hodges argue the government presented insufficient evidence identifying them as the bank robbers. Thomas claims and the proof against Hodges implies that association found Thomas guilty. The proof was much stronger. Hodges was ‘weak in some instances. This happened after a bank robbery was also ‘happenstance. United States v. Clayton. We apply this standard of review to direct and circumstantial evidence.

c: (2010) UNITED STATES of America, Plaintiff-Appellee, v. Paul Edward THOMAS; Derrick Van Hodges, Defendants-Appellants. Ireland, Carroll, Kelley for Thomas is P.C. Ireland, Carroll, Kelley for Thomas is Tyler. Ireland, Carroll, Kelley for Thomas is TX. Hodges had in his possession a $ 10 bait bill taken a week earlier during the robbery of a bank in Crockett, Texas. At that time, arrested. America Henderson, is Texas; 3. Austin Bank is Troup, Texas. Thomas, Hodges, were jointly tried before a jury. Thomas claims.the evidence against Hodges implies that Thomas was found guilty by association. The evidence against Hodges was much stronger.

d: LESLIE SOUTHWICK Circuit Judge Paul Edward Thomas and Derrick Van Hodges were convicted of counts of conspiracy bank robbery and weapons possession. Bank robberies were executed in manner. Each robbery was completed within two minutes. Derrick Van Hodges was arrested in Tyler Texas on state warrant. The basis for the warrant was evidence linking Hodges to glove dropped during a bank robbery in Henderson Texas. Bait bill was found in child’s bedroom. DISCUSSION Sufficiency of Evidence Thomas and Hodges argue government presented evidence identifying them as bank robbers. Thomas claims the evidence against Hodges was stronger and implies that Thomas was found guilty by association.

in the split sentences, we had expected that the Disaggregated system would perform better on question 4, which asks how difficult it was to identify referents of noun phrases.

An important question is why Disaggregated summaries included more ungrammatical sentences than Simplified summaries. The most likely cause seems to be overfitting. The modifications and the additional rules described earlier were based on a small number of gold standard sentences. Changes to the system intended only to allow modifiers to be repeated generated some side effects in initial testing; for example, some disaggregated sentences contained long strings of repeated conjunctions. These problems were fixed, in that they no longer occurred when tested on the gold standard sentences and basic modifications to them, before the summaries in this study were generated. However, perhaps the gold standard sentences and modifications to them did not expose a broad enough range of possible problems, and other side effects remained that could only have been discovered by a larger system test. Similarly, the new rules were developed to work on the sentences from the gold standard collection and variations on those. It is possible broader testing might reveal sentences that match the dependency patterns we found in those sentences, but are grammatically different enough that the rule application no longer makes sense.

Perhaps the most surprising result on the quality questions was the poor performance of the Compression system. The Clarke and Lapata ILP-based sentence compression algorithm that we used is widely considered state-of-the-art in sentence compression. We suspect that the problem may relate to the language model that the algorithm incorporates in its objective function. Maybe the trigram model built from the LDC Gigaword corpus of newswire articles does a poor job of representing n-grams that show up in legal cases and biomedical articles. Biomedical articles in particular are likely to contain a great many out-of-vocabulary terms. A simple follow-up study could test this hypothesis by building a language model using a corpus of sophisticated documents and checking if performance improved. It would not require evaluating entire summaries, but could be evaluated on a sentence-by-sentence basis.

Analysis of Results on Comprehension Questions

The Compressed condition seems to have suffered from the predicted problem: The compression algorithm does not know what information will be important to the summary overall, and so it sometimes omits words that are actually needed to make a sentence meaningful. Given the paragraph

> The mean time to loss of Engel Class II status after STL was 15.2 years (95% CI 13.2–17 years), and after mtg-SelAH it was 13.8 years (95% CI 11.9–16.2 years).
> The difference was not significant ($p = 0.536$).
> The mean time to loss of Engel Class I status after STL was 15.2 years (95% CI 13.2–17 years), and after mtg-SelAH it was 13.1 years (95% CI 11.9–16.2 years).
> The difference was not significant ($p = 0.536$).
> The mean time to loss of Class IA status after STL was 14.6 years (95% CI 12.2–17 years), and after mtg-SelAH it was 7.9 (95% CI 6.1–9.7 years).
> The difference was significant ($p = 0.034$) (Figure 1). (22)

the compression was

> The time to loss of Engel Class status was years years and it was years years.
> The difference was not significant.
> The time to loss of Engel Class I status was years years and it was years years.
> The difference was not significant.

DOI: 10.1002/asi

> The time to loss of Class IA status was years years and it was 7.9 years.
> The difference was. (23)

The original paragraph might have allowed the reader to infer that the answer to the comprehension question

> Which of the following best describes the differences in seizure outcomes between the group that underwent standard temporal lobectomy (STL) and the group that underwent selective amygdalohippocampectomy (SelAH)? (24)

is most likely b,

> b. The groups showed no difference in time to loss of Engel Class I or II status; STL performed better on time to loss of Engel Class IA status; and the SelAH group had more seizures during attempted medication withdrawal. (25)

since there was no significant difference between the groups on time to loss of Engel Class II or Engel Class I status, but there was a significant difference in time to loss of Class IA status. No such inference can be drawn from the compressed version of the paragraph, since important information is missing.

Since the Disaggregation condition generated the second-worst performance of all systems, however, the results on the comprehension questions do not support the hypothesis that the problem of compression removing important meaning will be solved by splitting, rather than compressing, the sentences. Instead, the similarity between the performance of Disaggregation and Compression suggests that they may suffer from similar problems.

There are two possible explanations for the poor performance of the Disaggregation and Compression systems on the comprehension measure. The first is that the information needed to answer the questions was simply not in the summaries, and the second is that the poor quality of the summaries obscured answers that were in fact present.

If the information needed to answer the questions was simply not in the summaries, we need to understand why it was missing. The maximum word count for each summary was fixed across conditions, so all summaries had the opportunity to include the same amount of information. But the Disaggregation condition tended to generate repetitive sections of text. The LexRank algorithm has a weakness for repetition of a phrase in multiple sentences; it construes this as similarity between the sentences, and the sentences therefore "vote" for each other to be included in the summary. Thus, when faced with a case that listed counsel as

> Laurel Franklin Coan, Jr., Asst. U.S. Atty. (argued), Robert James Middleton, Tyler, TX, for U.S.
> Deborah Johnson Race (argued), (Court-Appointed), Ireland, Carroll & Kelley, P.C., Tyler, TX, for Thomas. (26)

the disaggregator mistakenly combined the fragments, then split them into

> Laurel Franklin Coan, Jr. Asst.
> U.S. Atty.
> (argued), Robert James Middleton, Tyler TX for U.S. Deborah Johnson Race ( ), (Court-Appointed), Ireland, Carroll, Kelley for Thomas,.
> Tyler TX for U.S. Deborah Johnson Race ( ), is argued.
> Ireland, Carroll, Kelley for Thomas is P.C.
> Ireland, Carroll, Kelley for Thomas is Tyler.
> Ireland, Carroll, Kelley for Thomas is TX. (27)

And LexRank, noticing the repetition, included the last three lines in the summary. Flaws like this take up space that could be used to convey actually important information.

A second explanation is that, although the information is present, human readers could not extract it, because the generated summaries were so difficult to read and comprehend that they obscured the answers. Consider the comprehension question that was least often answered correctly:

> What effect did surgery type have on psychiatric outcomes?
> a. The STL group experienced increased depression
> b. The STL group had increased paranoia, while the SelAH group had decreased paranoia
> c. The SelAH group had increased paranoia, while the STL group had increased depression
> d. The SelAH group had increased depression and anxiety
> e. I can't answer this question using this summary. (28)

The correct answer is b. The only evaluator who answered it correctly had a summary from the Compressed condition. The summary included the following relevant sentences:

> Standard temporal was associated with higher scores on assessment of paranoia.
> Our concern was that STL cause rates of de novo psychosis as have been associated with it .9,25,27 not a patient in group was diagnosed psychosis.
> Gyrus SelAH be procedure for patients with high levels of disease paranoia. (29)

In this case, the information needed to answer the question—that STL was associated with increased paranoia—was in the summary; however, an evaluator could easily misunderstand the second sentence to mean the researchers were concerned that STL would cause psychosis, but not a single patient actually exhibited psychosis.

The significant, positive correlation between comprehension score and all of the quality question scores lends some support to this last explanation: Evaluators had more difficulty correctly answering comprehension questions as the subjective quality of the summary declined. Further study could ask annotators given a summary and a comprehension question with the correct answer to try to mark in the summary where the answer can be found, or indicate if the answer is not in the summary.

DOI: 10.1002/asi

Qualitative Analysis of Shortened Sentences

While our quantitative results focus on the entire summarization pipeline, a brief qualitative examination of the sentence shortening methods on their own is enlightening.

The sentence compression method sometimes removed unneeded phrases, as in a compression that removed "On the other hand" from

> On the other hand, argument of counsel is not evidence and is not to be considered as such by the jury. (30)

However, some compressions were ungrammatical; for instance,

> The DNA evidence and bait bills constitute sufficient evidence against Thomas to sustain convictions relating to the first and fifth bank robberies, and sufficient evidence against Hodges to sustain convictions relating to the first, second, fourth, and fifth bank robberies. (31)

was compressed to

> The evidence and bait bills constitute evidence against Thomas sustain convictions relating to the first and bank robberies and evidence against Hodges sustain convictions relating to bank robberies. (32)

Additionally, the compressions often excluded words that were important to the meaning of the sentence; for instance,

> Thomas and Hodges argue the government presented insufficient evidence identifying them as the bank robbers. (33)

was compressed to

> Thomas and Hodges argue government presented evidence identifying them as bank robbers. (34)

Thus, as predicted, compression may substantially alter sentences' meaning.

Simplification, too, succeeded on some sentences; for instance, it neatly split

> [W]e view the evidence and the inferences drawn therefrom in the light most favorable to the verdict, and we determine whether a rational jury could have found the defendant guilty beyond a reasonable doubt. (35)

into

> (W)e view the proof and the inferences drawn therefrom in the light most favorable to the verdict.
> And we determine whether a rational jury could have found the defendant guilty beyond a reasonable doubt. (36)

Yet the simplified sentences also suffer from some of the predicted problems. For example, sometimes lexical simplification changed the meaning, as when "challenging the lack of eyewitness identification" was simplified to "challenging the rarity of eyewitness identification." In addition, simplification from passive to active voice sometimes obscured meaning; for instance, the simplification containing "association found Thomas guilty" could confuse a reader, while the disaggregation of the same sentence maintained the clearer statement that "Thomas was found guilty by association." And as predicted, the simplifier used "this" while the disaggregator kept the entire noun phrases; for instance, where a simplification included

> This eyewitness stepped outside of his office to observe traffic. (37)

the corresponding disaggregation included

> Another eyewitness stepped outside of his office to observe traffic. (38)

as we hoped.

The disaggregation system still includes some simplification rules that should be changed for summarization. For instance, the original sentence

> Still, if a joint trial would prejudice a defendant, district courts may sever the defendants' trials. (39)

should not have been disaggregated into

> Suppose a joint trial would prejudice a defendant.
> Then still district courts may sever the defendants' trials. (40)

Also, disaggregation occasionally caused some severe grammatical problems, particularly involving missing conjunctions.

Strategies for Improvement

Determining the cause of the difficulty in answering the comprehension questions helps determine the best way to improve future systems. If the problem is that valuable information is being omitted from the summary because some characteristic of the disaggregated or compressed sentences causes suboptimal sentence selection, we might need to try using a different sentence selection algorithm. For instance, C-LexRank is an algorithm designed for multidocument summarization; it might handle the repetition that disaggregation adds to a document better than LexRank does.

If the problems are due to confusing, ungrammatical output of the disaggregator, however, then the focus of future work should be on improving disaggregation as a stand-alone function before doing further work on the summarization pipeline. Such future work could go in two different directions: improving the existing disaggregation method within the framework of the modified simplifier, or developing an entirely new method of disaggregation.

Certainly, there is room for improvement within the existing framework. Given that disaggregation introduced

DOI: 10.1002/asi

grammatical problems that simplification did not, it is apparent that testing of the disaggregation system on a broader collection of sentences is necessary to identify the source(s) of the problems.

However, there are several significant problems with the existing framework that make developing a new approach the better option. First, the disaggregator relies on a correct dependency parse to allow it to correctly split a sentence. However, we observed an unusually high rate of incorrect parses of sentences from the sophisticated documents; there were numerous problems with attachment, and some of the dependency relationships were not applied consistently across similar sentences. The problem may well arise because of the complexity of the sentences; the legal cases had a maximum parse tree depth of 47. Whatever the cause, the potential inaccuracy of the parser limits the ability to correctly disaggregate using this system.

Second, this framework cannot cope with a sentence structure that is very common to the longest sentences in legal cases: a numbered list within the sentence. For instance, our gold standard disaggregation transformed a sentence that began

> The request for a hearing must: (i) Provide a specific statement of the issue of law or fact to be raised or controverted . . . ; (ii) Provide a brief explanation of the basis for the contention; (iii) Demonstrate that the issue raised in the contention is within the scope of the proceeding; . . .. (41)

into a series of sentences

> The request for a hearing must provide a specific statement of the issue of law or fact to be raised or controverted.
> The request for a hearing must provide a brief explanation of the basis for the contention.
> The request for a hearing must demonstrate that the issue raised in the contention is within the scope of the proceeding. . . . (42)

The current framework has no way to iterate through list items. Thus, disaggregating a six-item list requires a rule for six-item lists, but that rule would not generalize to five-item lists. Writing rules for each number of items that might be on a list does not solve the problem, since each list item requires several variables, and the running time of the disaggregator becomes impractical when more than about 10 variables are used. This system thus cannot effectively disaggregate list sentences, yet ignoring them leaves intact many of the longest sentences in legal cases.

Finally, the current framework is purely syntactic, but disaggregation is not a purely syntactic task. Consider the sentence

> Rather than argue explicitly about the findings of the NRC, as to whether the portions of the contention met the reopening and/or the admissibility standards, in rejecting the Commonwealth's contention, Massachusetts devotes a substantial portion of its brief to arguing that the NRC acted arbitrarily and capriciously. (43)

Here, a purely syntactic approach cannot determine who rejected the Commonwealth's contention—the NRC or Massachusetts. Either semantic knowledge that "Massachusetts" and "the Commonwealth" refer to the same entity or some form of coreference resolution is necessary to know that a disaggregation containing

> In rejecting the Commonwealth's contention, Massachusetts did not argue explicitly about the findings of the NRC. (44)

would not be accurate, while a disaggregation containing

> Massachusetts does not argue explicitly about the findings of the NRC in rejecting the Commonwealth's contention. (45)

would be.

For these reasons, future work should explore other ways of disaggregating sentences involving the retention of semantic coherence, rather than continuing to rely on the modified simplification framework.

Conclusion

We found that sentence simplification, compression, and disaggregation before extractive summarization of sophisticated documents did not improve performance on extrinsic evaluations. The most likely reason is that, when applied to sentences from the legal and biomedical domains, all three sentence-shortening techniques produced some confusing or ungrammatical output. Future work should focus on improving the ability of sentence shortening techniques to handle sentences from these domains.

Acknowledgments

We would like to thank the members of the CLAIR research group at the University of Michigan for their advice and feedback about this work, as well as the 14 volunteers who evaluated summaries for this study. We gratefully acknowledge support from the Richard H. Orenstein Fellowship in Memory of Murray Orenstein.

References

Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In I. Mani & M. Maybury (Eds.), Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, Vol. 17 (pp. 10–17). Madrid, Spain: Association for Computational Linguistics.

Berg-Kirkpatrick, T., Gillick, D., & Klein, D. (2011). Jointly learning to extract and compress. In Y. Matsumoto & R. Mihalcea (Eds.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 481–490). Portland, Oregon, USA: Association for Computational Linguistics.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., & Tait, J. (1998). Practical simplification of English newspaper text to assist aphasic readers. Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology (pp. 7–10). Madison, Wisconsin: The AAAI Press, Menlo Park, California.

Chali, Y., & Hasan, S.A. (2012). On the effectiveness of using sentence compression models for query-focused multi-document summarization.

DOI: 10.1002/asi

In M. Kay & C. Boitet (Eds.), Proceedings of COLING 2012 (pp. 457–474). Mumbai, India: The COLING 2012 Organizing Committe.
Chandrasekar, R., & Srinivas, B. (1997). Automatic induction of rules for text simplification. Knowledge-Based Systems, 10(3), 183–190.
Clarke, J., & Lapata, M. (2007). Modelling compression with discourse constraints. In J. Eisner (Ed.), EMNLP-CoNLL (pp. 1–11). Prague, Czech Republic: Association for Computational Linguistics.
Clarke, J., & Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. Journal of Artificial Intelligence Research, 31, 399–429.
Clarke, J., & Lapata, M. (2010). Discourse constraints for document compression. Computational Linguistics, 36(3), 411–441.
Daumé, H., III, & Marcu, D. (2005). Bayesian summarization at duc and a suggestion for extrinsic evaluation. Proceedings of the Document Understanding Conference, DUC-2005. Vancouver, WA: Vancouver, B.C., Canada. See http://duc.nist.gov/pubs/2005papers/isi.daume.pdf
Erkan, G., & Radev, D.R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22(1), 457–479.
Graff, D., & Cieri, C. (2003). English Gigaword LDC2003T05.
Grefenstette, G. (1998). Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In E. Hovy & D. Radev (Eds.), Working notes of the AAAI Spring Symposium on Intelligent Text summarization (pp. 111–118). Menlo Park, California: The AAAI Press.
Hachey, B., & Grover, C. (2006). Extractive summarisation of legal texts. Artificial Intelligence and Law, 14(4), 305–345.
Haghighi, A., & Vanderwende, L. (2009). Exploring content models for multi-document summarization. In K.-Y. Su, J. Su, J. Wiebe & H.Z. Li (Eds.), Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 362–370). Suntec, Singapore: Association for Computational Linguistics.
Hand, T.F. (1997). A proposal for task-based evaluation of text summarization system. In I. Mani & M. Maybury (Eds.), Proceedings of the Association for Computational Linguistics/European Association for Computational Linguistics Summarization Workshop (pp. 31–38). Madrid, Spain: Association for Computational Linguistics.
Hovy, E., & Lin, C.-Y. (1998). Automated text summarization and the summarist system. In D.E.M. Penrose (Ed.), Proceedings of a Workshop on Held at Baltimore, Maryland: October 13–15, 1998 (pp. 197–214). Baltimore, Maryland, USA: Association for Computational Linguistics.
Jha, R., Coke, R., & Radev, D. (2015). Surveyor: A system for generating coherent survey articles for scientific topics. In B. Bonet & S. Koenig (Eds.), Proceedings of the Twenty-Ninth AAAI Conference (pp. 2167–2173). Austin, Texas USA: The AAAI Press.
Jonnalagadda, S., & Gonzalez, G. (2010). Sentence simplification aids protein-protein interaction extraction. arXiv preprint arXiv:1001.4273.
Klebanov, B.B., Knight, K., & Marcu, D. (2004). Text simplification for information-seeking applications. In T. Zahir & C. Angelo (Eds.), On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE (pp. 735–747). Agia Napa, Cyprus: Springer.
Knight, K., & Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. Artificial Intelligence, 139(1), 91–107.
Li, C., Liu, F., Weng, F., & Liu, Y. (2013). Document summarization via guided sentence compression. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu & S. Bethard (Eds.), EMNLP (pp. 490–500). Seattle, Washington, USA: Association for Computational Linguistics.
Lin, C.-Y. (2003). Improving summarization performance by sentence compression: A pilot study. Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages-Volume 11 (pp. 1–8). Sappro, Japan: Association for Computational Linguistics.
Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In S. Szpakowicz & M.-F. Moens (Eds.), Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.
Luhn, H.P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159–165.
Mani, I. (2001). Automatic summarization. Natural language processing. New York, USA: J. Benjamins Publishing Company.
Mani, I., & Maybury, M.T. (Eds.). (1999). Advances in automatic text summarization. Cambridge, MA: MIT Press.
Martins, A.F., & Smith, N.A. (2009). Summarization with a joint model for sentence extraction and compression. In J. Clarke & S. Riedel (Eds.), Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing (pp. 1–9). Boulder, Colorado: Association for Computational Linguistics.
McKeown, K., Passonneau, R.J., Elson, D.K., Nenkova, A., & Hirschberg, J. (2005). Do summaries help? Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 210–217). Salvador, Brazil: ACM.
Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into texts. In D. Lin & D. Wu (Eds.), Proceedings of EMNLP, Vol. 4 (pp. 404–411). Barcelona, Spain: Association for Computational Linguistics.
Morris, A.H., Kasper, G.M., & Adams, D.A. (1992). The effects and limitations of automated text condensing on reading comprehension performance. Information Systems Research, 3(1), 17–35.
Murray, G., Kleinbauer, T., Poller, P., Becker, T., Renals, S., & Kilgour, J. (2009). Extrinsic summarization evaluation: A decision audit task. ACM Transactions on Speech and Language Processing (TSLP), 6(2), 2.
Nenkova, A., Passonneau, R., & McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Transactions on Speech and Language Processing (TSLP), 4(2), 1–23.
Otterbacher, J., Radev, D., & Kareem, O. (2008). Hierarchical summarization for delivering information to mobile devices. Information Processing & Management, 44(2), 931–947.
Ou, S., Khoo, C.S.G., & Goh, D.H. (2007). Automatic multidocument summarization of research abstracts: Design and user evaluation. Journal of the American Society for Information Science and Technology, 58(10), 1419–1435.
Piwowarski, B., Amini, M.R., & Lalmas, M. (2012). On using a quantum physics formalism for multidocument summarization. Journal of the American Society for Information Science and Technology, 63(5), 865–888.
Radev, D.R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. Computational Linguistics, 28(4), 399–408.
Rastkar, S., Murphy, G., & Murray, G. (2014). Automatic summarization of bug reports. IEEE Transactions on Software Engineering, 40(4), 366–380.
Rath, G.J., Resnick, A., & Savage, T.R. (1961). The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines. American Documentation, 12(2), 139–141.
Reeve, L.H., Han, H., & Brooks, A.D. (2007). The use of domain-specific concepts in biomedical text summarization. Information Processing & Management, 43(6), 1765–1776.
Siddharthan, A. (2002). An architecture for a text simplification system. In O. Vikas (Ed.), Language Engineering Conference, 2002. Proceedings (pp. 64–71). Hyderabad, India: IEEE.
Siddharthan, A. (2006). Syntactic simplification and text cohesion. Research on Language and Computation, 4(1), 77–109.
Siddharthan, A. (2010). Complex lexico-syntactic reformulation of sentences using typed dependency representations. In J. Kelleher, B. Mac Namee & I. van der Sluis (Eds.), Proceedings of the 6th International Natural Language Generation Conference (pp. 125–133). Trim, Co. Meath, Ireland: Association for Computational Linguistics.
Siddharthan, A., & Angrosh, M.A. (2014). Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In S. Wintner, S. Goldwater & S. Riezler (Eds.), Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014) (pp. 722–731). Gothenburg, Sweden.
Siddharthan, A., Nenkova, A., & McKeown, K. (2004). Syntactic simplification for improving content selection in multi-document

DOI: 10.1002/asi

summarization. Proceedings of the 20th international conference on Computational Linguistics (pp. 407–414). Barcelona, Spain: Association for Computational Linguistics.

Thadani, K. (2014). Approximation strategies for multi-structure sentence compression. In K. Toutanova & H. Wu (Eds.), Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). (pp. 1241–1251). Baltimore, MD: Association for Computational Linguistics.

Thadani, K., & McKeown, K. (2013). Sentence compression with joint structural inference. In J. Hockenmaier & S. Riedel (Eds.), Proceedings of CoNLL (pp. 65–74). Sofia, Bulgaria.

Tiersma, P.M. (1999). Legal Language. Chicago: University of Chicago Press.

Turner, J., & Charniak, E. (2005). Supervised and unsupervised learning for sentence compression. In K. Knight, H.T. Ng & K. Oflazer (Eds.), Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 290–297). Ann Arbor, Michigan: Association for Computational Linguistics.

Vanderwende, L., Suzuki, H., Brockett, C., & Nenkova, A. (2007). Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. Information Processing & Management, 43(6), 1606–1618.

Wubben, S., Van Den Bosch, A., & Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In H.Z. Li, C.-Y. Lin, M. Osborne, G.G. Lee & J.C. Park (Eds.), Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (pp. 1015–1024). Jeju Island, Korea: Association for Computational Linguistics.

Yang, C.C., & Wang, F.L. (2008). Hierarchical summarization of large documents. Journal of the American Society for Information Science and Technology, 59(6), 887–902.

Zajic, D., Dorr, B.J., Lin, J., & Schwartz, R. (2007). Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. Information Processing & Management, 43(6), 1549–1570.

DOI: 10.1002/asi