

# Crowd-Sourced Iterative Annotation for Narrative Summarization Corpora

Jessica Ouyang and Serina Chang and Kathleen McKeown

Department of Computer Science, Columbia University, New York, NY 10027

{ouyangj, kathy}@cs.columbia.edu

sc3003@columbia.edu

## Abstract

We present an iterative annotation process for producing aligned, parallel corpora of abstractive and extractive summaries for narrative. Our approach uses a combination of trained annotators and crowd-sourcing, allowing us to elicit human-generated summaries and alignments quickly and at low cost. We use crowd-sourcing to annotate aligned phrases with the text-to-text generation techniques needed to transform each phrase into the other. We apply this process to a corpus of 476 personal narratives, which we make available on the Web.

## 1 Introduction

With the tremendous amounts of text published on the Web every day, automatic text summarization is more relevant than ever. Web content must compete for readers' attention, and the existence of click bait links shows that content providers are very aware that a short, appealing summary may be their only chance to attract readers. For the readers' part, summaries stating exactly what a piece of content is about protects them from wasting time on topics that do not interest them.

Research on summarization has long focused on extraction: selecting the most salient sentences from a text without any modifications. These summaries can be incoherent or incomprehensible due to unresolved pronouns and references, and sentences containing irrelevant information (Nenkova and McKeown, 2011), and this is particularly problematic with informal web text. Thus abstractive summarization is critical for the web, with rewriting of extracted sentences, as humans write summaries (Jing and McKeown, 1999).

To develop an abstractive summarization system, we need data: parallel corpora that align extractive summaries with abstractive summaries. Such corpora would allow researchers to develop text-to-text generation approaches to produce abstractive summaries from extractive ones. While there are many summarization corpora available, most provide abstractive summaries only (Meyer et al., 2016), extractive summaries only or unaligned abstractive and extractive summaries (e.g., as in (Over et al., 2007; Dang and Owczarzak, 2008)).

In this work, we present an iterative annotation process for producing aligned summaries annotated with text-to-text generation techniques. Figure 1 shows a human-written abstractive summary and human-selected extractive summary from our corpus. The extractive summary suggests the narrator was already uneasy and leaves the reader wondering why. This information is unimportant, but the extractive summary must include it because it is in the same sentence as the bloody woman, just as it must include an extra character: the man in medical attire. Text-to-text generation techniques, such as sentence compression, could be used to rewrite this extractive summary to more closely match the abstractive summary.

<p><b>Abstractive:</b> While driving home I saw a woman covered in blood standing by the side of the road. As I passed she attempted to launch herself at my car.</p> <p><b>Extractive:</b> As I'm looking around as to what the fuck is going on, we approach the roundabout and there is a man in medical attire next to a woman in white pyjamas, with blood covering her clothing. I go straight, and as we go past the woman attempts to launch herself at my car.</p>
---

Figure 1: Abstractive and extractive summaries.

While the extractive summary contains some extraneous information, it does include every-

thing present in the abstractive summary. We use crowd-sourcing with Amazon Mechanical Turk (AMT) to produce our extractive summaries, and workers are given the abstractive summaries as a prompt, ensuring high-quality extractive summaries despite using inexpensive crowd-sourcing. In the next stage of our annotation process, we use AMT workers (Turkers) to align phrases from the extractive summaries to the abstractive summaries. Finally, we use Turkers to annotate the aligned phrases with the five rewriting operations identified by Jing and McKeown (1999) – reduction (compression), combination (fusion), syntactic transformation, lexical paraphrasing, generalization/specification – indicating how best to rewrite each extracted phrase. We make our corpus available on the Web<sup>1</sup>.

## 2 Related Work

Text-to-text generation for abstractive summarization is the task of revising extracted sentences using techniques such as sentence compression (Knight and Marcu, 2000; Lin, 2003; Zajic et al., 2007; Liu and Liu, 2009) and fusion (Barzilay and McKeown, 2005). Unfortunately, corpora for text-to-text generation are rare and time-consuming to produce. Marcu (1999) created a corpus of nearly 7,000 abstractive and extractive summaries of news articles by automatically extracting sentences based on a human-written summary, building a large corpus at the cost of some noise. Murray et al (2005) created a corpus of 61 paired, human-written abstractive/extractive summaries of meeting transcripts, but the gain in summary quality achieved using human annotators is offset by the small size of the corpus.

This work uses personal narratives, widely found on social networks, weblogs, and online forums. The availability of online narrative begins to address a problem facing the text-to-text generation approach to summarization: lack of data. Gordon and Swanson (2009) trained a classifier to identify narratives in blog posts with 75% precision and built a corpus of 937,994 narratives. Ouyang and McKeown (2015) created a corpus of 4,647 narratives collected automatically from Reddit, achieving 94% precision in collecting only narrative text. They argue that the Most Reportable Event (MRE) is the most salient event

and thus the shortest possible summary; they annotated a subset of 476 narratives by extracting sentences that referred to MREs.

The Murray et al corpus includes alignments between phrases in the extractive summaries and sentences in the abstractive summaries. However, none of the corpora described above provides an analysis of how a summarizer might transform an extracted phrase into its abstractive form. While corpora exist for some rewrites in McKeown and Jing (1999), such as compression (Ziff-Davis, Filippova and Altun (2013), Kajiwara and Komachi (2016)), fusion (McKeown et al (2010)), and lexical paraphrasing/syntactic reordering (Ganitkevitch et al (2013)), these corpora exist in isolation. A human summarizer may apply multiple rewrites to a single phrase, and our work captures this information with annotations for all of the rewrites over each alignment.

## 3 Data Collection

We use the annotated subset of 476 personal narratives in Ouyang and McKeown (2015), although we do not use their annotations.

### 3.1 Stage One: Abstractive Summaries

We partitioned the 476 stories into 7 slices of 68 narratives. The narratives were written for 19 different prompts, which roughly correspond to topics (eg. “Your best ‘Accidentally Racist’ story?”). We randomly assigned an equal number of narratives from each prompt to each of the seven slices.

We trained four graduate student annotators from our university’s Department of English and Comparative Literature. Each was assigned four slices: one in common with each other annotator, and one among all annotators. Each participated in a 30-minute training session: they were told to imagine they were about to tell a story to a friend and wanted to ask, “Did I tell you about...?” They should write one or two sentences to complete the question and include any context they thought necessary for their friend to understand it.

We evaluated interannotator agreement on this task using an AMT HIT (Human Intelligence Task) where Turkers were shown summaries written by two different annotators, but not the narrative itself. We then asked the Turkers to decide whether or not the summaries described the same event, and if so, whether one or both of the summaries contained important information not found

<sup>1</sup>[www.cs.columbia.edu/~ouyangj/aligned-summarization-data](http://www.cs.columbia.edu/~ouyangj/aligned-summarization-data)

in the other. We required Turkers to complete a qualification test before working on the HIT, ensuring they had read and understood the task instructions. The test consisted of pairs of example summaries constructed so that the correct answers to our two questions were clear: the paired summaries were identical except for pieces of extra information that we inserted into one or both summaries. Three Turkers worked on each hit, and we considered a pair of summaries to be in agreement if at least two out of three Turkers indicated that the summaries described the same event.

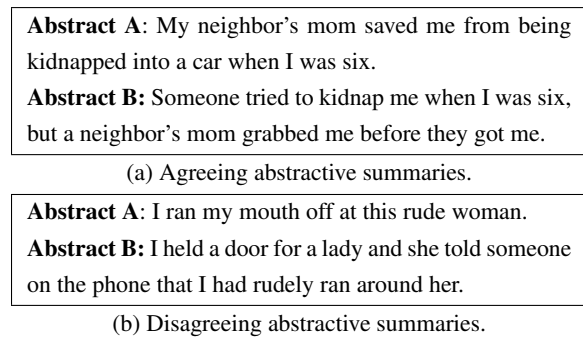


Figure 2: Examples of agreeing and disagreeing abstractive summaries for two different narratives.

Our annotators achieved 90.38% observed agreement, producing a total of 1088 different abstractive summaries. Figure 2 shows a pair of agreeing and a pair of disagreeing abstractive summaries. With the disagreeing summaries, we see that annotators A and B focused on different aspects of the narrative: A summarized the narrator’s confrontation with the rude woman, while B explains why the narrator was angry with the woman. Figure 3 shows a pair of agreeing summaries where Turkers indicated that both summaries contained important information not found in the other summary. We see that annotator A focused on the event’s emotional effect on the narrator, while annotator C emphasized the irresponsible friend’s bad behavior.

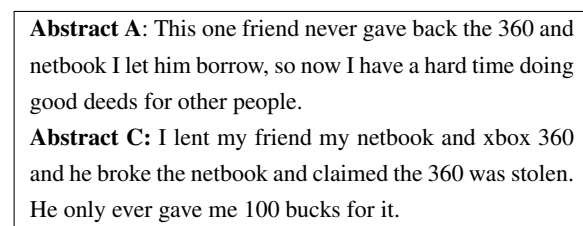


Figure 3: Extra information in a agreeing summaries.

### 3.2 Stage Two: Extractive Summaries

To produce the corresponding extractive summaries, we created another HIT that showed Turkers a narrative, one of its abstractive summaries, and instructions to compose an equivalent summary by selecting as few sentences as possible from the narrative. We once again required Turkers to complete a qualification test before working on our HITs. The test consisted of a single story and abstractive summary, written so that the summary was a word-for-word paraphrase of a single sentence in the narrative that did not overlap with any other sentences. We also required that Turkers be at least 18 years old and have completed at least 10,000 HITs with 98% acceptance on previous HITs. Three Turkers worked on each of our HITs, and Turkers achieved substantial agreement on which sentences they selected: Fleiss’s  $\kappa$  of 0.748. Figure 4 shows an extractive summary where they achieved perfect agreement.

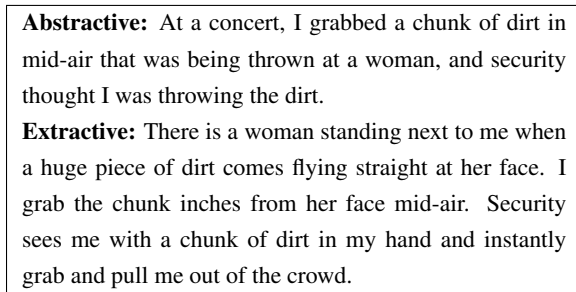


Figure 4: Perfect agreement among Turkers in constructing the extractive summary.

Combining our abstractive and extractive summaries, we have 476 narratives, 408 with two abstractive summaries and 68 with four. For each abstractive summary, we have six extractive summaries, one for each Turker and an additional three created by aggregating the Turkers’ summaries: sentences selected by at least one (*union*), two (*majority*), and all three (*intersect*) Turkers.

### 3.3 Stage Three: Phrase Alignments

We used another AMT HIT to produce phrase alignments between the extractive and abstractive summaries. We showed Turkers one of the abstractive summaries produced in Stage One and its corresponding extractive summary produced in Stage Two (using *union* aggregation). The task was to align phrases between the summaries, and to submit as many alignments as they could find.

To avoid confusing terminology, the instructions referred to the abstractive summary as the

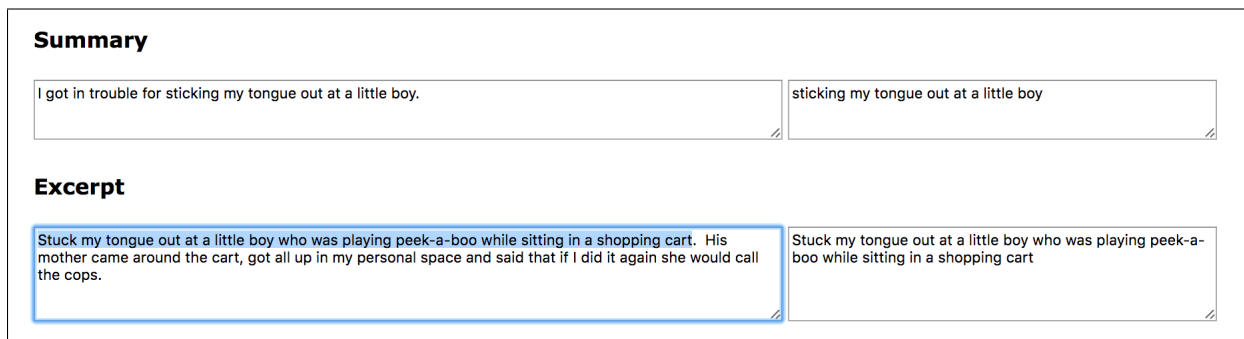


Figure 5: Highlighting interface for Phrase Alignment HIT.

“summary” and the extractive summary as the “excerpt.” We defined aligning as “matching phrases from the summary with phrases from the excerpt that effectively mean the same things.” The HIT interface (Figure 5), allowed Turkers to select phrases by highlighting, save alignments as they went along, and submit all their saved alignments at the end. Three Turkers worked on each HIT.

As in the previous stages, we required Turkers to complete a qualification test where we showed Turkers one phrase from an abstractive summary and four phrases from the corresponding extractive summary and asked them to decide which of the four extractive options would make a good alignment with the abstractive phrase. We also presented them with a link to a demo of the interface, so that they could try the highlighting and saving functions before working on the actual HIT.

### 3.4 Stage Four: Rewriting Operations

Our final HIT asked Turkers to review the alignments produced in Stage Three, and to identify the rewrite operation(s) involved in transforming the extractive phrase into the abstractive phrase. When performing the task, Turkers were only concerned with one rewrite at a time, and simply had to select whether the presented alignment employed that rewrite or not. We designed our task in this way because an alignment could employ more than one rewrite, and we wanted the Turkers to consider each rewrite independently.

We defined the rewrite operations for the Turkers as follows, and provided examples of each.

- **Reduction** keeps key parts word-for-word and removes less important information.
- **Lexical paraphrasing** replaces words or word sequences with paraphrases, ie. other words that have the same meaning.
- **Syntactic reordering** changes the grammatical structure (eg. passive vs active).

- **Generalization** replaces longer strings of detail with shorter, more general descriptions.
- **Specification** replaces short, general descriptions with longer strings of detail.

As in Stages Two and Three, we tested the Turkers’ understanding of the task before allowing them to work on the HITs. Since we ask about one rewrite operation at a time, we designed separate qualification tests for each rewrite. For each test, we selected one abstractive/extractive summary pair and constructed two different alignment examples where one alignment employed the rewrite in question and the other did not. The Turkers were asked whether or not the rewrite was used in each of the two alignments.

Rewrite Operation Counts			
Reduction	216	Generalization	3359
Lexical Para.	1218	Specification	1250
Syntactic Reor.	916		

Table 1: Rewrite operation counts.

For each alignment, we put up four HITs (we combined generalization and specification so that Turkers could choose one or neither, but not both). Table 1 lists each rewrite and how many alignments used it; we include an alignment when at least 2 out of 3 Turkers agreed it used the rewrite. We found that generalization was by far the most popular rewrite operation, and reduction was the least, likely because reduction’s definition was the most demanding, as it required word-for-word matching outside of the removed parts. Figure 6 shows an example each of generalization and its counterpart specification from our annotations.

Generalization: Very rarely do I ever get a “thanks” or a smile of appreciation. → I never get any thanks.  
Specification: I had the alien abduction dream. → I had a sleep paralysis dream where I was abducted by aliens.

Figure 6: Examples of the two most common rewrite operations, generalization and specification.

	Fusion	Reduction	Lexical Para.	Syntactic Reor.	Generalization	Specification
Fusion	<b>1052</b>	36	214	151	695	165
Reduction		<b>185</b>	34	32	113	24
Lexical Para.			<b>1068</b>	179	564	237
Syntactic Reor.				<b>772</b>	391	165
Generalization					<b>2802</b>	0
Specification						<b>1093</b>

Table 2: Rewrite co-occurrences produced from confident and precise alignments.

### 3.5 Discussion

We evaluated our Stage Three and Four data from the Turkers by assigning confidence levels to the alignments and judging annotator agreement on the rewrite labels. It would be difficult to determine interannotator agreement in Stage Three because Turkers could submit any number of alignments of any size for each HIT. Instead, we evaluated on the level of individual alignments. A *confident* alignment had to agree with another alignment, where two alignments agreed if (1) different Turkers submitted them; (2) the selected abstractive phrases overlapped enough that at least half of the shorter phrase was covered by the overlap; and (3) the selected extractive phrases overlapped enough that at least half of the shorter was covered. A *precise* alignment does not contain an extractive phrase that was over two sentences long, because the longer the alignment, the more difficult to identify the rewrite components involved. Thus a *confident* alignment is one where at least two different Turkers aligned the same spans, within a margin of error of a few words, while a *precise* alignment is one where it is easier to pinpoint the spans where rewrite operations apply.

Out of the 6173 alignments the Turkers produced, 5836 (95%) were *confident*, 5602 (91%) were *precise*, and 5281 (86%) were both. When we evaluated the rewrite labels produced for these confident and precise alignments, we found that many were labeled for multiple rewrite techniques at once, indicating that quality phrase transformations often involved stitching together rewrites instead of performing them separately. Figure 7 below displays an example of such an alignment, which was labeled for lexical paraphrasing (3/3 Turker agreement), syntactic reordering (2/3 agreement), and generalization (2/3 agreement).

Table 2 further displays the interactions between rewrites in the form of a co-occurrence matrix of the five rewrites we tested on AMT, plus fusion, which we identified automatically.

Extractive: **My SO at the time had been depressed/suicidal and I had been making posts in relevant subs with a different account asking for advice.** I didn't really have any experience with depression/suicide at the time, so it was a very scary situation for me . . .

Abstractive: My friend identified some of my Reddit posts **about my suicidal SO at the time**, and I was kind of relieved that I ended up getting to confide in him about the situation.

Figure 7: A confident and precise alignment (in bold) with multiple rewrite labels: lexical paraphrasing, syntactic reordering, and generalization. The extractive summary shown is truncated due to length.

## 4 Conclusion

We have presented a new corpus of 1088 aligned abstractive and extractive summaries, totaling 6173 phrase-level alignments, each annotated with rewrite operations, which we make available on the Web. Our iterative annotation process uses trained annotators to generate abstractive summaries and Amazon Mechanical Turk to produce extractive summaries, phrase alignments, and rewrite annotations. We found substantial agreement among annotators and Turkers for all tasks, demonstrating our ability to elicit high-quality summaries and alignments despite using inexpensive crowd-sourcing.

Our corpus provides summaries of a very different type of text from the traditional newswire articles: personal narratives, a genre that natural language processing research is just beginning to explore. This data is widely found on the Web and brings challenges such as informal language and extreme content. We hope that others will make use of these aligned, personal narrative summaries and their annotated rewrite operations, which we make available on the Web. Our next step will be to exploit this data to create an abstractive summarization system using text-to-text generation. We also hope that the success of our annotation method, using both trained annotators and crowd-sourcing, will encourage other researchers to create similar corpora.

## Acknowledgments

This paper is based upon work supported by the National Science Foundation under Grant No. IIS-1422863. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

- Regina Barzilay and Kathleen McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.
- Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. In *Proceedings of Text Analysis Conference*.
- Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.
- Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Proceedings of the 3rd International Conference on Weblogs and Social Media, Data Challenge Workshop*. Association for the Advancement of Artificial Intelligence.
- Hongyan Jing and Kathleen McKeown. 1999. The decomposition of human-written summary sentences. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 129–136. Association for Computing Machinery.
- Tomoyuki Kajiwaru and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, Osaka, Japan, December. The COLING 2016 Organizing Committee.
- Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization-step one: Sentence compression. In *Proceedings of the Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 703–710. Association for the Advancement of Artificial Intelligence.
- Chin-Yew Lin. 2003. Improving summarization performance by sentence compression — a pilot study. In *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages*, pages 1–8, Sapporo, Japan, July. Association for Computational Linguistics.
- Fei Liu and Yang Liu. 2009. From extractive to abstractive meeting summaries: Can it be done by sentence compression? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 261–264, Suntec, Singapore, August. Association for Computational Linguistics.
- Daniel Marcu. 1999. The automatic construction of large-scale corpora for summarization research. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 137–144. Association for Computing Machinery.
- Kathleen McKeown, Sara Rosenthal, Kapil Thadani, and Coleman Moore. 2010. Time-efficient creation of an accurate sentence fusion corpus. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–320, Los Angeles, California, June. Association for Computational Linguistics.
- Christian M. Meyer, Darina Benikova, Margot Mieskes, and Iryna Gurevych. 2016. MdsWriter: Annotation tool for creating high-quality multidocument summarization corpora. In *Proceedings of ACL-2016 System Demonstrations*, pages 97–102, Berlin, Germany, August. Association for Computational Linguistics.
- Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2005. Evaluating automatic summaries of meeting recordings. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 33–40, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5:103–233.
- Jessica Ouyang and Kathleen McKeown. 2015. Modeling reportable events as turning points in narrative. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2149–2158, Lisbon, Portugal, September. Association for Computational Linguistics.
- Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43(6):1506–1520.
- David Zajic, Bonnie Dorr, Jimmy Lin, and Richard Schwartz. 2007. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing & Management*, 43(6):1549–1570.