Adapting Automatic Summarization to New Sources of Information

Jessica Jin Ouyang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Adapting Automatic Summarization to New Sources of Information

Jessica Jin Ouyang

English-language news articles are no longer necessarily the best source of information. The Web allows information to spread more quickly and travel farther: first-person accounts of breaking news events pop up on social media, and foreign-language news articles are accessible to, if not immediately understandable by, English-speaking users. This thesis focuses on developing automatic summarization techniques for these new sources of information.

We focus on summarizing two specific new sources of information: personal narratives, first-person accounts of exciting or unusual events that are readily found in blog entries and other social media posts, and non-English documents, which must first be translated into English, often introducing translation errors that complicate the summarization process. Personal narratives are a very new area of interest in natural language processing research, and they present two key challenges for summarization. First, unlike many news articles, whose lead sentences serve as summaries of the most important ideas in the articles, personal narratives provide no such shortcuts for determining where important information occurs in within them; second, personal narratives are written informally and colloquially, and unlike news articles, they are rarely edited, so they require heavier editing and rewriting during the summarization process. Non-English documents, whether news or narrative, present yet another source of difficulty on top of any challenges inherent to their genre: they must be translated into English, potentially introducing translation errors and disfluencies

that must be identified and corrected during summarization.

The bulk of this thesis is dedicated to addressing the challenges of summarizing personal narratives found on the Web. We develop a two-stage summarization system for personal narrative that first extracts sentences containing important content and then rewrites those sentences into summary-appropriate forms. Our content extraction system is inspired by contextualist narrative theory, using changes in writing style throughout a narrative to detect sentences containing important information; it outperforms both graph-based and neural network approaches to sentence extraction for this genre. Our paraphrasing system rewrites the extracted sentences into shorter, standalone summary sentences, learning to mimic the paraphrasing choices of human summarizers more closely than can traditional lexicon- or translation-based paraphrasing approaches.

We conclude with a chapter dedicated to summarizing non-English documents written in low-resource languages – documents that would otherwise be unreadable for English-speaking users. We develop a cross-lingual summarization system that performs even heavier editing and rewriting than does our personal narrative paraphrasing system; we create and train on large amounts of synthetic errorful translations of foreign-language documents. Our approach produces fluent English summaries from disfluent translations of non-English documents, and it generalizes across languages.

Contents

List of Figures

List of Tables

Acknowledgements

To my advisor, Kathy McKeown, this thesis could not have been possible without you. You believed in me even when I was a naive undergrad who barely knew anything, applying to the PhD program with more opinions than sense. Since then, you have supported all of my wild ideas, even when they did not work out, or reviewers criticized them, or they had nothing to do with whatever I was already working on. Thank you for your guidance, your encouragement, and your confidence in me. I can only hope to someday be as fine an advisor to my own students as you have been to me, but I will try.

To the rest of my thesis committee – Julia Hirschberg, Smara Muresan, Ani Nenkova, and Sasha Rush – thank you for your time and advice.

Julia, thank you for always having a ready answer to all of my late-night emailed questions and for taking the time to serve on all three of my committees, candidacy, proposal, and defense! The initial ideas that went on to become this thesis started in your IGERT Fellowship seminar; thank you for giving me the opportunity and the freedom to explore them.

Smara, thank you for supporting and encouraging me in all of my goals, both here at Columbia and beyond. I remember hearing about your work on sarcasm when I was in my first year here and being excited by your use of distant labeling and genre-specific features. I went on to use those ideas over and over in the work that went on to become this thesis; thank you for inspiring me.

Ani, of all my committee members, I have known you the longest! Thank you for

always finding the time to hear about my work and tell me about yours and for all the pointers you have given me over the years. Large swaths of this thesis use your prior work; thank you for paving the way for me.

Sasha, I remember being scared of you when you were still a student here at Columbia because you would ask questions at talks that made me think, this guy is brilliant and I sure am glad I am not the one who has to answer him! Then my proposal happened, and it turned out not to be as scary as I had imagined. Thank you for your help with the neural side of things, and for leading the neural abstractive summarization revolution!

To Mark Dredze, who helped me to start thinking about life after this thesis, thank you for encouraging my interests, understanding my goals, and putting so much of your time and energy into advising and supporting me.

To those who got me here to Columbia – Bill Labov, Julie Anne Legate, Jim Mayfield, Ben Taskar, Ralph Weischedel, and especially Mitch Marcus: thank you for teaching me, mentoring me, and believing in me.

To those who helped me here at Columbia – Paul Blaer, Adam Cannon, Shih-Fu Chang, Mike Collins, Noemie Elhadad, Becky Passonneau, Drago Radev, and Andrew Rosenberg: thank you for teaching me, advising me, and inspiring me.

To my friends, fellow students, mentors, and mentees, who have helped contribute to the work in this thesis – Or Biran, Serina Chang, Chris Hidey, Alyssa Hwang, Chris Kedzie, Fei-Tzin Lee, Sara Rosenthal, Boya Song, Karl Stratos, Kapil Thadani, and Elsbeth Turcan: thank you.

To my friends and fellow students at Columbia, who have filled my seven years here with happy memories – Apoorv Agarwal, Emily Allaway, Maryam Aminian, Daniel Bauer,

Nishi Cestero, Oscar Chang, Tom Effland, Joe Ellis, Noura Farra, Katy Gero, Melody Ju, Kai-Zhan Lee, Sarah Ita Levitan, Weiyun Ma, Rimma Perotte, Yves Petinot, Vinod Prabhakaran, Anna Prokofieva, Mohammad Rasooli, Kevin Shi, Rose Sloan, Victor Soto, Lakshmi Vikraman, Yan Virin, Laura Wilson, Olivia Winn, and Brenda Yang: thank you.

And most of all, to my parents, thank you for everything you have done for me. I am who I am today because of you; I made it this far because of your love and support. You have always told me that I could do anything, as long as it made me happy. Thank you for always being there for me.

For my grandparents.

Introduction

In the first paper on automatic summarization, Luhn (1958) wrote that the goal of summarization is "to save a prospective reader time and effort in finding useful information." Since then, the field of automatic summarization has grown up around this goal. The scope of summarization research has broadened: where there earliest work focused on summarizing scientific articles, summarization research has grown to cover a wide range of domains, including meeting transcripts, email threads, and most importantly, news articles. While scientific article summaries are primarily of interest to scientists, and meeting and email summaries are primarily of interest to corporate workers, news summaries provide useful information for everyone.

However, English-language news articles are no longer necessarily the best source of information. The Web allows information to spread more quickly and travel farther: first-person accounts of breaking news events pop up on social media, and foreign-language news articles are accessible to, if not immediately understandable by, English-speaking users. More than forty percent of Americans get at least some of their news from social media, and over one-third of Americans between the ages of 18 and 29 consider social media to be the most helpful source of news, ahead of news articles or cable broadcasts (Pew Research Center 2016).

This thesis focuses on summarizing these new sources of information. Techniques developed for summarizing copy-edited, English-language news articles are often inappropriate for *informal genres*, such as blog entries and social media posts. In particular,

many English news articles can be summarized effectively by lightly paraphrasing, or even directly copying, the lead sentence or paragraph, but there is no lead in informal genres. Similarly, while foreign-language news articles might have lead sentences, an English summary of a non-English article (*cross-lingual summarization*) cannot directly copy them – the article must first be translated into English, introducing a potential source of noise via machine translation.

The copying, or *extractive*, summarization approach selects sentences from a document without modifying them. In the case of informal genres, the colloquialisms and grammatical or spelling errors common to such documents require heavy editing or paraphrasing if we are to form a coherent summary from them; similarly, in the case of cross-lingual summarization, we must correct or elide errors and disfluencies introduced by the machine translation system. These documents require an *abstractive* summarization approach that combines edited sentence fragments with new text generated from scratch.

In this thesis, we develop summarization techniques for two new sources of information: the bulk of the thesis focuses on a two-stage, extract-then-abstract approach for summarizing the genre of *personal narrative*, first-person accounts of exciting or unusual events, which are readily found in blog entries and other social media posts; the rest discusses an abstractive approach to cross-lingual summarization, focusing on producing English summaries of documents written in *low-resource* languages, where there are no summarization systems of any kind, and only the most basic machine translations are available.

Contributions

The main contributions of this thesis are as follows:

A human-quality summarization corpus for the genre of personal narrative that supports a two-stage, extract-then-abstract approach to summarization.

While there are many summarization and text-to-text generation corpora available, there are none for personal narrative, which is a relatively new genre of interest in natural language processing. Further, most existing corpora provide only a very specific set of annotations: only extractive summaries or only abstractive summaries, or sentences demonstrating only compression or only fusion and no other text-to-text generation techniques. This does not accurately capture how humans write summaries; human summarizers extract snippets of text, both with and without rewriting them, as well as generating new text, and they use a mix of text-to-text generation techniques. The work described in this thesis provides a corpus that captures all of the diverse operations human summarizers perform.

Our specific contributions in this area are

- a collection of 476 annotated and 4,396 unannotated personal narratives scraped from the Web;
- matching human-written abstractive summaries and human-selected extractive-summaries for each annotated narrative, supporting the extract-then-abstract summarization approach;

- phrase-level alignments annotated with text-to-text generation techniques between corresponding extractive and abstractive summaries, capturing the interaction between techniques.

A narratology-inspired approach to main event detection and extractive summarization for personal narrative.

Personal narratives are structured differently from the news articles that have been the focus of the majority of summarization research, and so new techniques are needed to identify important information in them. They do not have lead sentences, so position-based sentence extraction will not work; they are tightly written, with little to no repetition of information, so salience-based sentence ranking will not work. The work described in this thesis addresses the challenges of extracting summary sentences from personal narratives by drawing from contextualist narrative theory.

Our specific contributions in this area are

- a computational validation of the linguistic theory that the author of a narrative uses changes writing style to mark important content in the narrative;

- a change-based and completely content-agnostic view of narrative summarization that represents each sentence by a set of affectual and stylistic features;

- the use of this change-based view of narrative to perform both event detection and extractive summarization for personal narrative.

A thorough investigation of the task of paraphrasing for abstractive summarization.

While both paraphrase generation and abstractive summarization are active areas of research, the two have not previously been considered together. Further, prior work on paraphrase generation was primarily concerned with producing as many lexically diverse paraphrases as possible, while selecting the appropriate paraphrase and determining whether or not a given sentence ought to be paraphrased in the first place were neglected. The work described in this thesis focuses on the latter two tasks: determining which input document sentences a human summarizer would paraphrase and selecting natural, human-like paraphrases.

Our specific contributions in this area are

- a demonstration of the limitations of previous, "shotgun" approaches to paraphrasing when it comes to generating human-like paraphrases;
- a sentence-level neural classifier for predicting whether or not a document sentence ought to be paraphrased for summarization;
- a sentence-level neural paraphrasing system for rewriting a document sentence into an abstractive summary sentence;
- an analysis of word-level features that allow the above two systems to learn human-like paraphrasing.

A monolingual paraphrase alignment system capable of aligning phrases of any length and robust to minor semantic differences.

The task of monolingual paraphrase alignment branched out from the related task of alignment for machine translation and, as a result, inherited two of the latter's key assumptions: first, that alignments are between words or phrases that mean exactly the same thing, and second, that alignments are between single words or short phrases. However, these assumptions break down in monolingual alignment when we consider source/target sentence pairs that come not from parallel corpora, but from document/summary pairs. Prior work on monolingual alignment has been limited to aligning either identical words or very close synonyms, and alignments between longer phrases are beyond the capabilities of state-of-the-art systems; they are thus unable to align sentence pairs where the sentences are lexically diverse. The work described in this thesis addresses the two issues of aligning long phrases and aligning despite minor semantic differences between those phrases.

Our specific contributions in this area are

- a constituent-based view of alignment that is able to align phrases of arbitrary length as though they were single words;
- the use of a phrase-level embedding system to allow each phrase to be aligned based on its general meaning, rather than on the words it contains;
- a pointer-network-based alignment system that aligns a source phrase to the entire target sentence at once;

- a voting procedure for aggregating alignments using phrases of different lengths, which automatically discovers the optimal phrase length for a source/target sentence pair.

A fluent and generalizable approach to cross-lingual summarization for low-resource languages.

Prior work on cross-lingual summarization has focused on the high-resource language pair of English and Chinese. Because both the source language (English) and target language (Chinese) are very active in natural language processing research, summarization techniques can be applied to both the source language document and the translated target language document, and the translations are of good quality. With a low-resource source language, however, there is no data to train a source language summarization system, and translations into the target language are of much lower quality. The work described in this thesis is the first to address the low-resource language scenario in cross-lingual summarization.

Our specific contributions in this area are

- a robust, neural abstractive summarization system that produces fluent English summaries from low-quality, machine-translated documents from three low-resource languages: Somali, Swahili, and Tagalog;

- the trained, three-language model that works "out of the box" for new, previously unseen source languages, even without additional training;

- a straightforward methodology for extending the system to new languages to further tailor performance to a specific source language.

Outline

Five of the first six chapters of this thesis detail a two-stage, extract-then-abstract summarization system for personal narrative; Chapter 5 takes a brief detour into monolingual paraphrase alignment, and Chapter 7 concludes with cross-lingual summarization. Each chapter begins with an overview of previous work, either on the task discussed in the chapter or on a closely related task, and concludes with a summary of contributions and a discussion of how the chapter ties in to the overall structure of this thesis.

- **Chapter 1** introduces the genre of personal narrative, gives an overview of both linguistic and computational work on the genre, and motivates both our goal of summarizing personal narrative and our two-stage approach to doing it.

- **Chapter 2** details the change-based view of personal narrative, as well as the linguistic theory that inspired it, and then presents our early experiments on using this change-based view to detect the *most reportable event* of a narrative.

- **Chapter 3** describes the creation of our personal narrative summarization corpus, focusing on how we elicited matching pairs of human-quality extractive and abstractive summaries and how we captured multiple text-to-text generation techniques.

- **Chapter 4** applies the change-based approach of Chapter 2 to the task of extractive summarization, using the corpus in Chapter 3.

- **Chapter 5** presents the monolingual paraphrase alignment system that we use later in Chapter 6; it motivates the need for a system capable of aligning long phrases that do not necessarily mean exactly the same thing and details our solution for this task.

- **Chapter 6** discusses the task of paraphrasing for summarization, explores the limitations of previous approaches to paraphrase generation, and presents our approaches to addressing those limitations.

- **Chapter 7** shifts our focus to the task of cross-lingual summarization for low-resource languages, motivating the task and describing our training pipeline for producing summarization systems for any source language.

- **The Conclusion** summarizes our contributions, discusses the limitations of the work in this thesis, and suggests future directions for adapting automatic summarization to new sources of information.

1. Motivation

In the following chapters, we focus on the genre of *personal narrative*: first-person stories about exciting or unusual events in the authors' lives. We argue that personal narrative is an important and previously neglected genre for summarization research – when the archaeologists of the future want to know what it was like to live in the early twenty-first century, it is to personal narratives, not news, that they will turn.

1.1 A Brief History of (Computational) Narratology

The study of narrative (in general, as opposed to *personal* narrative in particular) spans across such diverse disciplines as philosophy (Ricoeur 1984), psychology (Goldman, Graesser, and Broek 1999), anthropology (Ochs and Capps 2001), and of course, linguistics and literature. In this section, we give a brief (noncomprehensive) overview of the field of narratology and its parallels in natural language processing research.

The formal study of narrative began with the work of Propp (1928), who noted that there is a limited set of actions or functions that form the building blocks of every fairy tale, and further that the order of these functions is always the same. While the names and other attributes of the characters vary across fairy tales, the sequence of actions (the hero leaves home, the hero is tested in some way, the hero acquires a magical item or follower, etc.) does not. Fittingly, the study of computational narrative began with the work of Lehnert (1981), who likewise identified fifteen primitive *plot units* constructed

from different combinations of positive or negative events and character mental states, along with the character motivations or actions that relate them to one another, and later work by Goyal, Riloff, and Daumé III (2010) tackled automatically identifying the plot units that make up a particular narrative. Lehnert's work was also the earliest attempt at summarizing narrative; she argued that a narrative is a graph of plot units, and that its summary can be constructed by selecting the most highly connected plot units.

Todorov (1969) argued that all narratives share an underlying structure, a *minimal plot*, where there is a shift from one equilibrium to another (Prince (1973) similarly defines a minimal story as an event that causes a change of state). Barthes (1975) defines three levels of structural analysis for narrative: (Proppian) functions; characters, their perspectives, actions, and relationships; and narration, including both the identity of the narrator and of the audience. These three levels of analyses are reflected in the computational narrative work of Elson and McKeown (2010) and Elson (2012), who encoded narratives in three graphs: the textual graph, containing spans from the narration; the timeline graph, containing predicate functions; and the interpretive graph, containing character goals, beliefs, and affect.

These works represent the *structuralist* view of narratology – that each narrative has two parts: a deep structure, consisting of characters and functions, and a surface form, consisting of the narration. The structuralist claim is that every narrative is the realization of some deep structure, and so identifying a narrative's deep structure is of paramount importance. This claim is reflected in the computational narrative work of Chambers and Jurafsky (2008), who extracted *narrative event chains* of functions centered around a common character, and of Manshadi, Swanson, and Gordon (2008), who learned probabilistic

sequences of predicate-argument pairs extracted from narratives.

The opposing *contextualist* view of narratology focuses not on the deep structure of a narrative, but on its surface form. Smith (1980) defines a narrative as a social act, a function of who the narrator is, who the audience is, the narrator's purpose in telling the story, and the audience's purpose in listening to the story. If any of those properties changes, the narrative produced will also change. Smith further argues that there is no such thing as a deep structure that exists independently of narration; stories cannot exist independently of their purpose in being told.

Polanyi (1981) notes that since storytelling is a social act, a narrative must be interesting and have a point, otherwise it will not hold the audience's attention. Narrators use *evaluation devices*, such as onomatopoeia, reported speech, and changes in sentence length, complexity, or verb tense, to emphasize important parts of their stories; these evaluated elements of a narrative form an *adequate paraphrase* of the point of that narrative.

1.2 Personal Narrative

The study of personal narrative is closely intertwined with contextualist narratology.

Labov (1966) founded the study of personal narrative partly by accident. A sociolinguist, Labov was trying to record variations in language across social classes in New York City. He found that, when confronted with an Ivy League professor and a microphone, people changed the way they spoke: "People said what they thought you wanted to hear, and said it in a way that they thought you wanted it to be said" (Labov 2013). Trying to combat this observer's paradox, Labov discovered that, when talking about important

personal experiences, people spoke more naturally. As a result, he collected hundreds of personal narratives and found a shared structure among them.

Here the terminology gets a bit confusing. Labov's *narrative structure* is not the deep structure of structuralist narratology; the Labovian view is a contextualist one. Labov defined the elements of narrative structure in terms of their role in organizing the narration, such as giving necessary background information that the audience needs to understand the narrative, or signaling the end of a narrative by relating it to the present. Labov's narrative structure is found in personal narratives across cultures (Labov 2010), suggesting that there is a universal, "right" way to effectively share a personal experience – a deep structure of narration, in a sense.

Labov and Waletzky (1967) defined the structure of narrative as consisting of three elements: the *orientation*, the *complicating action*, and the *evaluation*. Labov (1997) refined this structure to include three additional elements – the *abstract*, the *resolution*, the *coda* – and argued that every narrative is organized around a single *most reportable event*. Each clause of a narrative is assigned to one of the elements of narrative structure, but not all elements are necessary in every narrative – the original three are sufficient for a narrative, and the later three (abstract, resolution, and coda) are relatively rare.

Abstract. The abstract is an introduction to the story and may contain a description of the most reportable event. For example –

> Shall I tell you about the first man got kilt – killed by a car here... Well, I can tell you that.

(All examples in this section are from Labov (2013)).

Orientation. The orientation contains information on the time, place, and people involved in the story – the background information. It usually occurs at the beginning of the narrative, but some orienting information may be postponed until later in the narrative, just before it becomes relevant. An example of this is found in "Jacob Schissel's story": the orienting information

> When I let go his arm, there was a knife on the table,

is not given until close to the end of the narrative, just before the Schissel is stabbed with the knife.

Complicating Action. The complicating action is a chain of causal or instrumental events that culminates in the *most reportable event*. The complicating action chain tells what happened in the story. In "Jacob Schissel's story", the chain of complicating actions is as follows:

1. He saw a rat out in the yard
2. and he started talk about it [sic]
3. and I told him to cut it out.
4. ...I grabbed his arm
5. and twisted it up behind him.
6. he picked up [the knife]
7. and he let me have it.

Each event is causally related to the one before it, except for events 5 and 7, which are instrumentally related to events 4 and 6.

Evaluation. The evaluation is where the narrator gives his opinions on the events of the story, considers alternative outcomes, assigns praise or blame to the characters, or attempts to add credibility to the story. As suggested by its name, the evaluation serves to establish

the "point" of the narrative; it usually comes in a block immediately following the *most reportable event*, but like orientations, evaluations can be interjected among the events of the complicating action. For example, Jacob Schissel gave this evaluation on being stabbed:

> And the doctor just says, "Just that much more," he says, "and you'd a been dead."

This evaluation serves two purposes: first, it presents an alternative outcome in which the narrator did not survive the stabbing; second, it adds credibility to the stabbing by quoting a third party witness, the doctor.

Resolution. Some narratives extend the chain of events to a final resolution of the situation created by the most reportable event. For example, in a narrative about a fight, the narrator gives the resolution,

> An' they took us – they took us to the hospital.

Coda. The coda signals the end of the story by bringing the listener back to the present. For example, in the story about the fight, the lines

> An' that was it. That's the only fight I can say I ever lost.

relates the events of the narrative to the present.

Computational approaches to personal narrative are relatively few and recent. Rahimtoroghi et al. (2013), Ouyang and McKeown (2014), and Swanson et al. (2014) explored automatically labeling clauses in a narrative with Labov's structural elements. Personal narratives have also been used as data for learning commonsense causal relations (Gordon, Bejan, and Sagae 2011) and subjectivity (Sagae et al. 2013).

1.3 Summarizing Narrative

Aside from the early work of Lehnert (1981), summarization of narrative has been largely ignored in natural language processing research; there is far more interest in generating narratives than in summarizing them (McIntyre and Lapata 2009; McIntyre and Lapata 2010; Riedl and Young 2010; Montfort 2011; Rishes et al. 2013; Li et al. 2013). We argue that personal narrative is just as important a genre for summarization research as is news.

Storytelling is a fundamental and universal form of communication, and as an information source, the genre of personal narrative complements that of news. Where news articles are objective, personal narratives are subjective; where news articles convey facts, personal narratives convey experiences. Linde (1993) and Fivush, Bohanek, and Duke (2005), among others, argue that personal narratives are inextricably tied to one's sense of self. To share one's own experience in a personal narrative, one must evaluate the experience and construct a story that persuades the audience to take one's side; sharing in others' experiences provides new models of how the world works, new frameworks for behavior, and new perspectives on one's own experiences.

Anyone can author a personal narrative. A witness to a breaking news event can break the story him- or herself with a personal narrative far more quickly, for better or worse, than a news agency can compose, edit, and fact-check a news article. With the global reach of social media, anyone who has been affected by a newsworthy event can share his or her experience. Consider the famous statement usually attributed to Joseph Stalin: "The death of one person is a tragedy; the death of one million is a statistic." News may report the statistic, but personal narrative has the power to convey the tragedy of each of those one

million deaths.

Personal narrative summarization faces challenges not found in news summarization. First, content selection is more difficult. There is no headline or lead sentence in a personal narrative. While the abstract may paraphrase the most reportable event of a personal narrative and thus serve a similar role as a lead sentence, abstracts are rare among personal narratives, and not all abstracts explicity describe the most reportable event – one abstract from our personal narrative summarization corpus (described in Chapters 2 and 3) states simply, "This isn't exactly creepy, but it's one of the scariest things that's ever happened to me." Position-based sentence extraction techniques will not work for personal narrative.

Further, personal narratives are, by their nature, already quite minimal. Personal narratives are social acts and need to compete for their audience's attention; an overwritten story will be ignored in favor of any one of the thousands of shorter, punchier stories posted on the Web. There is thus little, if any, repetition of information in a personal narrative, so graph-based sentence extraction techniques will not work (we demonstrate this in Chapter 4).

The second challenge of personal narrative summarization is the need for very aggressive rewriting. Personal narratives posted on the Web do not have professional editors cleaning them up; spelling and grammatical errors are very common. In addition, while a narrative posted on the Web is capable of reaching vast audiences, it is often written for a small, specific audience: the users of the specific Web community in which it is posted. Thus, unexplained acronyms or other jargon may be present and need to be rewritten in order for a general audience to understand the summary. This problem extends to pronouns and other diectics; where a reporter writing a news article will avoid ambiguous

antecedents and remind his or her reader of who important named entities are, a layperson writing a personal narrative may not. This is all in addition to the basic rewriting required to produce any abstractive summary: deletion of extraneous information, fusion of mutually redundant sentences, and so on.

Sharing personal narratives online has had a tremendous impact on our society, and automatic summarization for personal narratives can, too. Topics such as mental health and sexual assault are stigmatized, and so people may be reluctant to discuss or seek help for them. However, people are willing to share personal narratives about their experiences with these topics online – the #PreexistingCondition and #MeToo movements. With these viral personal narrative movements, there are far too many narratives posted for a human to read them all; summarization can help psychologists, public health researchers, and sociologists more easily find those narratives that are of the most interest for their work. In mental health, automatic summarization can also help people to avoid reading certain narratives – people with post-traumatic stress disorder, for example, may have symptoms that are triggered by their being reminded of the source of their trauma. If such a person is browsing narratives online, he or she may not realize that a narrative contains such triggering content until it is too late; a summary can inform him or her that such content is present in the narrative, without going into enough detail to trigger symptoms.

In the following chapters, we describe how we address these two challenges separately. We develop a two-stage summarization system for personal narrative: a content selection system first identifies important sentences in a narrative by extracting *evaluated* sentences, sentences that are stylistically distinct from their neighbors; a paraphrasing system then rewrites the selected sentences into a coherent summary.

2. Main Event Detection

In one of the early linguistic analyses of storytelling, Prince (1973) defined a narrative as describing an event that causes a change of state. Prince's minimal narrative had three parts: the starting state, the ending state, and the event that transforms the starting state into the ending state. Prince gave this example of a minimal narrative:

> A man was unhappy, then he fell in love, then as a result, he was happy.

This is, of course, a toy example, but Prince's claim that a narrative is about a single key event was mirrored in the work of Labov, who defined a well-formed narrative as a series of actions leading to a *most reportable event* (MRE). The MRE is the point of the narrative – the most unusual event that has the greatest emotional impact on the narrator and the audience (Labov and Waletzky 1967; Labov 1997). The MRE is what the narrative is about, and without the MRE, there is no narrative. Thus, the MRE is also the shortest possible summary of the narrative; it is what one would say about the narrative if one could only say a single thing.

Our work on summarizing personal narrative began with the task of identifying the MRE of a narrative. In narratological terms, the MRE is part of the deep structure of a narrative; it is a real-world event that underlies and exists independently of the narration. Such a thing is is difficult to infer automatically, so instead, we focused on first extracting sentences in the narrative that described or referred to the MRE. This extractive approach is in keeping with Labov's own – a contextualist himself, Labov did not distinguish between

the MRE and the sentence in the narration that contains the MRE.

2.1 Related Work

Earlier natural language processing work drawing from Labovian narrative theory focused on classifying clauses in a narrative into Labov's elements of narrative structure. Rahimtoroghi et al. (2013) used 20 of Aesop's fables, totalling 315 clauses. They used two annotators to manually label clauses with the three main elements of narrative structure defined by Labov and Waletzky (1967): *orientation* (background information), *action* (events), and *evaluation* (author's perspective), which we discussed in the previous chapter. Like the work discussed in the rest of this chapter, Rahimtoroghi et al. avoided lexical features and instead represented sentences by their verb tenses, the presence of dialogue or questions, and other such structural features. They achieved accuracy and precision of about 90% on all three labels, as well as recall of about 90% on all but clauses orientation, which their classifier often confused with evaluation clauses. It is interesting to note that Rahimtoroghi et al. considered using two different datasets in their experiments: Aesop's fables and the Say Anything dataset (Swanson and Gordon 2012) of 267 personal narratives collected from weblog entries. They chose Aesop's fables over Say Anything largely because the latter was too challenging; the fables were shorter and used simpler language, and each clause served a clear narrative purpose.

Swanson et al. (2014) also worked on three-way classification of clauses into *orientation*, *action*, and *evaluation*. They used 50 personal narratives taken from the corpus of Gordon and Swanson (2009) (the Say Anything dataset was also taken from this cor-

pus). This corpus was automatically constructed using unigram lexical features to classify weblog entries as either narratives or not, with about 75% precision. Swanson et al. used three annotators to label their subset of 50 personal narratives, totalling 1,602 clauses. Unlike Rahimtoroghi et al.'s data, where over half of the clauses were actions and the rest were mostly evaluations, Swanson et al. found that nearly half of their clauses were evaluations, with orientations and actions making up about one quarter each. Swanson et al. used both lexical and structural features and achieved 69% overall F-score on three-way classification of clauses.

Swanson et al. echoed Rahimtoroghi et al.'s earlier conclusion that personal narratives collected from the Web are more challenging to work with than are Aesop's fables. While Swanson et al.'s annotators achieved substantial interannotator agreement (Fleiss's $\kappa = 0.63$) on the weblog narratives, they could not approach the near perfect agreement (Cohen's $\kappa = 0.82$)) of Rahimtoroghi et al.'s annotators on the fables. Unlike the fables – and indeed, unlike the personal narratives studied by Labov himself, which were transcribed from recordings of sociolinguistic interviews – weblog narratives often have long clauses that serve more than one purpose. Swanson et al. gave this example of a very long clause from their dataset of 50 personal narratives, which contains a mix of orienting information, narrator actions, and author evaluation:

> After leaving the apartment at 6:45 AM, flying 2 hours, taking a cab to Seattle, and then driving seven hours up to Whistler including a border crossing, it's safe to say that I felt pretty much like a dick with legs.

We likewise found this to be the case in the personal narrative data we used in this chapter.

Finally, our own prior work (Ouyang and McKeown 2014) focused on extracting the *action* chain of personal narratives. We used 49 narratives, consisting of 1,277 clauses, from Labov (2013), which were transcribed from sociolingusitic interview recordings – Labov himself annotated the clauses with his elements of narrative structure. Our features were a mix of narrative-inspired features, similar to but simpler than those we use later in this chapter, and discourse features inspired by the work of Pitler, Louis, and Nenkova (2009). We achieved 72 % F-score on the task of extraction action clauses and found that a clause's participation in certain classes of discourse relations was one of the strongest predictors of whether or not it was an action.

2.2 Modeling Personal Narratives

Despite the successes of Rahimtoroghi et al. (2013), Swanson et al. (2014), and Ouyang and McKeown (2014), we chose not to use Labov's elements of narrative structure directly. While Labov considered the *most reportable event* to be the culmination of the sequence of *actions* in a narrative (Labov 1997; Labov 2013), in practice this is not always the case: as discussed in the previous chapter, the sequence of actions in a narrative may be broken up by delayed *orientations* or early *evaluations*, or it may be extended to include a *resolution*.

Instead, we returned to Prince's claim that stories are about change. Polanyi (1985) observed that the climax or turning point of a story is often marked by a change in style, formality of language, or emphasis in the narration; Labov (2013) likewise observed that a change in verb tense often accompanies the MRE of a narrative. Thus, we hypothesized that the MRE sentence would be found at a point of change in the narration.

We modeled the narration using three categories of sentence-level scores: syntactic, semantic, and affectual.

Syntactic. Based on Polanyi's claim that a change in formality marks the changing point, we included metrics of sentential syntax, using the syntactic complexity of a sentence as an approximation for its formality. The complexity of a sentence also reflects emphasis – short, staccato sentences bear more emphasis than long, complicated ones. We used the length of the sentence, the length of its verb phrase, and the ratio of these two lengths; the depth of the sentence's parse tree, the depth of its verb phrase's subtree, and the ratio of these two depths. We also used the average word length for the sentence, as well as the syntactic complexity formula proposed by Botel and Granowsky (1972), which scores sentences based on the presence of specific syntactic structures, such as passives, appositives, and clausal subjects. Finally, we used the formality and complexity lexicons created by Pavlick and Nenkova (2015), which provide human formality judgments for 7,794 words and short phrases and complexity judgments for 5,699 words and phrases (we scored each sentence by averaging across all words and phrases in the sentence).

Semantic. As the MRE is an unsual and surprising event, we expect sentences describing it it to be dissimilar from the surrounding sentences. We used semantic similarity between a sentence and its immediate neighbors as a measure of surprisingness. Our semantic scores were the bag-of-words cosine score and the latent semantic similarity score (Guo and Diab 2012) between a sentence and its preceding and following sentences.

Affectual. Since the goal of a personal narrative is to share an important experience, we expect the MRE to occur at an emotional peak in the narration. We used the Dictionary of Affect in Language (DAL) (Whissell 1989), augmented with WordNet (Miller 1995) for

better word coverage. The DAL represents lexical affect with three scores: evaluation (*ee*, hereafter 'pleasantness' to avoid confusion with Labov's *evaluation*), activation (*aa*, activeness), and imagery (*ii*, concreteness). We also used a fourth score, the activation-evaluation (AE) norm, a measure of subjectivity defined by Agarwal, Biadsy, and McKeown (2009):

$$norm = \frac{\sqrt{ee^2 + aa^2}}{ii}$$

For each of these four word-level scores, we calculated a sentence-level score by averaging across the words in the sentence, accounting for negation by using the finite state machine described by Agarwal et al., which inverts the scores of words that follow a negation word, reverting to uninverted scores after the word "but" is encountered. We expect the sentences surrounding one containing the MRE to be more subjective and emotional as the impact of the MRE becomes clear. We also expect a build-up in activeness and intensity, peaking at the MRE sentence.

For example, the following personal narrative has its MRE sentence in bold:

> This isn't exactly creepy, but it's one of the scariest things that's ever happened to me. I was driving down the motorway with my boyfriend in the passenger seat, and my dad in the seat behind my own. My dad is an epileptic and his fits are extremely sporadic. Sometimes he goes extremely stiff and other times he will try to get out of places or grab and punch people. **Mid-conversation I felt his hands wrap around my throat as I was driving, pulling my head back and making it increasingly difficult to drive.** My boyfriend managed to help steer the car into the hard shoulder but it was one of the scariest experiences in my life.

To extract the MRE sentence, we looked for changes in the syntactic, semantic, and affectual scores of the narration. Figure 2.1 shows the activeness and pleasantness DAL

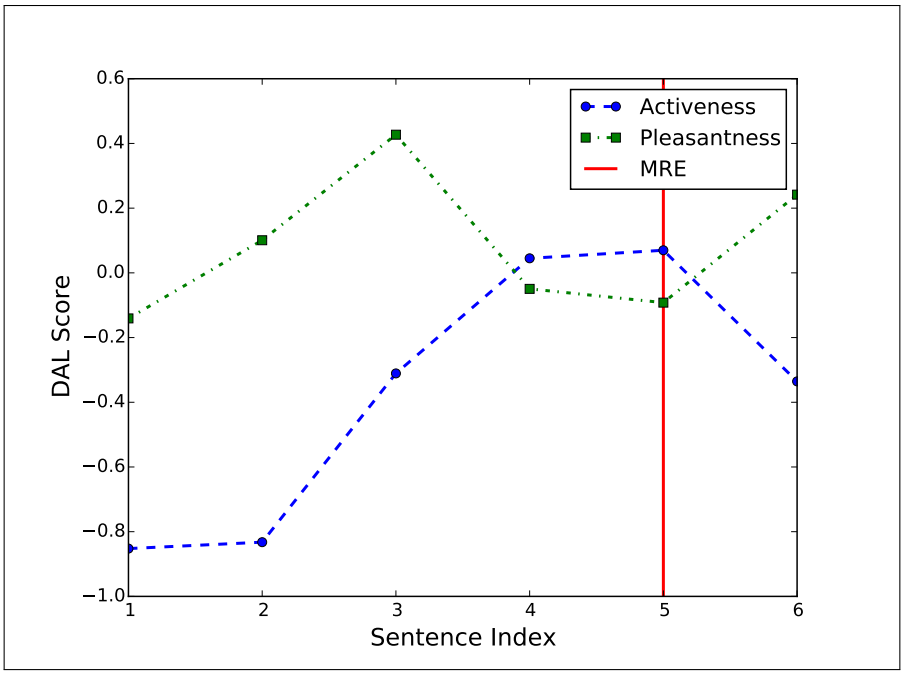

Figure 2.1: Activeness and pleasantness scores over the course of a personal narrative.

scores for the example narrative above. It shows that the MRE is the most exciting sentence in the narrative – global maximum in activation – as well as the most horrifying – global minimum in pleasantness. The overall shape of the activeness scores reflects Prince's three components of a minimal story: low initial activation (starting state) and low final activation (ending state) with a build up to a peak at the MRE (change in state) between them.

2.3 Data

Having developed this change-in-style view of the *most reportable event*, we needed narratives on which to test it. A large collection of personal narratives scraped from weblog posts already existed (Gordon and Swanson 2009), but we chose not to use it for two reasons. First, the Gordon and Swanson corpus consists of weblog posts that contain mostly, but not necessarily only, narrative content. Gordon and Swanson noted the rarity of weblog

posts consisting solely of narrative content and so decided to allow for some non-narrative content in their data, but non-narrative content would only serve to add noise to our change-in-style view of the MRE.

Second, while acquiring a large amount of personal narratives was possible, acquiring large amounts of human MRE annotations was less so. A human annotator must read each narrative, decide what he or she thinks the MRE is, and then identify the sentences that describe or refer to it. This is time-consuming work, and getting enough human MRE annotations to train even a non-neural model would be difficult. We realized that we would likely need to make use of unannotated personal narratives to train our model. Unfortunately, Gordon and Swanson's corpus was created by a trained classifier that predicted whether or not a given weblog post contained mostly narrative content; they reported 75% precision on the task, so 25% of the documents in their corpus were not personal narratives at all. A human annotator would be able to identify those non-narrative documents and remove them from the collection, but as we would need to use unannotated documents as well, we did not want to risk using so many non-narrative documents to train our model.

Thus, we needed to create a personal narrative corpus of our own – ideally, one that would address a third concern: how could we incorporate unannotated narratives in training our model? Since direct supervision would not be possible, distant supervision would be the next best option, so what data source would allow us to develop heuristics to label MRE sentences automatically?

2.3.1 Narrative Collection

We turned to Reddit, a social bulletin board website organized into topic-specific 'subreddit' communities. Each subreddit is a bulletin board where users make *posts*, consisting of a title and (optionally) a body, which can be text, an image or video, or a link to another website. Other users can make *comments* replying to these posts, or comments replying to other comments, forming a tree of comments rooted at the post. Users can also "upvote" (give a point to) and "downvote" (take a point from) any post or comment. Reddit automatically sorts posts in the same subreddit and comments rooted at the same post by by their accumulated points and hides comments with negative points, allowing users to filter out off-topic or low-quality comments.

Figure 2.2 shows an example of a post from the AskReddit subreddit, where users post questions for other members of the comunity, who reply with comments answering the questions. It also shows the advantages of collecting our personal narratives from Reddit:

- We could target narrative content by collecting comments replying to AskReddit posts that asked for stories.
- The points system ensured that most of the collected comments would be high-quality, on-topic narratives.
- We could capture each narrative as a social act: a comment replying to a prompt, with audience comments responding to it.

Using the Python Reddit API Wrapper[1] (PRAW), we scraped the top 50 AskReddit posts containing the keyword "story." Of these posts, 10 were tagged as NSFW ("not safe

[1] https://praw.readthedocs.org/en/v2.1.20/

Figure 2.2: Example of AskReddit post and comments.

for work"), indicating they contained adult content; we did not include those posts in the corpus, as we felt the language would be too different from that used in posts without the NSFW tag. Another 3 posts did not contain personal narratives, and instead were about fictional stories in movies or song lyrics. Table 2.1 shows some examples of AskReddit posts that we scraped, demonstrating the wide variety of story topics included in our corpus.

From the 37 remaining posts, we collected 6,000 *top-level* comments – those whose parent was the post, and not another comment – with positive vote score. We also collected

Post Title
Whats your creepiest (REAL LIFE) story?
Your best “Accidentally Racist” story?
What are your stories of petty revenge?

Table 2.1: Examples of AskReddit posts that serve as prompts for personal narratives.

any sub-comments replying to those top-level comments, discarding any top-level comments with no sub-comments responding to them; we used the number of sub-comments as an estimate for community engagement with, and thus quality of, the top-level comment. These two filters, vote score and presence of sub-comments, allowed us to achieve very high precision in collecting personal narratives, as we discuss in the next section.

We tokenized the top-level comments by sentence and removed all sentences following any variation of the word ‘EDIT’, as these were usually responses to readers’ comments and not part of the narrative. We discarded texts with fewer than three sentences, leaving 4,896 narratives with an average length of 16 sentences and a maxiumum of 198. While our experiments, both in this chapter and in Chapter 4, are conducted on these narratives collected from AskReddit, we expect the change-based model described in the previous section to apply to personal narratives from any source, whether from another Reddit community or another social media website; as discussed in the previous chapter, there is a universal “right” way to tell a story, and we show in this chapter and in Chapter 4 that the change-based model of narrative captures it.

2.3.1.1 Ethical Concerns

We take a brief detour here to discuss the ethics of scraping personal narratives from the Web without the explicit consent of the authors. Our collection and use of scraped personal

narratives described above was ethically sound for three following reasons.

First, we scraped the narratives from AskReddit, a public subreddit. Further, AskReddit is one of Reddit's "default subs," to which every Reddit user is automatically subscribed when he or she creates an account; content from the default subs is also shown to anyone browsing Reddit without being logged into an account. Thus, Reddit users know that content posted to AskReddit can – and very likely will – be seen by anyone.

Second, we did not target any particular user's content; we collected narratives based on their topics, not their authors. Further, we did not collect the usernames of any of our narrative authors, and narratives in our corpus are identified only by numbers reflecting the order in which they were collected, so it is difficult to discover the author's username (although not impossible; a determined snooper could search AskReddit for the text of a narrative).

Finally, we collected the narratives to train a summarization system. Automatic summarization is an innocuous task, the goal of which is to faithfully convey the important content in an input document; a summary does not change or misrepresent the content in its input document.

However, a researcher could easily stumble into ethically unsound territory, such as by scraping data from a private subreddit (or other similar, non-public community). Content on private subreddits cannot be seen by users who are not members of that subreddit; to become a member, a user must be approved by the moderating team of that subreddit. Since scraping Reddit consists of using a bot that is associated with a user account to access posts, it is possible for a researcher to gain access to a private subreddit and use a bot to scrape it. However, unless the researcher had informed that subreddit community of his or her intent,

and had received permission to collect the data, it would be deceptive and highly unethical to do so.

Similarly, targeting a specific user's data, rather than collecting an aggregate of many anonymous users' data, or using a user's data to train for a more targeted task, rather than a general task like summarization, would be invasive and unethical. Tracking a single user's content – for example, to capture changes in his or her writing style over time – or training a system based on that user's content – for example, to use style transfer to rewrite text so that it appears to be authored by him or her – requires the user's consent.

2.3.2 Human MRE Extraction

We partitioned our data into development, seed, and tuning sets of 100 narratives each; a testing set of 200 narratives; and a training set of 4,396 narratives. Narratives from 36 of the 37 AskReddit posts were randomly assigned to each of the development, seed, tuning, and training sets, such that the proportion of narratives from a given post in each of these data sets matched the proportion of narratives from that post in the overall corpus. The test set consisted of 103 narratives from the 37th post, "What's a story you've been wanting to tell on here but no one has submitted a viable question?" and 97 narratives from the other 36 posts, distributed like those in the other data sets. The development, seed, tuning, and testing sets were manually annotated by a native English speaker, a graduate student in our research group, who was instructed to extract all sentences that contained or referred to the *most reportable event*.

To measure interannotator agreement, we also had a second annotator (also a native

English speaker and and graduate student in our research group) extract MRE sentences from the 100 narratives in our development set. We found substantial agreement (Cohen's $\kappa = 0.729$) between the two annotators; the two classes, MRE sentence and not, are highly unbalanced – only 15% of sentences in the human-annotated data sets were MRE sentences – so observed agreement percentage between the two annotators was extremely high (95%).

In addition to labeling the MRE sentences, our first annotator identified and discarded 31 texts out of 500 that were not personal narratives, but rather Reddit-specific inside jokes or meta-discussions on how interesting the stories in the thread were. From this, we can see that the precision of our story collection method is very high. Gordon, Cao, and Swanson (2007) found that narratives comprised only 17% of the weblog posts that they collected, while 94% of our 500 top-level comments that were read by a human were narratives.

2.3.3 Heuristic MRE Extraction

It took several weeks for our human annotator to extract 500 personal narratives' worth of MRE sentences, after which we experimented on the development set with seven heuristics, defined below, for automatically extracting MRE sentences from the 4,396 unannotated narratives in the training set. We measured the performance of each heuristic by selecting the index s_h of the sentence in each narrative with the highest heuristic score and calculating the root-mean-square error (RMSE), which measured the standard deviation of how

many sentences away the heuristic fell from a true MRE sentence.

$$\text{Let } N = \text{number of narratives}$$

$$S_i = \text{the set of human-annotated MRE sentences in narrative } i$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \min_{s \in S_i}\Big((s - s_{h_i})^2\Big)}$$

We used a linear combination of the three heuristics that achieved the lowest RMSE (lower is better, with 0 indicating a perfect match between the heuristic and the true MRE sentences), averaged across all narratives in the development set, to automatically extract MRE sentences from our training set.

Similarity to comment. The maximum bag-of-words cosine similarity[2] between sentence s_h and any of the comments replying to the narrative (Figure 2.3). We expected comments to refer to the MRE because of its shocking nature and importance to the narrative. This heuristic achieved RMSE of 5.5 sentences on the development set.

Comment: Yeah, this happened to me too, almost exactly the same. "⟨Friend⟩, what's your Chinese name?" ⟨awkward pause⟩ "...I'm from South Korea. I don't have a Chinese name." Beyond awkward.

Narrative: I was meeting people on my floor at one of those beginning of the year Floor Meetings your RAs put on (last year) and I was talking to an Asian girl named "Sunny" and she said that was her "American Name" because her real name was too hard for Americans to say. *So I asked "So how do you say your Chinese name?" and she said "...I'm Korean."* I wanted to die.

Figure 2.3: Example of the similarity to comment heuristic, with its highest-scoring narrative sentence *italicized*.

[2]Everywhere we use bag-of-words cosine similarity, we first stemmed the sentences and removed all stopwords.

Similarity to tl;dr. The latent semantic similarity between sentence s_h and the *tl;dr* of the narrative, if there was one (Figure 2.4). The tl;dr (which stands for "too long; didn't read") is a very short paraphrase of a post or comment written by its author. They are relatively rare – only 663 narratives, or 14% of our training set, had a tl;dr. Since the MRE is the central event of the narrative, we expect it to be included in the tl;dr. We calculated the similarity using the weighted matrix factorization algorithm described by Guo and Diab (2012). This heuristic achieved RMSE of 5.8 sentences.

Tl;dr: Kid in his underwear flips me off and trash talks me.

Narrative: Playing Xbox at my house, I look to see that my controller needs new batteries. Searching through my home, I find no batteries and decide to walk to the nearby liquor store to buy some. Down the street I see a small figure starting to walk towards me. Puzzled, I walk toward him hoping to see who he was. The figure was a kid of around 5 to 6 completely naked except for his underwear. *He flips me off and starts to trash talk me.* "Get over here bitch, come fight me!" "I will fuck you up bitch come on!" I was so confused I didn't know what to say. I started to run toward him to stop him from shouting. He looked and ran, screaming at the top of his lungs, and disappeared into the nearby alley.

Figure 2.4: Example of the similarity to tl;dr heuristic, with its highest-scoring narrative sentence *italicized*. The narrative is slightly edited for length.

In contrast, bag-of-words cosine similarity to the tl;dr performed poorly (RMSE of 13.2 sentences; this was very high, considering the average length of a narrative was only 16 sentences). This was due to the tl;dr being both very short and a paraphrase of its narrative. There are very few words in the tl;dr, and those words are often synonyms of, but not the same as, words in the narrative. Guo and Diab's latent semantic similarity score addresses this word sparsity problem by modeling not the words that are present in the input text, but the words that are missing from the input text. We also experimented with latent semantic similarity for the similarity-to-comment and similarity-to-prompt heuristics, but in these

two cases, it did not perform as well as bag-of-words cosine similarity.

Similarity to prompt. The bag-of-words cosine similarity between a sentence and the AskReddit post that prompted the narrative (Figure 2.5). The narrative should be relevant to the prompt, so we expect the MRE to be similar to the prompt text. This heuristic achieved RMSE of 6.3 sentences.

Prompt: Does anyone have any Bridezilla stories?

Narrative: I knew a woman who was a bridesmaid in a relative's wedding. She was married and had been trying to get pregnant for a while. Finally, her and her hubby got lucky and she conceived. *The bridezilla got furious and kicked her out of the wedding because she would be pregnant in the pictures.* 3 months later, sadly, the woman miscarried. The bride called her with a response along the lines of "good, well now you can be back in the wedding." Needless to say, she did not even attend it.

Figure 2.5: Example of the similarity to prompt heuristic, with its highest-scoring sentence *italicized*.

We used the heuristic with the fourth lowest RMSE as one of the baselines in our experiments:

Last sentence. The last sentence in the narrative. Since the events of a narrative build up to the MRE, the MRE sentence should occur near the end. This heuristic is the personal narrative equivalent of the lead baseline in news summarization. The last sentence heuristic achieved RMSE of 6.9 sentences.

Other heuristics. We also tried the following:

- Single-sentence paragraph (RMSE of 8.7 sentences). This heuristic was meant to capture emphasis, as an MRE might be placed in its own, separate paragraph to draw attention to it.

- First sentence (RMSE of 13.7 sentences). As previously discussed, a news-style lead baseline could work well on personal narratives with *abstracts* that mention the MRE; this heuristic was meant to capture a reference to the MRE in abstract. Unfortunately, most narratives do not have abstracts, and not all abstracts mention the MRE.

We extracted MRE sentences from the training set automatically, using a linear combination of the three best-performing heuristics: similarity to comment, similarity to tl;dr, and similarity to prompt.

$$h_{\text{label}} = 0.2 * h_{\text{comment}} + 0.5 * h_{\text{tldr}} + 0.3 * h_{\text{prompt}}$$

This linear combination outperformed each of the three alone, achieving an RMSE of 5.1 sentences (lower is better). The weights for each heuristic were tuned on the development set. For stories without a tl;dr, that heuristic was set to 0. The sentence in the narrative with the highest heuristic score was selected as the MRE sentence. We extracted only one MRE sentence per narrative using the heuristics, in constrast with the human annotation, which allowed any number of MRE sentences to be extracted from a narrative; we chose to extract a fixed number of MRE sentences automatically to avoid having to use a tuned threshold that might result in some narratives having dozens of MRE sentences while others had none.

In 52 of the 99 stories in the development set, we found that multiple, consecutive sentences were labeled by our annotator as MRE sentences. The average number of consecutive MRE sentences was 2.5, so to reflect this, we labeled our training set in three-sentence

		Number of Sentences	
Data Set	Stories	MRE	Total
dev*	99	169	1528
seed*	82	184	958
tuning*	95	212	1301
testing*	193	444	2771
training	4178	11205	67954

Table 2.2: Distribution of MRE sentences. * denotes human annotations.

blocks. The sentence selected by our heuristics, along with the immediately preceding and following sentences, were all considered MRE sentences. The result was the weakly-labeled training set in Table 2.2. Just over 16% of the weakly-labeled sentences were MRE sentences.

2.4 Experiments

Using this weakly-labeled Reddit dataset and the change-based view of narration described in Section 2.2, we conducted two experiments on automatically extracting MRE sentences. We compared our results with three baselines: random, our extraction heuristic, and the last sentence of the narrative (the next-best heuristic).

As described in the previous section, we labeled our training set in blocks of three consecutive MRE sentences, centered on the sentence from each narrative that was selected by our heuristic. To account for this, in our experiments and baselines, we predicted the presence of an MRE sentence in a three-sentence block: in testing, we considered a predicted block to be correct if it contained at least one human-extracted MRE.

2.4.1 Features

Stylistic Features. For each sentence in a narrative, we generated 176 sentence-level features capturing changes in the narration. We first scored each sentence using each of the sixteen metrics shown in Table 2.3[3]. The semantic metrics, *cossimilarity* and *lssimilarity*, refer to bag-of-words cosine similarity and latent semantic similarity of a sentence to the preceding sentence.

Type	Metric Names
Syntactic	sentlength, vplength, lengthratio, sentdepth, vpdepth, depthratio, wordlength, structcomplexity, wordformality, wordcomplexity
Semantic	cossimilarity, lssimilarity
Affectual	pleasantness, activation, imagery, subjectivity

Table 2.3: The sixteen narration-style metrics.

We then smoothed the scores across sentences in a narrative by applying a Gaussian filter. We also tried weighted and exponential moving averages, as well as a Hamming window, but the Gaussian performed best in experiments on our tuning set. Finally, we generated eleven features for each metric at each sentence: the sentence score; whether or not the sentence is a local maximum or minimum; the sentence's distance from the global maximum and minimum; the difference in score between the sentence and the preceding sentence, the difference between the sentence and the following sentence, and the average of these differences (approximating the incoming, outgoing, and self slopes for the metric); and the incoming, outgoing, and self differences of differences (approximating the second derivative).

[3]While lexical formality and complexity scores are not properly features of the syntax of a sentence, we considered them part of the same category as the truly syntactic features, whose goal was to capture the formality and complexity of the sentence.

Other Features. We included an additional ten features inspired by Labov's theory of narrative structure:

- The tense of the main verb and whether or not there is a shift from the previous sentence. Labov (2013) suggests a shift between the past and the historical present near the *most reportable event*.
- The position of the sentence in the narrative; the MRE usually appears near the end. We implemented position as four binary features by dividing the narrative into four sections.
- The bag-of-words cosine similarity and latent semantic similarity between the sentence and the first and second sentences in the narrative. While the MRE sentence usually appears near the end of the narrative, Labov (2013) notes that the abstract, a short introduction that occurs in some narratives, often refers to the MRE.

It is important to note that we did not use any lexical features in our experiments – the system we trained does not in any way model what a narrative is about. This was a deliberate choice in the design of our system. Because our Reddit narratives were collected from just 39 different prompts, a system that did use lexical features might learn a set of words topically related to each prompt and classify MRE sentences based on those words. Such a system would perform poorly on narratives about previously unseen topics. We designed our features to capture only the similarities and differences between a sentence in a narrative and its neighbors, keeping them independent of the actual content – the deep structure – of the narrative. Our hope was that the narration alone contained enough information to allow our system to automatically extract MRE sentences.

2.4.2 Distant Supervision

Our first experiment used distant supervision with our weakly-labeled training set: the heuristically extracted MRE sentences were treated as if they were gold standard labels. We classified blocks of three sentences as containing an MRE sentence or not. The two classes, *MRE* and *no-MRE*, were weighted inversely to their frequencies in the weakly-labeled training set, and all features were normalized to the range $[0, 1]$. We trained a support vector machine with margin $C = 1$ and an RBF kernel with $\gamma = 0.001$ (these parameters were tuned using grid search on our human-annotated tuning set).

Trial	Precision	Recal	F-Score
Last sentence baseline	0.208	0.112	0.146
Heuristic baseline	0.107	0.333	0.162
No change-based features*	0.146	0.378	0.211
Random baseline	0.185	0.586	0.281
Change-based features only*	0.351	0.685	0.466
All features*	**0.398**	**0.745**	**0.519**

Table 2.4: Distant supervision experiment results (* indicates significant difference from baselines, $p < 0.01$).

The results of the distant supervision experiment are shown in Table 2.4. We trained three different systems using different sets of features: the change-based features included all 176 stylistic features except for the metric scores themselves; the non-change-based features included the other Labov-inspired features and the metric scores, but none of the other stylistic features, such as slopes and distance from global extremes.

Our best results use both sets of features, but notably, using the change-based features alone achieved significant improvement over the three baselines ($p < 0.00005$). The no-change feature set was outperformed by the random baseline ($p < 0.0024$), supporting our

hypothesis that it is change in a stylistic metric, rather than the metric score itself, that predicts MRE sentences.

2.4.3 Self-Training

The distant supervision approach treats heuristically-labeled data as if it were human-labeled, gold standard data. The hope is that a large amount of noisy data would allow one to train a more general model than would a very small amount of clean data. However, there is the risk that some of these noisy training labels are so egregiously bad that they could lower the overall performance of the trained model.

Our second experiment used a self-training approach, where a classifier uses a small, labeled seed set to label a larger training set. In pure self-training, there is the risk that these predicted labels are incorrect or reflect strange outliers or biases in the seed set. We addressed the risks of both the distant supervision and self-training approaches by adding an additional quality control step to self-training: to ensure the quality of a self-training predicted label, we required that it agree with the heuristic weak label for that sentence in order to be added to the training set.

With the same parameter settings as in the distant supervision experiment, we trained an SVM on our hand-labeled seed set of 958 sentences. We used this initial model to predict labels for the training set, and all sentences where this labeling agreed with the heuristic labeling were added to the seed set and used to train a new model, which was in turn used to label the remaining unused sentences (ie. sentences whose predicted labels from the previous round had disagreed with the heuristic labels), and so on until none of the current

model's labels agreed with any of the remaining heuristic labels. Figure 2.6 shows the learning curve for the self-training experiment, along with the growth of the self-training set.

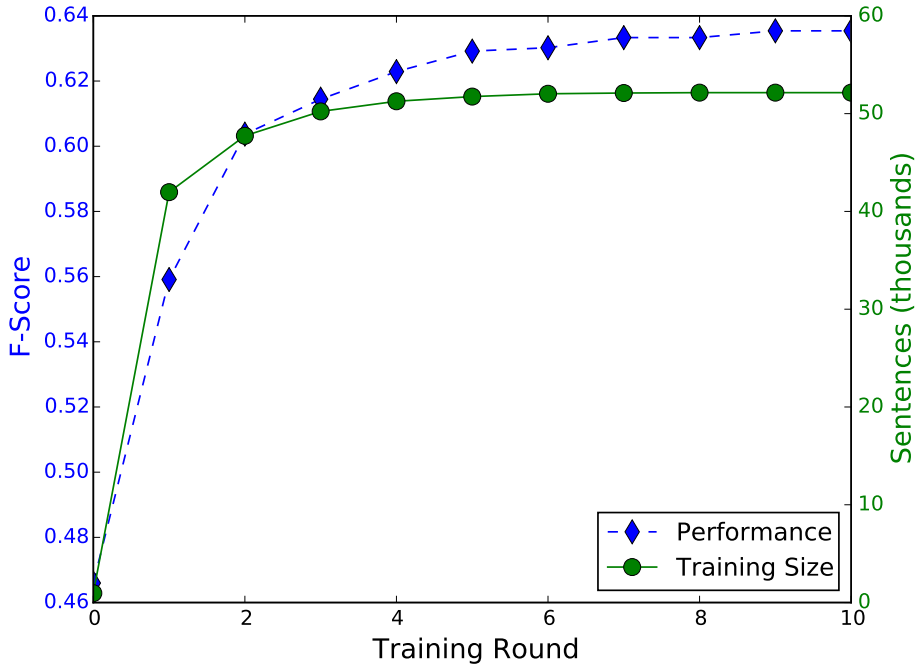

Figure 2.6: Learning and training set size curves for self-training.

The results of the self-training experiment are shown in Table 2.5. We achieved the best performance, F-measure of 0.635, after 9 rounds of self-training; self-training terminated after 10 rounds, but the 10$^{\text{th}}$ round had no effect on model performance.

Trial	Precision	Recall	F-Score
Random baseline	0.185	0.586	0.281
Seed training only*	0.374	0.617	0.466
Distant supervision*	0.398	0.745	0.519
Self-training*	**0.478**	**0.946**	**0.635**

Table 2.5: Self-training experiment results (* indicates significant improvement over the baseline, $p < 0.01$).

The initial model, trained only on the seed set, performed nearly as well as our distant supervision experiment. This illustrates that sheer quantity of data may not overcome the use of accurate manual labels on a small dataset. As described in Section 2.2, the distant supervision labels were based on a linear combination of three heuristics that achieved at best an RMSE of 5.1 sentences. However, with quality-controlled self-training, we were better able to exploit the noisy heuristic labels by using only those that agreed with the seed-trained model, thus reducing the amount of noise. 52,147 of the 67,954 total heuristically-labeled sentences were used in our quality-controlled self-training experiment – roughly 27% of our heuristic labels were too noisy to use.

2.5 Discussion and Error Analysis

Because we used an non-linear kernel in our SVMs, we were not able to examine feature weights directly. Instead, Table 2.6 shows the results of a logistic regression model trained on our features. The 10 best features are shown, along with their weights and 95% confidence intervals.

From feature 8, we see that MRE sentences tend to occur just before the narrative's global minimum in imagery. This fits Labov's obvservation that the *most reportable event* is usually followed by the *evaluation*, where the narrator praises or blames characters for their roles in the narrative and discusses possible alternative outcomes. Feature 1 indicates that MRE sentences show a sharp increase in imagery compared to previous sentences; the MRE is described in a burst of vivid language, followed by the more abstract author opinions.

	Feature Name	Weight	Confidence Interval
1.	incomingd2_imagery	4.174	(4.062, 4.287)
2.	distancefrommin_wordformality_neg	4.109	(3.952, 4.265)
3.	cossimilarity_adjacent	3.618	(3.425, 3.812)
4.	distancefrommin_activeness	3.377	(2.855, 3.298)
5.	sentdepth	3.364	(3.138, 3.590)
6.	distancefrommin_wordlength_neg	3.321	(3.018, 3.624)
7.	distancefrommin_vpdepth	3.034	(2.823, 3.247)
8.	distancefrommin_imagery_neg	2.790	(2.524, 3.056)
9.	wordformality_neg	2.329	(2.226, 2.432)
10.	incomingd2_vplen	2.128	(1.938, 2.318)

Table 2.6: Top 10 features in self-training experiment.

Features 2 and 9 indicate that MRE sentences tend to use informal language – a textual echo to Labov's observation that the subjects of his sociolinguistic interviews spoke less formally and more colloquially as they relived the climaxes of their stories.

Feature 3 suggests that MRE sentences are similar to the surrounding sentences. While we expected MRE sentences to be different from their neighbors due to the unusual and shocking nature of the MRE, this feature seems instead to reinforce the idea, initially suggested by our human annotator's labels, that MREs tend to described over the course of multiple, consecutive sentences, rather than in a single sentence. From feature 4, we see, as expected, that the MRE is far from the narrative's global minimum in activeness, as it is the end of a chain of events, far away from the stative *orientation*.

Finally, features 5 and 10 suggest that MRE sentences are not only long, but much longer than the preceding sentences, and feature 6 indicates that MRE sentences are close to the global minimum in average word length. Shorter average word length is expected, as an indicator of both informal word choice and emphasis. Long MRE sentences, however, were unexpected. Looking over our development set, we found that many authors combined the

description of the MRE with evaluative material in a single sentence, resulting in a longer and more syntactically complex MRE sentence than is found in Labov's data.

Our experiments showed that our change-based view of narration is effective for identifying MRE sentences, and this model provides evidence supporting the change-in-state definition of narrative suggested by Prince (1973). We achieved high recall with the self-training experiment (95%), but precision was low across the board, suggesting that, while MRE sentences do occur at extremes in syntactic complexity, semantic similarity, and emotional activation, there may be many non-MRE local extremes throughout a narrative.

Examining our results, we found a few common sources of error. False positive sentences tended to have high imagery and activeness. In Table 2.6, we saw that imagery and activeness alone do not indicate the presence of the MRE. An MRE sentence is not just active; it is separated from the stative introduction by the other events of the narrative. Nor is it enough for a sentence to have high imagery; the MRE must be even more vividly described than the preceding events – demonstrating again the importance of change in our view of narrative. False negative sentences tended to have high scores in syntactic complexity and formality.

2.6 Conclusion

The main contributions described in this chapter (based on work published in Ouyang and McKeown (2015)) are as follows:

- We created a new approach to modeling narrative, based on linguistic theories narration (Prince 1973; Polanyi 1985; Labov 1997). Our model tracks stylistic sentence

characterstics over the course of a narrative, capturing changes in complexity, meaning, and emotion.

- We created a new corpus of 4,896 personal narratives, taking advantage of AskReddit, a online community where members often prompt each other for stories. The corpus consists of both high-quality human annotations of sentences describe or refer to the *most reportable event* of the narrative, as well as noisy labels automatically generated using heuristic rules that leverage the comment-thread structure of Reddit content.

- We showed that our content-agnostic model is able to identify MRE sentences purely by modeling changes in stylistic features among sentences in a personal narrative. We also demonstrated that large quantities of hand-labeled data are not required for this task: the noisy heuristic labels were sufficient.

- We combined the distant supervision and vanilla self-training approaches into quality-controlled self-training, a learning scheme that successfully filtered out low-quality noisy labels.

This first foray into personal narrative summarization was intended as an extractive first step for a summarization system. The MRE is a real-world event that exists outside of the narration of a personal narrative, and thus an abstractive second step to rewrite the extracted MRE sentences into self-contained, standalone summaries followed naturally.

Unfortunately, we soon realized that the MRE sentence selection task as it was conceived in this early work was not a good parallel for extractive summarization. Our human

annotator selected all sentences that referred to the MRE, introducing two problems for extractive summarization. First, the extracted MRE sentences often did not include enough context to understand them in isolation. Second, extracting all sentences referring to the MRE often resulted in redundancy among selected sentences and the inclusion of extraneous information; many MRE sentences contained irrelevant text, such as narrator opinions. Thus, while the work described in this chapter was successful in detecting emotional impact in personal narratives, it did not produce usable extractive summaries.

Now obviously my balls drop and instinct kicks in as I speed up, it's quite a small roundabout so I need to either slow down enough to turn left, or speed up and nope the fuck straight out of there.

Figure 2.7: Example MRE sentence that does not work as an extractive summary.

Figure 2.7 shows the human-annotated MRE sentence for one of the narratives in our corpus. One can see from the colorful language in the MRE sentence that our annotator captured a very emotionally-charged moment. Unfortunately, the MRE sentence is difficult to understand without the rest of the narrative – what is the narrator afraid of?

Another problem with this first approach to personal narrative summarization was that it was unclear how to proceed from extracting MRE sentences to rewriting them into an abstractive summary. We had no reference summaries for what such a thing should look like – in fact, we had no gold standard annotations for what the MREs, the real-world events underlying the MRE sentences in the narration, were.

In the next chapters, we describe how we addressed the problem of creating a true personal narrative summarization corpus, containing both extractive and abstractive versions of the same summary for a narrative, as well as how we adapted the change-based model

of narrative for true extractive summarization of personal narrative.

3. A Summarization Corpus

To develop a two-stage summarization system with extractive content selection and abstractive editing, we need personal narratives paired with matching extractive and abstractive summaries. We have as a starting point our Reddit personal narrative corpus, with its *most reportable event* (MRE) sentence annotations. It is one of the two personal narrative corpora in existence; the other is the weblog narrative corpus of Gordon and Swanson (2009), which is mostly unannotated, although small subsets of it have been labeled with Labov's elements of narrative structure (Swanson et al. 2014). As discussed in the previous chapter, the MRE sentence annotations had several limitations that made them unsuited for the abstractive summarization task.

In this chapter, we describe how we address these limitations and transform the Reddit personal narrative corpus into a summarization corpus. The reference summaries in this new corpus are organized in matching pairs of human-written abstractive summaries and extractive summaries explicitly constructed to match the abstractive summaries in content; only a handful of corpora provide paired extractive and abstractive summaries of this kind. Our corpus consists of 1,088 unique abstractive summaries, written by trained human annotators, paired with crowdsourced extractive summaries. Also using crowdsourcing, each pair is annotated with phrase-level alignments, and each alignment is annotated with the six summarization rewriting operations identified by Jing and McKeown (1999): reduction (more commonly called compression), combination (more commonly called fusion), syntactic reordering, lexical paraphrasing, generalization, and specification. This is the only

summarization corpus in existence that provides not only paired, aligned extractive and abstractive summaries for each input document, but also an analysis of which text-to-text generation techniques are needed to rewrite an extractive summary into its corresponding abstractive summary.

3.1 Related Work

While there are many single-document summarization copora available, ours is the only one for personal narrative, as well as the only one to include text-to-text generation technique annotations. Corpora do exist for some of the individual rewriting operations described by Jing and McKeown (1999), such as compression (Filippova and Altun 2013; Kajiwara and Komachi 2016), fusion (McKeown et al. 2010), and lexical paraphrasing/syntactic reordering (Ganitkevitch, Van Durme, and Callison-Burch 2013), but the rewriting operations in these corpora are captured in isolation, whereas human summarizers may apply multiple operations to a single phrase. In this section, we do not give an exhaustive list of existing single-document summarization corpora; instead we discuss in detail those corpora that we either used or considered using in the work described in the rest of this dissertation, as well as any corpora with interesting properties similar to those of our personal narrative summarization corpus. We organize the corpora we discuss by genre: news, discussions, and web text.

3.1.1 News Summarization Corpora

While the corpora described in this section provide only abstractive reference summaries, they have all been used to train and evaluate both extractive and abstractive summarization systems. The number and size of news summarization corpora greatly outstrips those of any other text genre.

3.1.1.1 Document Understanding Conferences

Year	Documents	Summary Length
2001	300	100
2002	600	100
2003	640	10
2004	500	~12.5*

Table 3.1: DUC single-document news summarization datasets. *The summary length for DUC 2004 was 75 bytes, which is approximately 12.5 words.

Single-document news summarization was been a recurring task in the early Document Understanding Conferences (DUC), usually paired with multi-document news summarization. However, single-document summarization was removed after DUC 2004, and later DUCs retained only multi-document summarization tasks. Table 3.1 shows the four DUC single-document summarization datasets: the number of summarized documents provided for system development and training and the lengths of reference summaries. Of particular note are the provision of multiple reference summaries for each document, and the small sizes of the datasets and the short lengths of the DUC 2003-4 summaries.

3.1.1.2 The Gigaword Corpus

Rush, Chopra, and Weston (2015) created a summarization corpus of news article headlines (the summaries) and first sentences (the documents) using the annotated Gigaword corpus (Napoles, Gormley, and Van Durme 2012). They first collected about 9.5 million news articles from the Gigaword corpus and created pairs of headlines and first sentences, then filtered these 9.5 million pairs to remove spurious pairs, such as those where the headline and first sentence shared no content words in common, resulting in 4.2 million document-summary pairs. The Gigaword corpus was the first large-scale corpus designed for the training of end-to-end neural abstractive summarization systems. Both the documents and summaries in this corpus consist of single sentences and are extremely short: documents average 31.4 words, while summaries average 8.3 words.

3.1.1.3 The CNN/Daily Mail Corpus

Hermann et al. (2015) collected a corpus of 311,672 news articles from the CNN and Daily Mail websites, along with human-written bullet-point highlights, which they used to create cloze-style queries to train a neural question-answering system. Nallapati et al. (2016) modified this corpus for summarization by skipping the query creation step of Hermann et al.'s web crawling script and using the intact highlights for an article as its summary, treating each highlight as a sentence. Thus the CNN/Daily Mail corpus consists of multi-sentence summaries, in constrast to the single-sentence summaries in DUC 2003-4[1] and Gigaword: the average summary length in the CNN/Daily Mail corpus is 53 words (3.72

[1] DUC 2001-2, which had multi-sentence summaries, are no longer used in recent work on summarization.

sentences).

3.1.1.4 The New York Times Annotated Corpus

The New York Times (NYT) annotated corpus (Sandhaus 2008) contains 664,998 articles paired with summaries written by the New York Times Indexing Service. (While the NYT corpus properly refers to the entire set of nearly two million articles, hereafter when we discuss the NYT corpus, we mean the subset of articles that are paired with summaries). Paulus, Xiong, and Socher (2018) were the first to use this corpus for end-to-end neural abstractive summarization, noting that the summaries in the NYT corpus tend to be shorter (40 words on average) and more heavily rewritten than those in the CNN/Daily Mail corpus, where the summaries tend to be worded more similarly to their articles.

3.1.1.5 The Newsroom Corpus

The largest and most recent news summarization corpus was collected by Grusky, Naaman, and Artzi (2018), who scraped 1,321,995 articles from 35 news websites, using HTML metadata to identify the human-written summaries that are often used as social media or search result descriptions. Grusky et al. evaluted their document-summary pairs using three metrics: *coverage*, the percentage of summary words that appear in the document[2]; *density*, the average length of the shared word sequences between the summary and the document that contain each summary word; and *compression*, the ratio of document words to summary words. They found that, in general, the greater the summary compression rate for a given news website, the less diversity in coverage and density – that is, shorter

[2]This definition may seem counterintuitive, but the metric was intended to measure the degree to which the summary copied from the document, ie. how extractive (as opposed to abstractive) it was.

summaries were correlated with more standardized summarization strategies within a news organization.

Grusky et al. also compared their document-summary pairs with those of the DUC 2003-4, CNN/Daily Mail, and NYT corpora, finding that DUC had by far the highest median compression ratio: a document is 47 times longer than its summary. The NYT corpus, as Paulus et al. qualitatively observed, was quantifiably more abstractive than the CNN/Daily Mail corpus, with lower average coverage and density scores than CNN/Daily Mail. In comparison, the Newsroom corpus summaries cover a wider range of density and coverage scores than either NYT or CNN/Daily Mail.

3.1.2 Meeting Summarization Corpora

Meeting transcripts were the first major non-news genre to receive attention in text summarization research. By their nature, meeting summarization corpora are time-intensive to produce, requiring human annotators to review long recordings and transcripts. As a result, meeting summarization corpora are relatively small in terms of the number of document-summary pairs, and the documents and summaries are very long. Email threads are a related genre that introduce an additional difficulty: most collections of email are not publically available due to privacy concerns.

3.1.2.1 The ISCI Meeting Corpus

Janin et al. (2003) recorded and transcribed 75 naturally-occurring meetings at the International Computer Science Institute. Shriberg et al. (2004) segmented the meeting transcriptions into discourse units called *dialogue acts*, and Murray et al. (2005) created extrac-

tive and abstractive summaries for each meeting. Their annotators first wrote abstractive summaries of up to 800 words, in which they were instructed to cover certain pieces of information: the general purpose of the meeting; any decisions, progress, or achievements made; and any problems discussed. The annotators were then asked to select any number of dialogue acts from the segmented meeting transcription to cover the information in their abstractive summary, thus producing an extractive version of the summary. Finally, the annotators aligned the extractive summaries to the abstractive summaries by annotating each abstractive summary sentence with the dialogue acts that supported the correctness of that sentence. The result is a summarization corpus that is the most similar to our own, with three key differences:

- The size of the corpus is small, consisting of only 75 document-summary pairs.
- The documents and summaries in the ISCI corpus are very long. While the lengths of the meeting transcripts are not reported, each recording was roughly one hour; the summaries were up to 800 words long, which is too long for current neural abstractive summarization systems to generate.
- While this corpus is the only one besides ours to provide alignments between the extractive and abstractive summaries for a meeting transcript, it does not provide text-to-text generation technique annotations for the alignments.

3.1.2.2 The AMI Meeting Corpus

In contrast to the naturally-occurring meetings of the ISCI corpus, the AMI meeting corpus (Carletta et al. (2005) and McCowan et al. (2005)) consists of 141 *elicited* meetings, in

which participants role-played a design meeting for a new type of remote control, and one participant, assigned to be the project manager, was asked to write a report summarizing the meeting. The recorded meetings were, like those in the ISCI corpus, transcribed and segmented into dialogue acts, and extractive summaries were created by selecting dialogue acts to support the project managers' abstractive summaries. Unlike ISCI's extractive summaries, AMI's were not constructed by the authors of the abstractive summaries, and they were not aligned back to the abstractive summaries.

3.1.2.3 The Enron Email Dataset

Carenini, Ng, and Zhou (2007) created extractive summaries for 20 email threads from the publically available Enron email dataset, including both single-chain threads, where each email in the thread had only one reply, and tree-structured threads, where some emails in the thread had multiple replies. Annotators were asked to extract sentences to construct a summary roughly 30% as long as its email thread, as well as to classify each extracted sentence as either *essential* or *optional*, where essential sentences must be included in the summary, and optional sentences are helpful for understanding the essential ones and should be included if the length limit allows.

Carenini et al. used five annotators per email thread and scored each sentence based on the number of annotators who classified it as either essential or optional. They found that about 12% of sentences were *overall essential*, ie. agreed upon by all five annotators, with at least two annotators marking the sentence as essential and at least two others marking it as essential or optional. Unfortunately, this one agreement statistic leaves the reader wanting to know more: do the overall essential sentences form a complete summary, or are

there other sets of equally essential sentences, where each sentence in the set provides the same essential piece of information, such that the annotators disagreed on which sentence to extract from the set? (This question was addressed by Nenkova and Passonneau (2004).)

3.1.2.4 The BC3 Email Corpus

Ulrich, Murray, and Carenini (2008) annotated 40 email threads from the W3C email corpus. Three annotators worked on each email thread, producing first a 250-word abstractive summary, then aligning each sentence in their abstractive summary to a sentence in the email thread, and finally creating an extractive summary by selecting important sentences from the thread. Like the ICSI meeting corpus, the BC3 email corpus is structurally similar to the personal narrative corpus we describe in this chapter, except for its small size, long summaries, and its alignments being between abstractive summary and the input document rather than the extractive summary – Ulrich et al. treated the extractive summary as a completely separate task from the abstractive summary.

3.1.3 Web Corpora

Many web-based summarization tasks – such as summarizing live blogs (Avinesh, Peyrard, and Meyer (2018)) and Twitter topics (Sharifi, Inouye, and Kalita (2013)) – are framed as multi-document, rather than single-document, summarization. Relatively few dedicated corpora for either single- or multi-document web summarization of the kind described in the previous two sections exist (that is, pairs of documents and human-constructed gold standard summaries), partly because of the relative ease of scraping large amounts of web

text, and partly because early work on web summarization used mainly unsupervised approaches.

3.1.3.1 The Webis-TLDR-17 Corpus

Völske et al. (2017) created a large corpus of "found" summaries using the social bulletin board website Reddit. They filtered posts and comments from a publically available Reddit web crawl, retaining only those that contained some variation of the acronym "TL;DR", which stands for "too long; didn't read." The acronym TL;DR is usually followed in a post by a very short, often humorous summary written by the author of the post; it was first used in natural language processing by Ouyang and McKeown (2015), as we described in the previous chapter.

Post: Doing some traveling this year and I am looking to build the ultimate travel kit . . . So far I have a Bonavita 0.5L travel kettle and AeroPress. Looking for a grinder that would maybe fit into the AeroPress. This way I can stack them in each other and have a compact travel kit.

TL;DR: What grinder would you recommend that fits in AeroPress?

Figure 3.1

Völske et al. further filtered the posts to remove those written by summarization bots, for a final corpus of 4 million document-summary pairs. However, as with any heuristically-labeled dataset, Völske et al.'s summaries contain some noise. Figure 3.1 shows a document-summary pair from the Webis-TLDR-17 corpus. The TL;DR here is not really a summary of the post; the post contains the contextual information a reader needs to answer the question posed in the TL;DR.

3.2 The Reddit Personal Narrative Corpus

In this section, we briefly review the Reddit corpus, which we discussed in the previous chapter, and which served as our starting point for building a personal narrative summarization corpus.

We collected 4,896 personal narratives from AskReddit, which had been posted in response to 19 different prompts, along with readers' comments posted in response to the narratives. A human annotator labeled a subset of 476 narratives with *most reportable event* (MRE) sentences; she was instructed to first decide what she thought was the climactic event of the narrative and then label all sentences that referred to it. Our intent was that the MRE sentences should serve as an extractive summary of the narrative, but our instructions for our annotator did not elicit good summaries. Figure 3.2 shows the MRE sentence for a narrative, compared to the abstractive and extractive summary for that narrative from the corpus whose creation we describe in this chapter. The MRE sentence makes a poor summary because it is impossible to figure out what is happening in the narrative from reading the MRE sentence on its own – it is impossible even to tell that the MRE sentence comes from the same narrative as the summaries!

Another limitation of the original Reddit corpus is that we had no explicit representation of the MRE itself, only the sentences in the narrative that referred to it. Even if the MRE sentences had made reasonably good extractive summaries, we still had no example of what a good, self-contained, standalone description of an MRE – an abstractive summary – should look like. Since abstractive summarization is our ultimate goal, we begin building the new Reddit personal narrative *summarization* corpus by eliciting human-written

MRE: Now obviously my balls drop and instinct kicks in as I speed up, it's quite a small roundabout so I need to either slow down enough to turn left, or speed up and nope the fuck straight out of there.

Extractive: As I'm looking around as to what the fuck is going on, we approach the roundabout and there is a man in medical looking attire next to a woman in what looked like white pyjamas, with blood covering her clothing. I go straight, and as we go past the woman attempts to launch herself at my car as the man looked like he was trying to stop her.

Abstractive: 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.

Figure 3.2: An MRE sentence compared with extractive and abstractive summaries.

descriptions of the MRE.

3.3 Abstractive Summaries

We use the same subset of 476 narratives as in in the previous chapter, partitioned into 7 slices of 68 narratives each. We trained four graduate student annotators from the Department of English and Comparative Literature. Each was assigned four slices, three of which were shared with one other annotator and one slice that was common to all annotators (Table 3.2). Thus, for our 476 narratives, there are 408 with two different abstractive summaries written by two different annotators, and 68 with four different abstractive summaries written by all four annotators.

Annotator	Slices
Annotator A	1, 3, 4, 7
Annotator B	1, 2, 5, 7
Annotator C	2, 3, 6, 7
Annotator D	4, 5, 6, 7

Table 3.2: The four annotators' narrative slice assignments.

Each annotator participated in a 30-minute training session: they were told to identify the *most reportable event* of each narrative by imagining they were about to tell the story to a friend and wanted to ask, "Did I tell you about the time that…?" They were instructed to write one or two sentences to complete the question and include any context they thought necessary for their friend to understand it (e.g., "I slipped on a banana peel" is different from "I slipped on a banana peel *that my rival planted*"). The annotators were given a tutorial, consisting of five narratives accompanied by questions about which events and pieces of contextual information should be included in the summary, as well as an example summary they could read after answering the questions. The annotators worked independently, summarizing their assigned narratives over the course of three weeks; they produced a total of 1,088 different abstractive summaries.

The abstractive summaries allow us to address another limitation of our previous MRE sentence annotation scheme: the difficulty of measuring agreement. While we had found substantial agreement (Cohen's $\kappa = 0.729$) between two annotators on identifying MRE sentences on a subset of 100 personal narratives, we could not determine the cause of disagreement. We had asked the annotators to visualize the MRE internally and to select sentences based on that, but we had no way of knowing what each annotator thought the MRE was, and so there were two possible sources of disagreement. The annotators could have disagreed on which sentences best represented the MRE; one of the challenges of evaluating extractive summaries is that there is often not a single "correct" set of sentences, but rather many different sets of sentences that all convey roughly the same information (Nenkova, Passonneau, and McKeown 2007).

Alternatively, the annotators could have disagreed on what the MRE itself was; what

was shocking and unusual to one annotator might not have been shocking and unusual to the other, depending on his or her background. In Figure 3.2, for example, the author of the abstractive summary focuses on one event, the woman throwing herself at the narrator's car, while the MRE sentence annotator focuses on another, the narrator's reacting with a dangerous driving maneuver. With abstractive summaries explicitly stating what each annotator thinks the MRE is, however, we are able to measure agreement on the MRE itself. This is an important sanity-check, as low agreement on MREs would suggest a fundamental problem with our belief in a single climactic event that can serve as the summary of an entire personal narrative.

> You will be shown two short texts, each describing an event or situation. The task is to determine 1) whether or not the two texts describe the same event or situation, and 2) if so, how similar the descriptions are.
>
> Assume that the characters in both texts are the same, eg. if a character called "Joe" in one text is a soccer player, and a character referred to as "he" in the other text is also a soccer player, assume they are the same person.

Figure 3.3: Turker instructions for the abstractive summary agreement HIT.

We evaluate interannotator agreement on both the MRE and the context required to understand it using an Amazon Mechanical Turk HIT (Human Intelligence Task). Turkers were shown two abstractive summaries of the same narrative, written by different annotators – but not the narrative itself – along with the instructions shown in Figure 3.3. We then asked the Turkers to decide whether or not the summaries described the same event. 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 question was clear: the paired

summaries were either about completely different topics, or identical except for some lexical variation or extra pieces of information that we inserted into one or both summaries. Figure 3.4 shows one of the qualification test questions we used.

Feedback from Turkers indicated that answering both questions for the agreement HIT took about 15-20 seconds per; they were compensated at a rate of $0.05 per HIT. 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.

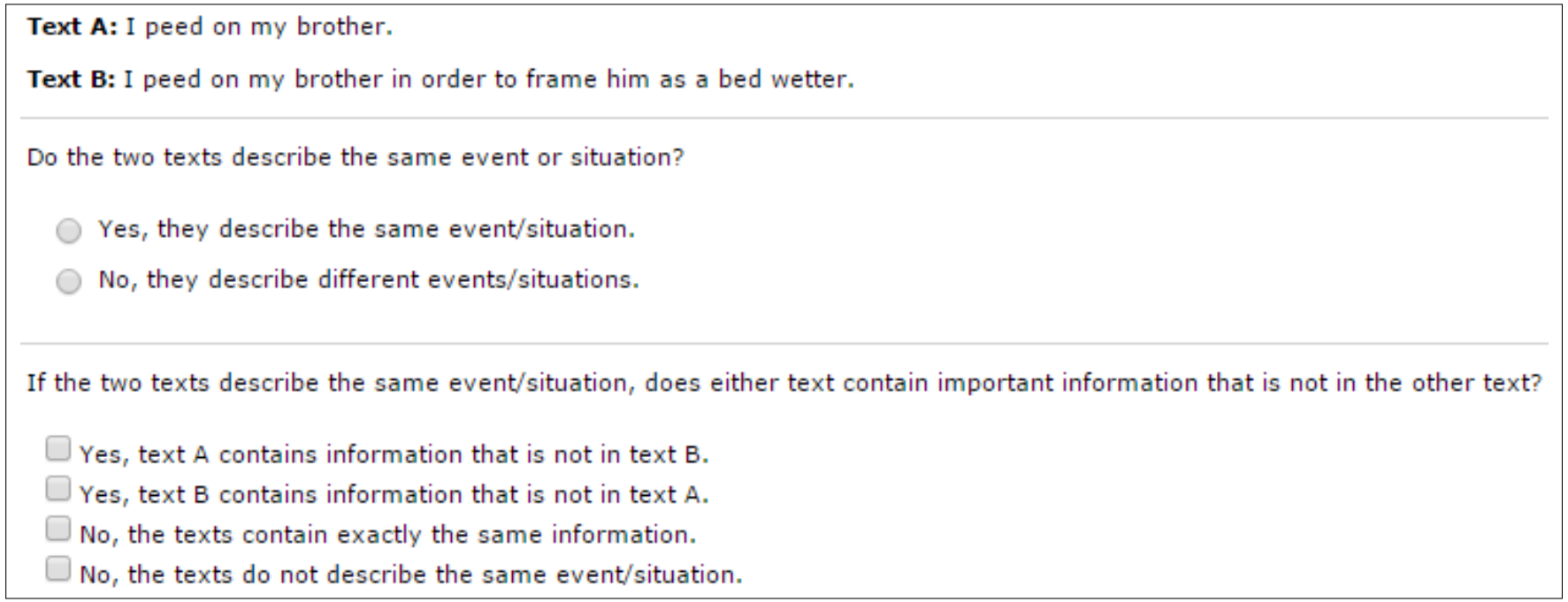

Figure 3.4: Example of qualification test question for the abstractive summary agreement HIT.

Table 3.3 shows each of our four annotators' average agreement with the other three annotators. The right two columns correspond to the percentage of narratives on which each annotator agreed with the other three annotators, depending on whether two (*consensus*) or all three Turkers (*unanimous*) agreed on whether or not the two summaries were about the same general event; each annotator shared 136 narratives in common with each other annotator. We wondered if narrative length might affect agreement, since longer narratives contain more events and contextual information for annotators to choose from, but we find that length has minimal effect: for narratives of more than 32 sentences, more than one standard deviation above the average narrative length of 16 sentences, *consensus* agreement among all annotators was 75.73%, and *unanimous* agreement was 88.41%.

	Abstractive Summary Agreement	
Annotator	Consensus	Unanimous
A	79.17%	91.05%
B	82.85%	91.25%
C	77.70%	87.80%
D	82.85%	91.33%

Table 3.3: Observed agreement among annotators on the main events described by their abstractive summaries.

Figure 3.5 shows a pair of agreeing and a pair of disagreeing abstractive summaries. With the disagreeing summaries, annotators A and B focus on different aspects of the narrative. As with the abstractive summary and MRE sentence in Figure 3.2, one annotator focuses on what happened to the narrator, while the other focuses on what the narrator did in response; B explains why the narrator was angry with the rude woman, while A summarizes the narrator's confrontation with the her.

If a Turker indicated that the two summaries they had read described the same event,

Abstract A: My neighbor's mom saved me from being kidnapped into a car when I was six.

Abstract B: Someone tried to kidnap me when I was six, but a neighbor's mom grabbed me before they got me.

(a) Agreeing abstractive summaries.

Abstract A: I ran my mouth off at this rude woman.

Abstract B: I held a door for a lady and she told someone on the phone that I had rudely ran around her.

(b) Disagreeing abstractive summaries.

Figure 3.5: Examples of agreeing and disagreeing abstractive summaries for two different narratives.

we asked them a second question: whether or not one or both of the abstractive summaries contained important information that was not in the other. This part of the HIT only applied to the abstractive summaries for narratives in slices 1-3, written by annotators A-C; annotator D was recruited last and had not finished writing summaries when we ran this HIT.

Of the 408 summary pairs in slices 1-3, the Turkers achieved consensus on 338 pairs. We find that, despite our annotators generally choosing the same MRE for their summaries, they rarely agree exactly on what contextual information should be included in the summary – the Turkers indicated that only 11.39% of agreeing summary pairs contained exactly the same information. In 23.72% of pairs, both summaries in the pair contained information not in the other, suggesting that annotators consider different aspects of the narrative to be important.

From Table 3.4, we see that annotators A and C choose the same MRE for their summaries for almost all of their shared narratives; annotator B differs slightly from A and C, but still chooses the same MRE for most narratives. Annotator B includes more infor-

	Extra Information			
Turkers	Abstract A	Abstract B	Both	No Difference
Consensus	18	38	18	6
Unanimous	8	23	3	2

(a) Comparing annotators A and B.

	Extra Information			
Turkers	Abstract B	Abstract C	Both	No Difference
Consensus	8	42	35	4
Unanimous	4	31	15	0

(b) Comparing annotators B and C.

	Extra Information			
Turkers	Abstract A	Abstract C	Both	No Difference
Consensus	17	25	42	5
Unanimous	5	13	13	0

(c) Comparing annotators A and C.

Table 3.4: Variation among annotators A, B, and C in abstractive summary content.

mation than annotator A in nearly half of their summary pairs (Table 3.4a, Figure 3.6a). Annotators A and C tend to focus on different aspects of the narrative: in nearly half of their pairs, both summaries contain information not in the other (Table 3.4c). For example, in Figure 3.6b, we see that annotator A focuses on the event's emotional effect on the narrator, while annotator C emphasizes the irresponsible friend's bad behavior.

3.4 Extractive Summaries

With human-written abstractive summaries in hand, we now need to construct extractive versions of the abstractive summaries. Because our ultimate goal is to rewrite extractive

Abstract A: My friends and I discovered a chilling porcelain doll on a bed in a creepy abandoned house.

Abstract B: My friends and I broke into an abandoned farmhouse and among a bunch of creepy things we found a life-sized porcelain doll set up in the bed.

(a) Extra information in abstractive summary B.

Abstract A: This one friend never gave back the 360 and netbook I let him borrow, so now I have a hard time doing good deeds for other people.

Abstract C: I lent my friend my netbook and xbox 360 and he broke the netbook and claimed the 360 was stolen. He only ever gave me 100 bucks for it.

(b) Extra information in both summaries.

Figure 3.6: Examples of extra information in pairs of agreeing abstractive summaries.

summaries into fluent, coherent, and easy to understand abstractive summaries, the ideal extractive summary for a narrative already matches its abstractive summary as closely as possible – most importantly, any information not present in the extractive summary cannot be recovered by a rewriting system, so we need to ensure that the extractive summaries cover the same information as their corresponding abstractive summaries.

We create another HIT that shows Turkers a narrative, one of the abstractive summaries written for that narrative, and instructions to compose an equivalent summary by selecting as few sentences as possible from the narrative (Figure 3.7). The abstractive summary thus serves as a guide for constructing the extractive summary.

You will be shown a short story where each sentence has been numbered. You will also be shown 1-2 sentences summarizing an event or situation from the same story.

The task is to construct a new summary equivalent to the one you are shown, using only the sentences in the story. Select the smallest possible number of sentences from the story such that, when read together, they give the same information as the summary sentences.

Your summary does not need to explain who characters are, eg. a character called “Joe” in the provided summary may be referred to only as “he” in your summary.

Figure 3.7: The Turker instructions for the extractive summarization HIT.

We again required Turkers to complete a qualification test before working on our HITs. The test consists of a single story and abstractive summary, written so that the summary is a word-by-word paraphrase of a single sentence in the narrative that did not overlap with any other sentences (Figure 3.8).

The Turkers reported spending about two minutes per HIT. We received a good deal of positive feedback on this HIT – the Turkers enjoyed reading the narratives and did not find the sentence selection task difficult. The Turkers also left positive feedback on their compensation rate of $0.20 per summary, suggesting that this HIT could be completed even less expensively than it was. Three Turkers worked on each HIT, and all of the HITs were completed within 24 hours of being posted.

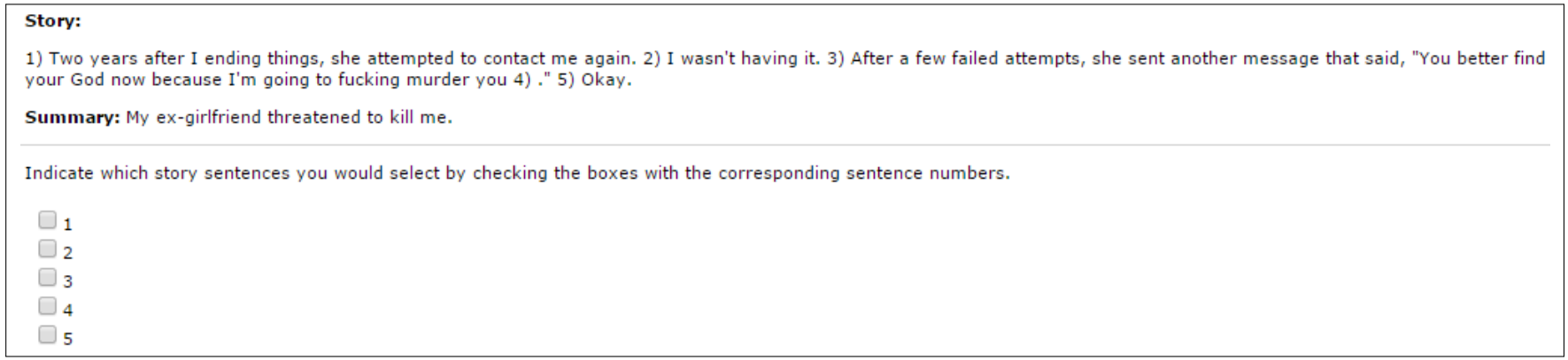

Figure 3.8: Example of qualification test question for the extractive summary sentence selection HIT.

Summary Set	Fleiss's κ
Abstract A	0.7364
Abstract B	0.7507
Abstract C	0.7604
Abstract D	0.7475

Table 3.5: Turker agreement on extractive summary sentence selection.

The Turkers achieved substantial agreement on which sentences they selected: Fleiss's κ of $0.7364 - 0.7604$, depending on which annotator wrote the abstractive summaries the Turkers used as a guide. Figure 3.9 shows an extractive summary where the three Turkers achieved perfect agreement.

Abstractive: At a concert, I grabbed a chunk of dirt in mid-air that was being thrown at a woman, and security thought I was throwing the dirt.

Extractive: There is a woman standing next to me when a huge piece of dirt comes flying straight at her face. I grab the chunk inches from her face mid-air. Security sees me with a chunk of dirt in my hand and instantly grab and pull me out of the crowd.

Figure 3.9: An example of perfect agreement among Turkers in constructing the extractive summary.

From the results of this HIT, we construct six different extractive summaries for each abstractive summary. There is one for each of the three Turkers and an additional three created by aggregating the Turkers' summaries: sentences selected by at least one (*union*), two (*majority*), and all three (*intersect*) Turkers. For 16.12% of narratives, the union and intersect summaries are identical (i.e., perfect agreement among the Turkers); only two narratives have no majority summary (i.e., no agreement at all among the Turkers). The average length of an aggregated summary is 4.15 sentences for union, 2.69 for majority, and 1.67 for intersect.

3.5 Phrase Alignments

With both the abstractive and extractive summaries finished, we turn to the rewriting process. One of our goals for applying text-to-text generation techniques to rewriting an extractive summary into an abstractive summary is targeted rewriting: learning not just how to perform a rewriting operation, but when it is appropriate to do so. Thus, we need to annotate not only which rewriting operations are present in a given extractive-abstractive summary pair, but roughly where in the two summaries the operations are applied. We break down this annotation into two steps: first, phrase-level alignments between corresponding phrases in the extractive summary and abstractive summary, and second, binary yes/no annotations for the presence of each rewriting operation in each pair of aligned phrases.

We use another AMT HIT to produce the phrase-level alignments between the extractive and abstractive summaries. We showed Turkers one of the abstractive summaries and its corresponding extractive summary (using *union* aggregation, to err on the side of recall over precision). 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 to the Turkers referred to the abstractive summary as the "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 3.10), displayed the extractive and abstractive summaries in text boxes and allowed Turkers to select phrases by directly highlighting them. Turkers could submit as many alignments per HIT as they could find,

so the interface allowed them to save alignments as they went along and submit all their saved alignments at the end.

As in the previous stages, we required Turkers to complete a qualification test to check their understanding of the task. However, since the alignment interface allowed them to highlight any string of characters, the results obtained from the interface would be too open-ended to judge straightforwardly in an automated qualification test. Instead, we created a multiple-choice 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.

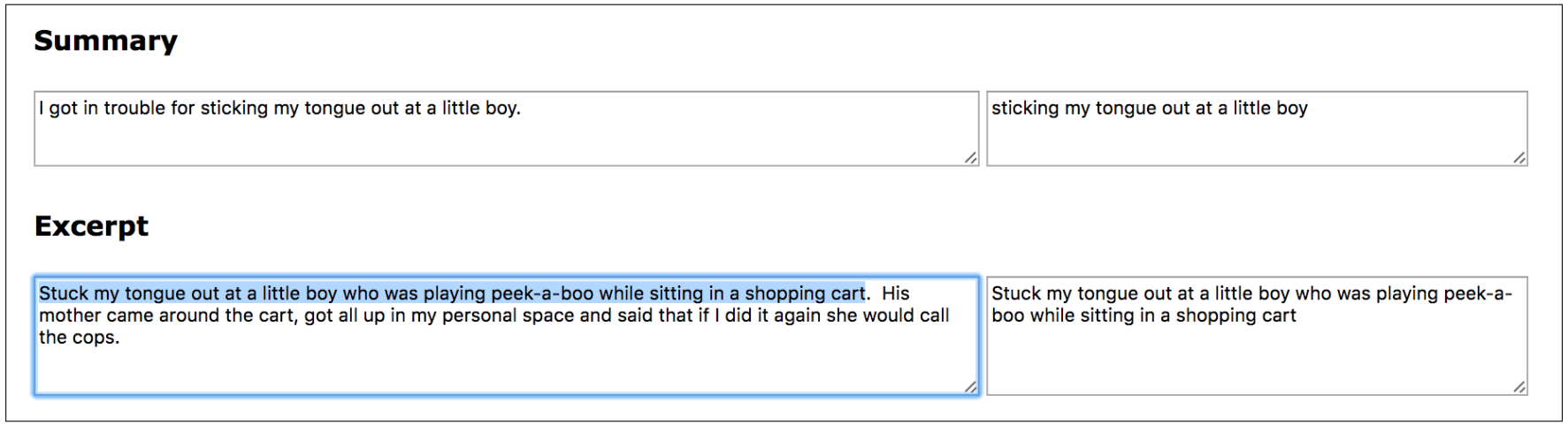

Figure 3.10: Highlighting interface for the phrase alignment HIT.

Annotator	Alignments
A	1424
B	1792
C	1514
D	1443
Total	6173

Table 3.6: Number of phrase alignments by abstractive summary annotator.

Three Turkers worked on each HIT, and they were compensated at a rate of $0.20 per HIT. Feedback from the Turkers again indicated that they enjoyed the HITs, both as a chance to read interesting material and for a good rate of pay; all of the HITs were returned within three days. Table 3.6 shows the breakdown of the 6173 total phrase alignments the Turkers generated across the four different abstractive summary annotators.

Worker A1PAY3X73PQ16S
Abstractive: My crazy ex-boyfriend *killed my kitten.*
Extractive: My crazy ex-boyfriend told me one day while we were still dating that I loved my 10 week old kitten more than him. *The next morning he called to tell me that he woke up late, jumped out of bed, and "accidentally crushed him with both feet" in his exact words.*

Worker A31Q7PWLBCU2X0
Abstractive: *My crazy ex-boyfriend killed my kitten.*
Extractive: My crazy ex-boyfriend told me one day while we were still dating that I loved my 10 week old kitten more than him. *The next morning he called to tell me that he woke up late, jumped out of bed, and "accidentally crushed him with both feet" in his exact words.*

Figure 3.11: Example of two agreeing phrase alignments, shown in *italics*.

It is difficult to determine interannotator agreement on phrase alignments because each Turker could submit multiple alignments for each summary pair, and alignments could be between any two substrings of the summaries. We consider two phrase alignments to be in agreement if 1) different Turkers submitted them, 2) the selected abstractive phrases

overlap enough such that at least half of the shorter of the two is covered by the overlap, and 3) the selected extractive phrases overlap enough such that at least half of the shorter of the two is covered by the overlap. Figure 3.11 shows an example of two agreeing phrase alignments. We consider a phrase alignment that is in agreement with at least one other phrase alignment to be a *confident* alignment – out of the 6,173 phrase alignments, 5,836 (95%) were *confident*.

3.6 Rewriting Operations

Finally, we annotate each phrase alignment with text-to-text generation techniques. Our final HIT asked Turkers to review the phrase alignments described in the previous section, and to identify which of the rewriting operations defined by Jing and McKeown (1999) were involved in transforming an extractive phrase into its aligned abstractive phrase. Because an alignment can employ more than one operation, we did not want to confuse Turkers by showing them all of the operations at once, so we simplified the task by creating separate HITs for each combination of phrase alignment and rewriting operation. When performing the task, Turkers were only asked about with one rewriting operation at a time and simply had to indicate whether the current alignment employed that operation or not.

We defined the following rewriting operations for the Turkers, 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.

We do not include fusion in this task, instead detecting it automatically from multiple alignments between a given phrase from a single abstractive sentence and multiple other phrases from multiple extractive sentences. Figure 3.12 shows the full instructions for the *reduction* HIT, which includes an example. As in the previous section, the instructions to the Turkers referred to the abstractive summary as the "summary" and the extractive summary as the "excerpt" to avoid confusing terminology.

> You will be shown two versions of the same story, the first an excerpt from the original and the second a summary. You will see each passage has one phrase in bold. The task is to examine the bold phrases and decide whether reduction was used to transform the bold phrase from the excerpt into the bold phrase from the summary.
>
> **Definition**: Reduction keeps the key parts word-for-word, and removes less important information.
>
> **Excerpt**: *Already late for my meeting,* I waited impatiently at the bus stop *and hoped the crosstown would arrive soon.* $\rightarrow$ **Summary**: I waited impatiently at the bus stop.

Figure 3.12: Instructions for *reduction* rewriting operation HIT.

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

For each alignment, we put up four HITs (we combined generalization and specification so that Turkers could choose one or the other or neither, but not both). Four HITs per alignment totaled 24,696 HITs, and we had three Turkers work on each HIT. Since the task was very quick, we compensated them for $0.01 per HIT, and with this rate of pay, the entire set of HITs took one week to finish.

Rewriting Operation Counts			
Fusion	2123	Reduction	216
Syntactic Reordering	916	Generalization	3359
Lexical Paraphrasing	1218	Specification	1250

Table 3.7: Number of phrase alignments employing each rewriting operation.

Table 3.7 lists each rewriting operation and how many alignments used it; we include an alignment when at least 2 out of 3 Turkers agreed it used the operation. We see that generalization is by far the most popular rewriting operation, and reduction is the least, likely because reduction's definition was the most restrictive, as it required word-for-word matching outside of the removed parts. Figure 3.13 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 3.13: Examples of the two most common rewriting operations, generalization and specification.

We did not ask Turkers to indicate where in a phrase alignment a rewriting operation occurred, only that it was present somewhere in the alignment, because the presence of multiple rewriting operations within a phrase alignment could make it very difficult to

identify exactly where each operation occurred. We define a *precise* alignment to be one that does not contain an extractive phrase over two sentences in length (the abstractive summary is already limited to two sentences in length). The longer the alignment, the less likely the two halves of the alignment were to be a perfect match on meaning, and the more difficult it could be to identify the rewriting operations involved. Of the 6,173 phrase alignments, 5,602 (91%) were *precise*, and 5,281 (86%) were both *confident* and *precise*.

Extractive: I came back to her introducing me to her new boyfriend and me subsequently being kicked to the curb. *2 years later we got to talking over facebook and I initiated a booty call.* Afterwords, She told me she had feelings for me again and i responded with ”cool, i’m late for dinner, talk to you later” BEST MASHED POTATOES EVER WERE HAD.

Abstractive: My girlfriend dumped me for some other guy, but *we hooked up two years later* and I had the best mashed potatoes ever.

Figure 3.14: Example of a confident and precise alignment (in *italics*) with multiple rewriting operation labels: lexical paraphrasing, generalization, and reduction.

Examining the rewriting operation labels produced for the confident and precise phrase alignments, we find that many alignments are labeled for more than one operation, indicating that quality phrase transformations for summarization often involve stitching together multiple rewriting operations. Figure 3.14 shows an example of such an alignment, which is labeled for lexical paraphrasing (perfect agreement from 3/3 annotators), generalization (perfect agreement from 3/3 annotators), and reduction (majority agreement from 2/3 annotators). Table 3.8 shows the interactions between rewriting operations in the form of a co-occurence matrix of the five rewriting operations we labeled using Amazon Mechanical Turk, plus fusion, which we identify automatically.

	Fusion	Reduction	Lexical Paraphrasing	Syntactic Reordering	Generalization	Specification
Fusion	**1052**	36	214	151	695	165
Reduction		**185**	34	32	113	24
Lexical Paraphrasing			**1068**	179	564	237
Syntactic Reordering				**772**	391	165
Generalization					**2802**	0
Specification						**1093**

Table 3.8: Rewriting operation co-occurences produced from confident and precise alignments. Because an alignment can contain multiple rewriting operations, the sums of the rows are greater than the total counts for each individual rewriting operation.

3.7 Conclusion

The main contribution described in this chapter (based on work first published in Ouyang, Chang, and McKeown (2017)) is our creation of a personal narrative summarization corpus consisting of 1,088 aligned abstractive and extractive summaries, totaling 6,173 phrase-level alignments, with each alignment annotated with rewriting operations. For each abstractive summary, our summarization corpus also includes information from our original Reddit corpus: the AskReddit prompt in response to which the personal narrative was written, as well as any comments posted in response to the narrative. The prompts and comments were key to both the MRE sentence classification system described in the previous chapter and extractive summarization system we describe in the next chapter.

Our corpus provides crucial data that other existing summarization corpora do not:

- Paired abstractive and extractive summaries explicitly constructed to be as similar as possible to each other, as well as phrase-level alignments between matching phrases in each pair. Existing corpora provide only gold standard extractive summaries (Enron), only abstractive summaries (DUC 2003-4, Gigaword, CNN/Daily Mail, NYT, Newsroom, and Webis-TLDR-17), or separate abstractive and extractive summaries that do not necessarily cover the same information (BC3). Only the ISCI and AMI meeting corpora provide extractive summaries explicitly constructed based on an abstractive summary; of those two, only ISCI provides alignments between corresponding extractive and abstractive summaries, and it does not provide rewriting operation labels for those alignments.

  Direct, end-to-end abstractive summarization is a difficult task, especially with longer

input documents and target summary lengths. Summarization corpora that provide only abstractive summaries force systems to try to learn both the content selection and editing subtasks at once, while corpora that provide paired extractive and abstractive summaries allow systems to learn first to select content in the form of an intermediate extractive summary, and then to edit the selected content into a fluent, coherent abstractive summary.

- Text-to-text generation technique annotations for each phrase alignment. Rewriting operations can oppose each other, most obviously in the case of generalization and specification, but also in the case of a de facto seventh operation: *do nothing*, which opposes every other operation. An abstractive summarization system trained on data that does not distinguish among rewriting operations must either learn some representation of them on its own, or commit to an operation-agnostic approach that either favors one operation over another or performs all of them poorly. No other existing corpora provide annotations for all six of the rewriting operations identified by Jing and McKeown (1999); our contribution fills a gap among summarization corpora.

Except where otherwise stated, we use this dataset for the rest of the work described in this thesis, occasionally augmented by other, larger datasets for training neural networks. As we discuss in the next chapters, our extractive and abstractive summarization systems are designed for the highly emotional style of these narratives, in contrast to the traditional summarization genre of newswire text.

4. Extractive Summarization

Using our completed Reddit personal narrative summarization corpus, we now revisit our approach to MRE sentence detection and adapt it to perform extractive summarization. The *most reportable event* (MRE) sentence prediction task to which we originally applied the change-based approach was directly based on the same contextualist narratological theories that inspired the approach itself: authors use *evaluation devices*, such as changes in sentence length, sentence complexity, or verb tense, to mark important content in their narratives, and every narrative has an MRE, the most important and most heavily evaluated event in the narrative.

The new extractive summaries in our Reddit summarization corpus differ from the earlier MRE sentences in two ways. First, the Amazon Mechanical Turk workers (Turkers) who constructed them were not given any instructions about the MRE at all; the annotators who wrote the abstractive summaries were instructed to write about the MRE, and the Turkers were told only to match the abstractive summary as closely as possible using as few sentences as possible. Second, the annotators were instructed to include any non-MRE orienting information that was required to understand the MRE, so both the abstractive and extractive summaries could include non-MRE information. Thus, it is possible that the change-based approach may not work well for extractive summarization; the sentences in an extractive summary may not be the heavily evaluated sentences that the approach was designed to detect, but rather high-coverage sentences containing many different pieces of information, or orientation sentences giving essential background information.

Despite these differences, we believe that the change-based approach will still work for extractive summarization. A key claim of contextualist narratology is that the evaluated clauses of a personal narrative form an *adequate paraphrase* of that narrative (Polanyi 1981) – what is a summary if not an adequate paraphrase? In this chapter, we describe how we tested this hypothesis by applying the change-based approach to modeling narrative to true extractive summarization.

4.1 Related Work

The task of single-document extractive summarization has been studied in natural language processing for over sixty years. In this section, we do not give an exhaustive list of extractive summarization systems, but focus instead on the main novelty of our change-based approach: the way in which a document and its sentences are represented. With the change-based approach, we represent each sentence in a document by comparing its scores on several stylistic metrics to the scores of neighboring sentences. We organize our discussion of other document representations as follows: topic-based representations, graph-based representations, structural representations, and neural representations.

4.1.1 Topic-Based Representations

The simplest method of determining the main topic of a document is to take normalized frequency counts for each word; these were used in the very first attempt at automatic extractive summarization (Luhn 1958). The idea is that words that occur frequently in a document are associated with the main topic of that document, and so sentences contain-

ing many such high-frequency words are good candidates for extraction. Luhn filtered out stopwords and manually tuned frequency thresholds for determining the set of words that was characteristic of a document, and later work refined the approach, using more sophisticated metrics for word importance, such as TF*IDF (term frequency * inverse document frequency) and the log-likelihood ratio (Nenkova and McKeown 2011). These metrics compare the frequency of a given word in a document with its frequency in a large set of documents belonging to the same domain; words that occur more frequently in the document to be summarized but less frequently in the other documents are more representative than words that are common in all documents. Some notable systems that used word frequency approaches are Lin and Hovy (2000), Nenkova, Vanderwende, and McKeown (2006), and Vanderwende et al. (2007).

Once the topics of a document have been identified, an extractive summarizer can choose one or more sentences to represent each topic in the summary. Other approaches to identifying the topics of a document include lexical chains (Barzilay and Elhadad 1999), latent semantic analysis (Gong and Liu 2001), and topic modeling (Daumé III and Marcu 2006). The advantage of topic-based approaches is that they are unsupervised; they require no human-written reference summaries, only collections of in-domain documents for calculating TF*IDF, fitting a topic model, and so on (or in the case of lexical chains, a lexicon of semantic relations to build the chains). These approaches are also language-independent, assuming the required resources (document collections or semantic lexicons) are available.

4.1.2 Graph-Based Representations

Like topic-based representations, graph-based representations focus on the words in each sentence. Each document is represented as a graph whose vertices are sentences and whose edges are weighted by the similarity between the two endpoint vertices. Any sentence-level similarity measure can be used, such as normalized word overlap counts, string kernels, or cosine similarity between TF*IDF vectors. In this graph, high-weight cliques roughly correspond to topics, and graph centrality algorithms can be applied to rank sentences in order of importance (Mihalcea and Tarau 2004; Erkan and Radev 2004).

Because they depend on similarity between sentences, graph-based representations are at a disadvantage with short documents (fewer sentences available to provide supporting evidence of a given sentence's importance) and documents with little repetition among sentences (low similarity scores throughout); personal narratives are short documents with little repetition among sentences, and we show later in this chapter that a graph-based baseline summarizer performs poorly on this genre.

4.1.3 Structural Representations

These representations include the position of a sentence in the document, its position within its paragraph, the position of its paragraph in the document, and its proximity to section headers, such as "Introduction" or "Conclusion" (Edmundson 1969), as well as the length of the sentence (Osborne 2002) and the presence of named entities and discourse markers (eg. "further" or "for example") (Fuentes Fort, Alfonseca, and Rodríguez Hontoria 2007), and of quoted speech (Wong, Wu, and Li 2008) in the sentence.

Positional features can be very strong in certain genres. In many sub-genres of news, such as newswire and world news, selecting the first sentence of an article (the *lead*) is an extremely powerful baseline for extractive summarization, to the point where human-engineered features often underperform it (Nenkova 2005). Later in this chapter, we show that positional features work for summarizing personal narrative – although not in the same way as in news – and other structural features, such as sentence length, are also useful.

4.1.4 Neural Representations

Recent work on extractive summarization has turned to neural models, where each word is represented by a *word embedding*, a high-dimensional vector in a distributional semantic space. A sentence is a sequence of words, and so the *encoding* of a sentence is the combination of the embeddings of its words (another high-dimensional vector, possibly of a different size than the word embeddings); similarly, a document is a sequence of sentences. The main approaches to encoding sentences and documents for neural extractive summarization are training convolutional or recurrent neural networks to produce sentence encodings from word embeddings and document encodings from sentence encodings (Cheng and Lapata 2016; Nallapati, Zhai, and Zhou 2017), although simply averaging the word embeddings often works just as well as trained sentence encodings (Kedzie, McKeown, and Daume III 2018).

Neural representations are difficult to interpret, compared to topic-based, graph-based, or structural representations. In end-to-end training, neural extractive summarizers learn the sentence and document representations that give the best performance on their valida-

tion sets, and these learned representations do not necessarily correspond to any linguistic or rhetorical theories or human intuitions about which sentences should or should not be included in a summary. A further disadvantage of neural representations is their need for large numbers of document/reference summary pairs to learn.

One final interesting note is that existing neural extractive summarization systems implicitly make use of some positional information in that they treat sentence selection as a sequence tagging task. As previously discussed, the *lead*, or first-sentence, baseline is very strong in the news genre, and Kedzie, McKeown, and Daume III (2018) found that neural extractive summarizers trained on news learn to identify the first sentence despite having no explicit positional features: performance dropped significantly when summarizers were trained on documents with the order of the sentences shuffled.

4.2 Using the Reddit Summarization Corpus

One of the goals of the work described in this chapter is to compare the performance of the contextualist-inspired, change-based approach on extractive summarization to its previous performance on the MRE sentence detection; we use the same development, seed, tuning, and testing set split of the 476 annotated personal narratives as in our previous experiments.

4.2.1 Reference Summary Construction

The Reddit personal narrative summarization corpus provides either two or four human-written abstractive summaries for each narrative and up to six crowdsourced or aggregated extractive summaries for each abstractive summary. Thus, our first task is to construct a

single extractive reference summary for each narrative in the corpus. Choosing a single extractive summary for each abstractive summary is relatively straightforward: we use the *majority* aggregation scheme, which included any sentence selected by at least two out of three Amazon Mechanical Turk workers. Of the other two aggregation schemes, *intersect*, which required sentences to be selected by all three workers, is too strict – in about 12% of narratives, the intersect-aggregated summary is empty – while *union*, which combines all sentences selected by all workers, is likely to include redundant sentences. Further, the average length of a majority-aggregated extractive summary is 2.69 sentences, very close to the average of 2.5 MRE sentences per narrative in our previous experiments.

How best to combine the majority summaries corresponding to two different annotators' abstractive summaries into a single extractive summary for each narrative is less clear. Differences among the majority summaries reflect, to a large extent, differences in the content of their corresponding abstractive summaries. In the previous chapter, we used crowdsourcing to examine differences in content in the abstractive summaries in our corpus, finding three possible cases:

1. The abstractive summaries are about completely different main events. In this case, the majority summaries are completely disjoint sets of sentences.

2. The abstractive summaries are all about the same main event and cover exactly the same information. In this case, the majority summaries are ideally identical, although in practice, there are multiple sets of sentences that can reconstruct the information in the abstractive summaries.

In both of these cases, neither of the abstractive summaries is any more or less "correct" than the other, so we take the union of the sentences in the majority summaries.

3. The abstractive summaries are about the same event, but one or both of them covers information not in the other. In this case, there is likely to be a shared subset of sentences between the two majority summaries, with some extra sentences in either one or both.

In this last case, there is some piece of information that one annotator considered essential to a reader's understanding of the summary, but that another annotator considerd unimportant. As we have no way to tell which annotator is "correct," we experiment with two different methods for handling extra information: keeping it or dropping it.

- *Keep:* We take the union of the sentences in the majority summaries, thus retaining any extra information present in either summary.
- *Drop:* We do not include any sentences present only in the more-informative majority summary. If both majority summaries contain information not present in the other, we take the intersection of the sentences in the two summaries.

For the 68 narratives with four different abstractive summaries, we begin building the extractive reference summary using two of the majority summaries as described above and then working in the other two, adding or removing sentences as needed, following the same rules. The average length of a completed extractive reference summary is six sentences, using the *keep* method of handling extra information, while the average length of a completed extractive reference summary using *drop* is four sentences.

4.2.2 Pyramid Weighting

Because the extractive reference summaries are constructed by merging multiple majority-aggregated Turker summaries, it is not necessarily the case that all sentences in a reference summary are equally important. Up to twelve Turkers contributed to each reference summary, but we only require a sentence to be selected by two Turkers to be included. A natural next step to ensure the quality of our reference summaries is to apply the *pyramid method* of Nenkova and Passonneau (2004).

The pyramid method for summarization evaluation was designed to address the possibility of there being multiple sentences in a document that express overlapping or identical information – in other words, there may be multiple, equally correct extractive summaries that express the same information using different sets of sentences. The pyramid method addresses this concern by breaking each summary into *summary content units* (SCUs), roughly clause-level units of meaning. A system summary is evaluated based on the SCUs it contains, where each SCU is weighted according to the number of human summaries that contain it. The set of weighted SCUs taken from all human summaries forms the *pyramid*, the gold standard reference summary.

While we do not have SCU annotations for our data, we can still use the pyramid method to weight the sentences in our extractive reference summaries. A sentence selected by all twelve Turkers, for example, likely expresses crucial information not present in any other sentence in the document; a system-generated summary should be penalized more heavily for failing to include such a sentence than for omitting another sentence selected by only two Turkers, the information in which might well be present in some other sentence

that the system did extract.

In the development, seed, tuning, and testing sets, we assign each sentence a pyramid weight equal to the number of Turkers who selected it; because we use majority aggregation to construct the extractive reference summaries, the pyramid weight for all sentences in the extractive reference summary is at least 2, while the weight of the other sentences is 1.

4.2.3 Heuristic Labeling

We use a new linear combination of the same similarity metrics from our previous experiments (semantic similarity to comment, tl;dr, and prompt) to label each sentence as extracted or not, using a threshold tuned on our pyramid-weighted development set:

$$h_{\text{comment}} + h_{\text{tldr}} + h_{\text{prompt}} \stackrel{?}{>} 1.544$$

The 67,954 training sentences labeled using this heuristic are assigned a pyramid weight of 1.

4.3 Experimental Results

We test our change-based model of narrative using both distant supervision and quality-controlled self-training on the same 193-narrative test set as in our previous experiments. Our task is to classify a given sentence in a narrative as either extracted or not. As in the previous chapter, the distant supervision experiment trains on the full set of 67,954 heuristically-labeled sentences (using the new heuristic shown above). The quality-controlled

self-training experiment is first trained on the human-labeled seed set of 82 narratives, then after each round of training, compares its own predicted labels on the larger training set against the heuristic labels and trains again on those training sentences for which the heuristic agreed with its predicted labels.

Table 4.1 shows the "mass" distribution of unweighted and pyramid weighted extracted sentences in the test set; while the positive class, extracted sentences, is relatively rare, the pyramid weighting scheme does a great deal to balance the positive and negative classes, for both *drop* and *keep* reference summaries.

Reference Summary	Unweighted Extracted	Pyramid Weighted Extracted	Not Extracted
Drop	444	1214	2,391
Keep	675	1721	2160

Table 4.1: Distribution of unweighted and pyramid weighted labels in the test set.

Table 4.2 shows the results of our distant supervision and quality-controlled self-training experiments on the *keep* reference summaries, which takes the union of sentences in the majority-aggregated extractive summaries for a given narrative. For both distant supervision and self-training, training and testing on the *keep* reference summaries outperforms the *drop* summaries, which takes the intersection sentences in the majority-aggregated summaries, by about 0.1 F-score, suggesting that for the amount and quality of data points we have for the positive class (extracted sentences), more is better. We use an SVM with RBF kernel, $\gamma = 0.001$, and the margin C was set to 100 for distant supervision and 10 for self-training, tuned on the tuning set.

The quality-controlled self-training experiment terminates after five rounds and uses 81% of the weakly-labeled training sentences. Our previous self-training experiment used

Trial	Precision	Recall	F-Score
TextRank	0.281	0.384	0.324
Distant supervision	0.362	0.821	0.502
Self-training	**0.465**	**0.956**	**0.626**

Table 4.2: Binary sentence classification results using *keep* reference summaries.

only 73% of the training sentences over 10 rounds, suggesting that our new heuristic labels more closely match our reference summaries than our old heuristic labels matched the MRE sentences. We compare our results to a graph-based summarizer (Table 4.2), the Sumpy[1] implementation of TextRank (Mihalcea and Tarau 2004); to construct the TextRank summary, we select the top k ranked sentences, where k is the number of sentences in either the distant supervision or self-training summary for that narrative, whichever is longer.

Trial	ROUGE-1	ROUGE-2	ROUGE-L	Summary Length
Kedzie et al.	0.208	0.054	0.189	5.67
Distant supervision	0.441	0.306	0.397	8.67
Self-training	0.567	0.423	0.513	7.70

Table 4.3: ROUGE-1, -2, and -L scores using *keep* reference summaries, and the average number of sentences in produced summaries.

We also compared our results to a neural network summarizer, Kedzie, McKeown, and Daume III (2018)'s reimplementation of Cheng and Lapata (2016)'s neural extractive summarizer, using word embedding averaging for the encoder[2]. This comparison is somewhat unfair in our favor. Neural approaches require large amounts of training data, so one should not expect a neural model trained on our relatively small corpus to perform extremely well;

[1] https://github.com/kedz/sumpy

[2] Kedzie et al. also experimented with other neural extractive summarizers and encodings, but averaging and the Cheng and Lapata system performed the best on our corpus.

for comparison, the same neural model trained and tested on the much larger CNN/Daily Mail corpus achieved ROUGE-1 of 0.389, ROUGE-2 of 0.177, and ROUGE-L of 0.366.

Predicted Summary: About a year ago, an online friend's cat ran away. He was inconsolable, since he grew up with the cat and it was basically one of his best friends. I vaguely recalled him telling me a long time ago that the cat had run away for three days once, and had gotten stuck in a local barbecue restaurant while it was closed for the weekend. *I didn't mention it, because I figured it would make my friend more upset.* Three days later, the cat comes back, smelling like barbecue sauce because it got stuck in the restaurant over the weekend. My friend swears the cat had never run away before, and I could find no record of the supposed "cat ran away" conversation in any of our chat logs. He never even mentioned there was a barbecue restaurant near his house.

TextRank Summary: About a year ago, an online friend's cat ran away. *He was inconsolable, since he grew up with the cat and it was basically one of his best friends.* I vaguely recalled him telling me a long time ago that the cat had run away for three days once, and had gotten stuck in a local barbecue restaurant while it was closed for the weekend. I didn't mention it, because I figured it would make my friend more upset. Three days later, the cat comes back, smelling like barbecue sauce because it got stuck in the restaurant over the weekend. My friend swears the cat had never run away before, and I could find no record of the supposed "cat ran away" conversation in any of our chat logs. He never even mentioned there was a barbecue restaurant near his house.

Figure 4.1: Extractive summaries: the whole narrative is shown, with extracted sentences in blue or red; the *italicized* sentences are false positives.

Figure 4.1 shows an extractive summary produced by our self-trained model; this summary has perfect recall and one false positive sentence. On this particular example, TextRank also performs very well; it generates a similar summary with a different false positive sentence.

4.4 Discussion

While the change-based model works well for both extractive summarization and MRE sentence detection, the two tasks focus on different subsets of the features. We compare

the top 10 features for our quality-controlled self-training experiment (obtained from logistic regression due to our nonlinear SVM kernel) to those in our previous self-training experiment in Table 4.4. Some of the features are similar, such as a sharp increase in imagery compared with preceding sentences. For several metrics – activeness, verb phrase parse tree depth (a measure of sentence complexity), and semantic similarity to adjacent sentences – one of the two tasks relied more on a sentence's distance from the global minimum for that metric, while the other relied on the metric score itself; both features targeted sentences with relatively high scores on that metric.

Formality features, such as low average word formality and proximity to the document's global minimum in formality, perform well in MRE sentence detection but not in extractive summarization; for extractive summarization, but not MRE sentence detection, semantic similarity features for the first two sentences in a narrative perform well ("lssimilarity2" and "cossimilarity1"), reflecting the inclusion of background information, which usually appears near the beginning of a narrative.

Examining false positive sentences, we find mostly sentences with high imagery and activeness, and low semantic similarity to neighboring sentences, similar to true positives. However, false positives tend to be short and syntactically simple. Because the Turkers who constructed our reference summaries were instructed to select as few sentences as possible, they likely chose longer sentences that contained as many pieces of information as possible. In contrast, our model predicts whether or not a given sentence should be extracted in isolation, allowing it to choose several short, one-factoid sentences where a human would choose a single long, multiple-factoid sentence (as evidenced by the strength of the sentence length feature in extractive summarization).

	MRE Sentence Detection			Extractive Summarization	
	Feature Name	Weight		Feature Name	Weight
1.	incomingd2_imagery	4.174	1.	distfrommin_imagery	4.543
2.	distfrommin_wordformality_neg	4.109	2.	distfrommin_sentdepth	3.885
3.	cossimilarity_adjacent	3.618	3.	sentlength	3.376
4.	distfrommin_activeness	3.377	4.	incomingd2_imagery	3.283
5.	sentdepth	3.364	5.	distfrommin_vplength	3.262
6.	distfrommin_wordlength_neg	3.321	6.	activeness	3.220
7.	distfrommin_vpdepth	3.034	7.	lssimilarity2	3.186
8.	distfrommin_imagery_neg	2.790	8.	distfrommin_lssimilarity_adjacent	3.142
9.	wordformality_neg	2.329	9.	cossimilarity1	2.922
10.	incomingd2_vplen	2.128	10.	vpdepth	2.964

Table 4.4: Comparison of the top 10 features for MRE sentence detection and extractive summarization.

We also find that false positives tended to score higher on subjectivity than neighboring sentences. The predicted summary in Figure 4.1 demonstrates this: the false positive sentence describes the narrator's concerns about his friend's feelings. Human summarizers trying to limit summary length are likely to omit such sentences, as they are less important to a reader's understanding of the story.

4.5 Conclusion

The main contributions of this chapter are as follows:

- We explore different methods for combining multiple human-constructed extractive summaries into a single reference summary, including aggregating individual workers' summaries, removing or retaining information present in one summary but not another, and using a weighting system based on the pyramid method to account for the relative importance of sentences extracted by different workers.

- We apply the change-based model of narrative to the task of extractive summarization and show that, despite significant differences between this task and that of detecting *most reportable event* sentences, for which it was originally designed, the change-based model performs very well and outperforms extractive summarization approaches not specifically designed for the personal narrative genre.

It is interesting to note that while the precision of our extractive summarizer is lower than its recall, this is not nessarily an undesirable outcome; the reverse, high precision and low recall, would be far worse. The extractive summaries produced by this system are, in

the next chapters, used as inputs to a paraphrasing system that rewrites them into abstractive summaries. Any information not preserved in the extractive summary is lost forever; the rewriting system cannot recover it. Thus, while there is some room for improvement in our change-based extractive summarization system, its performance is satisfactory – and better than that of any other extractive summarizer.

In the next chapters, we move on to the second stage of our two-stage summarization system: editing and rewriting. We focused on *lexical paraphrasing*, one of Jing and McKeown (1999)'s six rewriting operations; we develop an abstractive system that learns not only how to perform paraphrasing for summarization, but also when it is appropriate to paraphrase.

5. Sentential Paraphrase Alignment

Before we can address the task of *lexical paraphrasing* for summarization, we must wrestle once again with the problem of data. As discussed in Chapter 3, our Reddit summarization corpus contains only 1218 examples of lexical paraphrasing – nowhere near enough to train the neural sequence-to-sequence model we plan to use for the paraphrasing task. We need more data, but hiring human workers to create, align, and annotate more summaries would be time-consuming and expensive.

Human annotators would not be necessary if we had some way to automatically detect paraphrasing in abstractive summary sentences. Our goal is to learn to rewrite extractive summary sentences into abstractive summary sentences using lexical paraphrasing, but we do not need full extractive summaries to do so; we are only interested in those extractive summary sentences that demonstrate lexical paraphrasing. If we could automatically determine, for a given document and abstractive summary, whether or not each sentence in the summary was a paraphrase of some sentence in the document, then we could use those sentences as additional examples of the lexical paraphrasing operation.

However, the type of paraphrasing that we would like to detect is very different from the strict, word-matching style of paraphrasing for which existing alignment systems are designed. Paraphrases between a document/extractive summary sentence and an abstractive summary sentence tend to be worded very differently from each other, and they tend to be much longer than the one- or two-word paraphrases for which existing systems are designed. Further, a summary, by definition, is not a perfect reproduction of all information

present in the document, but rather contains only the most important information from the document. As a result, in a paraphrase between a document/extractive summary sentence and an abstractive summary sentence, there is likely to be a mismatch in the amount of information present: both sentences may express roughly the same ideas, but the abstractive summary sentence is likely to be more general and less detailed.

In this chapter, we take a brief detour to explore the task of *sentential paraphrase alignment*, designing a system that aligns paraphrases of arbitrary length and of approximately equivalent meaning that fills this gap in existing paraphrase alignment research.

5.1 Motivation

Monolingual paraphrase alignment (as opposed to alignment for machine translation) has applications in many natural language processing tasks, such as text-to-text generation (Barzilay and Elhadad 2003; Barzilay and McKeown 2005), natural language inference (MacCartney, Galley, and Manning 2008), and recognizing textual similarity (Sultan, Bethard, and Sumner 2014b). In our case, we use it to collect additional paraphrase pairs to train our paraphrasing system.

Madnani and Dorr (2010) define three levels of paraphrasing:

- *Lexical*, where individual words are replaced by synonyms or hypernyms.
- *Phrasal*, involving equivalent idiomatic phrases, such as verb-preposition combinations (eg. “take over” or “assume control of”), or syntactic transformations, such as active versus passive voice.

- *Sentential*, which can be trivially achieved by performing lexical and phrasal paraphrasing on parts of a sentence, but Madnani and Dorr note that more interesting paraphrases, such as "He needed to make a quick decision in that situation" and "The scenario required him to make a split-second judgment," are challenging.

It is important to note here that the term *lexical paraphrasing* is overloaded. Madnani and Dorr use it to refer specifically to word-level paraphrasing, while Jing and McKeown (1999) use it to refer to paraphrasing in general – they give as examples both a word-level paraphrase ("point out" and "note") and a phrase-level paraphrase ("fits squarely into" and "hits the head on the nail") when discussing lexical paraphrasing for summarization. In this thesis, we follow Jing and McKeown in using the term *lexical paraphrasing* to refer to paraphrasing in general; we use *word-level paraphrasing* to refer to Madnani and Dorr's definition.

We were confronted with two challenges when we first attempted to use existing monolingual alignment systems to find paraphrases in document-summary pairs. First, past work on alignment had focused on word-level and short phrase-level alignments, to the exclusion of longer phrase-level or sentence-level alignments. In fact, alignment systems designed for phrase-level alignments were practically limited in the lengths of the phrases they could align – the decoding time complexity of Yao et al. (2013b)'s state-of-the-art aligner, JacanaAlign, the fastest existing phrase-based aligner, grows linearly in the length of the source phrase and quadratically in the length of the target phrase. These limitations are likely due in part to the lack of longer alignments in existing paraphrase corpora: Yao et al. report that 95% of alignments in the Microsoft Research Recognizing Textual Entail-

ment (MSR RTE) (Brockett 2007) and Edinburgh++ (Cohn, Callison-Burch, and Lapata 2008; Thadani, Martin, and White 2012) paraphrase corpora are single-token, word-level paraphrases, and phrases of four or more words make up less than 1% of MSR RTE and 3% of Edinburgh++ – there was no need for alignment systems to handle longer phrases.

Pat **ate**. ⇔ Pat **did not starve**.
Pat **teaches** Chris. ⇔ Chris is Pat's **student**.
The Marines are **fighting** the terrorists. ⇔ The Marines are **eliminating** the terrorists.
The **government** declared victory in Iraq. ⇔ **Bush** declared victory in Iraq.

Figure 5.1: Examples of quasi-paraphrases from Bhagat and Hovy (2013).

Second, past work had avoided what Bhagat and Hovy (2013) call *quasi-paraphrases* (Figure 5.1). Bhagat and Hovy argue that linguists generally accept paraphrase pairs with approximate, rather than exact, semantic equivalence. *Quasi-paraphrases*, "sentences or phrases that convey approximately the same meaning using different words," include pairs where there are pragmatic differences between the two text spans, where one implies the other, or where real-world knowledge is required to understand that the pair are synonymous. As was the case with length, the lack of existing work on quasi-paraphrases is likely due to the type of data available; while both the MSR RTE and Edinburgh++ paraphrase corpora contain quasi-paraphrases marked as *possible* alignments, most previous work chose to ignore them in favor of *sure* alignments, which hold between words that are identical or close to identical (we show examples in the next section).

The paraphrases in our Reddit summarization corpus, however, are long quasi-paraphrses (Figure 5.2). The Amazon Mechanical Turk workers who created them had been instructed to align "phrases from the [abstractive] summary with phrases from the [extractive summary] that effectively mean the same things." The word "effectively" in the instructions

Extractive: Very rarely do I get a “thanks” or a smile of appreciation.
Abstractive: I never get any thanks.

Extractive: I had a sleep paralysis dream that I was abducted by aliens.
Abstractive: I had the alien abduction dream.

Figure 5.2: Examples of long quasi-paraphrases from the Reddit summarization corpus.

allowed workers to align *quasi-paraphrases* rather than strict paraphrases, and because one sentence is roughly a summary of the other, some amount of compression and generalization between the two sentences is to be expected. The workers were free to align phrases of any length, including the full sentences shown above; just over 99% of the paraphrases in the Reddit summarization corpus involve phrases of four or more words, and the average length of aligned phrases is 11 for abstractive summary sentences and 25 for extractive summary sentences. Thus, to align the type of paraphrases found in the Reddit summarization corpus, we need to design our own alignment system to address the two challenges of aligning phrases of arbitrary length and aligning based on approximate, rather than exact, semantic equivalence.

5.2 Related Work

Because the design of alignment systems was strongly influenced by the data that was available during their development, in this section, we first give an overview of important paraphrase corpora and lexicons, and then discuss the alignment systems that were designed using those resources.

5.2.1 Paraphrase and Alignment Corpora

5.2.1.1 The Microsoft Research Recognizing Textual Entailment (MSR RTE) Corpus

The development of monolingual alignment as an independently evaluated natural language processing task began with the release of the MSR RTE corpus (Brockett 2007), which consists of 1600 sentence pairs, divided evenly into training and testing sets, annotated with alignments. Before it was annotated for alignment, the data had been used in the 2006 Pascal Recognizing Textual Entailment Challenge (Bar-Haim et al. 2006); the sentence pairs are thus premise-hypothesis pairs, where the premise sentence either does or does not entail the hypothesis sentence.

He met U.S. President, George W. Bush, in **Washington** and British Prime Minister, Tony Blair, in **London.** **Washington** is part of **London.**

(a) *Sure* alignments in a sentence pair where the first does not entail the second.

Although they were born on different planets, Oscar-winning actor **Nicolas Cage's** new **son** and Superman have something in common – both **were named Kal-el.** **Nicolas Cage's son is called Kal-el.**

(b) *Sure* alignments in a sentence pair where the first does entail the second.

Tilda Swinton *has a prominent role as* **the White Witch** in "The Chronicles of Narnia: The Lion, The Witch and The Wardrobe", coming out in December. **Tilda Swinton** *plays the part of* **the White Witch.**

(c) Both *sure* (shown in bold) and *possible* (shown in italics) alignments in a sentence pair.

Figure 5.3: Examples of sentence pairs from the MSR RTE corpus, with alignments shown in bold.

In the case where the premise does not entail the hypothesis, the two sentences do not mean the same thing (Figure 5.3a), and so the alignments on these pairs are simply between shared words. In the case where the premise does entail the hypothesis, the alignments may

still be between identical or nearly identical words (Figure 5.3b), or occasionally, there may be alignments marked as *possible*, rather than *sure* (Figure 5.3c) – these are the quasi-paraphrases of the MSR RTE corpus. Aligners designed for use on MSR RTE, which we review later in this section, are trained and evaluated on *sure* alignments only, effectively omitting the quasi-paraphrases in the corpus.

5.2.1.2 The Edinburgh++ Corpus

The original Edinburgh corpus (Cohn, Callison-Burch, and Lapata 2008) consisted of sentence pairs from three sources:

- The Microsoft Research Paraphrase corpus. This is not to be confused with the MSR RTE corpus; it consists of 5800 sentence pairs that were automatically extracted from news websites and annotated by human judges as either semantically equivalent or not. Cohn et al. removed pairs where the longer sentence was more than eight times as long as the shorter sentence but otherwise did not modify this sub-corpus.

- The Multiple-Translation Chinese (MTC) corpus, consisting of 105 Mandarin Chinese news articles, each of which was translated into English by eleven different translators. Cohn et al. manually selected four of these translators who produced the most fluent English translations and used different translations of the same sentence as paraphrase pairs.

- The Leagues corpus, consisting of two different English translations of Jules Verne's *Twenty Thousand Leagues Under the Sea*, with paragraph-level alignments. Cohn et

al. filtered these alignments, keeping only those that were single sentence to single sentence.

From each of these three sources, Cohn et al. randomly selected 300 sentence pairs (the corpus totaled 900 sentence pairs). They first aligned them automatically using Giza++ (Och and Ney 2003) and then used two trained annotators to correct the automatic alignments. As was the case with the MSR RTE corpus, the Edinburgh corpus contains both *sure* and *possible* alignments; Cohn et al. explained that "...sure alignments are clear-cut decisions and typical of genuinely substitutable words or phrases, whereas possible alignments flag a correspondence that has slightly divergent syntax or semantics."

"the *woman* **was taken to** New **Charing Cross Hospital** by ambulance and her *condition* is *critical*. *she* **was taken to Charing Cross Hospital**, where *she remained critically ill* last night.

(a) A sentence pair from the MSR Paraphrase subset of the corpus.

"we do not intend to stay until *the* **term** *is over*. **"we do not plan to stay** in office **up till** the *expiry* of *our* **term**.

(b) A sentence pair from the MTC subset of the corpus.

"are these *lenses* **capable of resisting such pressure?"** *"is glass* **capable of resisting such pressures?"**

(c) A sentence pair from the Leagues subset of the corpus.

Figure 5.4: Examples of sentence pairs from the Edinburgh++ corpus, with *sure* alignments shown in bold and *possible* alignments shown in italics.

Thadani, Martin, and White (2012) built upon the Edinburgh corpus by using human annotators to correct the automatic word tokenization, as well as performing truecasing and normalizing quotation marks. They also added dependency parses and named entity tags for each sentence. This enhanced version of the corpus is known as Edinburgh++. Figure 5.4 shows examples of alignments from each of the three subsets of the Edinburgh++

corpus. As we saw in the MSR RTE alignments, *sure* alignments are word-for-word identical or close to identical, while *possible* alignments capture *quasi-paraphrases*.

An important distinction between Edinburgh++ and MSR RTE, however, is that the two sentences in an Edinburgh++ pair are very similar to each other, while the sentences in MSR RTE pairs are less so. This is due to differences in how the two corpora were constructed: the sentence pairs in Edinburgh++ were drawn from monolingual parallel text, and so the sentences in a pair must have the same meaning; the pairs in MSR RTE were designed for the entailment task, and so the sentences in a pair almost never mean exactly the same thing – even in entailed pairs, the hypothesis sentence covers only a subset of the information in the premise sentence. Of the two corpora, MSR RTE is the larger and more commonly used.

5.2.1.3 Paraphrase Lexicons

There has also been significant work on building paraphrase lexicons, which can be used by alignment systems to look up potential paraphrase pairs. Rather than collecting sentences that contain paraphrases, as in the case of the two corpora described above, paraphrase lexicons collect pairs of synonymous words, phrases, or syntactic structures, without the surrounding context. Early work on creating paraphrase lexicons used human annotations or thesauruses to identify interchangeable synonyms, while the first corpus-based approaches were limited to certain types of paraphrases, such as morphological or syntactic word variants (Jacquemin, Klavans, and Tzoukermann 1997) or adjective-noun phrases (Lapata 2001).

The first general-purpose, corpus-based paraphrase lexicon was the work of Barzilay

undertone ⇔ low voice sudden appearance ⇔ apparition

Figure 5.5: Examples of paraphrases from Barzilay and McKeown (2001).

and McKeown (2001). They used monolingual parallel text, as did the Edinburgh++ corpus: eleven different English translations of five different non-English novels. They predicted whether or not a given pair of words or phrases were a paraphrase pair using cotraining between one classifier that looked at the words or phrases themselves and another that looked at the surrounding contexts. Figure 5.5 shows examples from Barzilay and McKeown's lexicon of 9483 paraphrase pairs. Other lexicons constructed using monolingual parallel text include Barzilay and Lee (2003), Pang, Knight, and Marcu (2003), Ibrahim, Katz, and Lin (2003), Dolan, Quirk, and Brockett (2004), Dolan and Brockett (2005), and Lan et al. (2017).

thrown into jail ⇔ imprisoned especially concerned about the situation of ⇔ particularly concerned at the situation of

Figure 5.6: Examples of paraphrases from the PPDB.

The other corpus-based technique for constructing paraphrase lexicons is *bilingual pivoting*. Because monolingual parallel texts are relatively rare, this approach instead uses bilingual parallel texts, which come readily available – and already aligned – in the form of machine translation corpora. The bilingual pivoting approach was introduced by Bannard and Callison-Burch (2005), who observed that, if two different English phrases are aligned to (two different occurrences of) the same non-English phrase in a machine translation corpus, then the two English phrases are likely to be a paraphrase pair. The most well-known and widely used lexicon constructed using bilingual pivoting is the Paraphrase Database

(PPDB) (Ganitkevitch, Van Durme, and Callison-Burch 2013; Pavlick et al. 2015); Figure 5.6 shows examples of paraphrase pairs from this lexicon.

How often do earthquakes occur? ⇔ How frequently are earthquakes happening? This myth involves three misconceptions. ⇔ There are three misconceptions in this mythology.

Figure 5.7: Examples of paraphrases from ParaBank.

The most recent approach to paraphrase lexicon construction uses neural machine translation to generate, rather than discover, paraphrase pairs. Instead of aligning phrases in bilingual parallel texts, this approach instead translates the non-English text into English and treats an original English sentence and its translated English partner as a paraphrase pair. Work that uses this approach includes Mallinson, Sennrich, and Lapata (2017), Wieting, Mallinson, and Gimpel (2017), Wieting and Gimpel (2018), and Hu et al. (2019); Figure 5.7 shows examples of paraphrase pairs from ParaBank (Hu et al. 2019). One of the stated goals of the neural machine translation approach is to generate *sentential* paraphrases, the long paraphrases that were lacking in previous work; however, the paraphrases that current systems are able to generate consist of multiple lexical or short phrasal paraphrases put together, not the more interesting and challenging whole-sentence paraphrases envisioned by Madnani and Dorr (2010).

5.2.2 Monolingual Alignment Systems

To date, there are only five phrase-based monolingual aligners in existence, not including the work discussed in this chapter. The first was **MANLI** (MacCartney, Galley, and Manning 2008), which was developed for the MSR RTE corpus and which set a precedent for

following research: the *possible* alignments were not used. MacCartney et al. made this decision based on conclusions drawn in machine translation research that training using *possible* alignments did not improve the performance of machine translation systems, but as we show in Section 5.4, this decision, which has been followed by subsequent MSR RTE systems, removed from consideration nearly all of the long alignments (four or more words) in the corpus. MANLI was a phrase-based system, designed to be capable of aligning multiple source tokens to multiple target tokens. However, MacCartney et al. found that constraining it to align only at the word level (i.e., setting a maximum phrase length of one) decreased the system's F-measure by only 0.2%, suggesting that this early work was not yet able to represent the meanings of multi-word phrases as well as it could represent the meanings of single words.

Thadani and McKeown (2011) extended MANLI by introducing syntactic constraints on alignment, improving the system's precision, and used integer linear programming to perform faster, exact decoding, rather than the slower, approximate search used by the original system. **Thadani, Martin, and White (2012)** later added dependency arc edits to MANLI's phrase edits, again improving the system's performance. Interestingly, Thadani et al. used both the *sure* and *possible* alignments in the Edinburgh++ corpus and showed that training on both gave better performance than training only on *sure* alignments, but no subsequent monolingual alignment systems took advantage of *possible* alignments until we did so in the work described in this chapter.

The current state-of-the-art phrase-based monolingual alignment system is **JacanaAlign-phrase** (Yao et al. 2013b), the phrase-based extension of JacanaAlign-token (Yao et al. 2013a). Yao et al. used a semi-Markov conditional random field to tag each token or se-

quence of tokens in the source sentence with the indices of aligned target tokens. To train this system, they synthesized phrasal alignments by merging consecutive word-level alignments among the MSR RTE *sure* alignments; however, even after doing so, they found that long alignments involving phrases of four or more words still made up less than 1% of the corpus. Yao et al. found that the phrase-based JacanaAlign performed slightly worse than the token-based version, likely due to the overwhelming majority of alignments in their test set being at the token level and the token-based alignments in the test set penalizing their phrase-based output.

JacanaAlign-phrase is the fastest existing phrase-based aligner (there are only four others: MANLI, its two extensions, and SemAligner, all described in this section), but Yao et al. noted that it is roughtly 30-60 times slower than the word-based version, JacanaAlign-token. Of particular note is that the decoding time of JacanaAlign-phrase is $\mathcal{O}\left(L_s L_t^2 M N^2\right)$, where L_s and L_t are the maximum allowed phrase lengths, and M and N are the sentence lengths, for the source and target, respectively. The longer the phrases being aligned, the longer Jacana-Align would need to run.

Finally **SemAligner** (Maharjan et al. 2016) is a semantic aligner designed for the semantic textual similarity task. It chunks input sentences into phrases, aligns the phrase chunks, and labels each alignment with a semantic relation, such as equivalence or specification. However, it was designed for and evaluated on the semantic textual similarity task, so its published performance could not be directly compared with those of monolingual alignment systems.

5.3 Models

Our goal in designing our own alignment system is to overcome two challenges of monolingual paraphrase alignment that previous work on alignment did not address: aligning long phrases of four or more words and aligning phrases that were only approximately synonymous. We first experimented with a rule-based approach that used the output of both an existing aligner and a dependency parser; the approach began with aligned words in the input sentences' dependency parses and attempted to "grow" the alignments by traversing the edges of the dependency graphs. However, this first approach was unsuccessful. It was difficult to determine whether or not a given dependency graph edge should be used to grow an alignment; because the two input sentences did not necessarily mean exactly the same thing, there was not always a match between edges in their graphs. It was not clear which edge matches should be allowed, and it was further unclear how to handle alignments between phrases of different lengths, for which there would not be a one-to-one edge mapping.

Our solution and the key to our second, successful approach, is the observation that neither words nor phrases are necessarily the best unit of alignment; the best unit of alignment could be word, phrase, clause, or full sentence, depending on how similar in meaning the input sentences are. We need an alignment system that can not only align units of arbitrary size, but also discover the best unit size for each input sentence pair.

To accomplish these two goals, our system first chunks the source and target sentences of an input pair several times, at different levels of granularity, from mostly single words to phrases to whole clauses, then computes a chunk embedding in a distributed semantic

space for each chunk (Section 5.3.1). We call any segmentation of a sentence into chunks a *chunking* of that sentence; each chunking represents a different possible set of alignment units for that sentence. We then pair a source chunking with a target chunking and use a pointer-network (Vinyals, Fortunato, and Jaitly 2015) to perform a preliminary alignment from a source chunk to the target chunks (Section 5.3.2). Finally, we combine the preliminary alignments from all source/target chunking pairs using a voting system to produce the final alignment from the source sentence to the target sentence (Section 5.3.3).

5.3.1 Chunkings and Chunk Embeddings

Our system first chunks the source and target sentences using constituent parsing (Bauer 2014). We consider all nodes with phrase-level tags (XP) to be *constituents*. Beginning with the leaves, we move up the tree, deleting any node that is wholly contained in a larger constituent but that is neither a constituent itself, nor the sibling of a constituent. Figure 5.8 shows such a simplified constituent tree.

Constituents and their siblings are the smallest possible *chunks* that we consider. In the example constituent tree above, there are eight such small chunks. Our system can also merge any number of consecutive small chunks to form a larger chunk: "offered," "no dinner," "at," and "the reception," for instance, can be merged to form "offered no dinner at the reception." In a sentence with i of these smallest chunks, there are $i - 1$ potential chunk boundaries (Figure 5.9); since merging two adjacent chunks is equivalent to ignoring the chunk boundary between them, there are 2^{i-1} unique *chunkings* of the sentence. It is important to note that each token in the sentence is contained in exactly one chunk in each

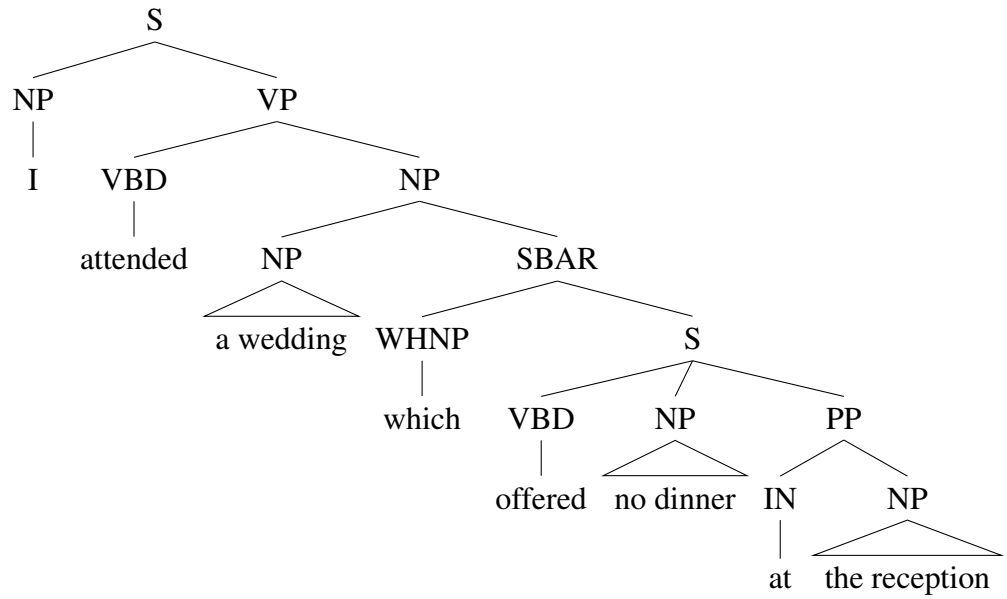

Figure 5.8: An example simplified constituent tree.

chunking of that sentence.

I	attended	a wedding	which	offered	no dinner	at	the reception

Figure 5.9: All potential chunk boundaries for the example sentence.

From the example sentence above, our system would obtain 128 unique chunkings. The coarsest consists of a single chunk containing the entire sentence, and the most fine-grained has each leaf of the constituent tree as a separate chunk. Each chunk is a possible unit of alignment, and so each chunking is a different possible way to break down an input sentence into units of alignment. However, at this point there is no way for our system to know which chunking is the "correct" one. Instead of arbitrarily choosing a single chunking to align, we perform the preliminary alignment described in the next section using all possible chunkings of the source and target sentences; the alignment scores from the preliminary alignments and the voting system that finally combines those scores will decide the "correct" units of alignment.

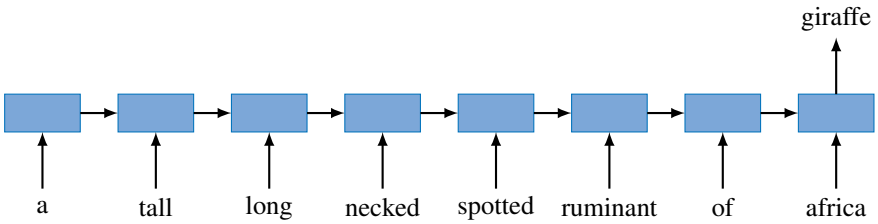

Figure 5.10: Hill et al.'s LSTM language model for sentence-level embeddings.

In order to align a chunk, rather than a word, we need some way to represent the meaning of the chunk as a whole. We look to recent work in composing word embeddings into phrase- or sentence-level embeddings. Since the early work of Mitchell and Lapata (2008), there has been a great deal of interest in learning phrase embeddings (Baroni and Zamparelli 2010; Zanzotto et al. 2010; Yessenalina and Cardie 2011; Socher et al. 2012; Grefenstette et al. 2013; Mikolov et al. 2013; Yu and Dredze 2015). In the work discussed in this chapter, we generate chunk embeddings using the LSTM language model of Hill et al. (2016)[1] . The model is trained on dictionaries: it takes as input a dictionary definition, in the form of a sequence of word embeddings, and produces as output the embedding of the word to which the definition belongs, thus learning to compose the embeddings of the words into a single embedding representing the entire phrase or sentence (Figure 5.10)[2]. By representing each chunk by a single chunk embedding, we are able to align chunks of arbitrarily large size with only the language model's run time as overhead.

[1]We also experimented with simply averaging word embeddings, but this approach underperformed the language model.

[2]A limitation of this approach – also a limitation of context-independent word embeddings in general – is that there is no accomodation for words with multiple senses.

5.3.2 Preliminary Alignment

Once all possible chunkings of the source and target sentences have been assembled and chunk embeddings calculated for each chunk, our system takes the Cartesian product of the set of source sentence chunkings and the set of target sentence chunkings, pairing each source chunking with each target chunking. Then, for a given source/target chunking pair, we perform a preliminary alignment from each source chunk in the source chunking to the entire target chunking using a neural network aligner inspired by the pointer network of Vinyals, Fortunato, and Jaitly (2015). For a source sentence with M different chunkings, each of length l_m where $m \in [1, M]$, and a target sentence with N different chunkings, we perform $\prod_{m \in [1,M]} l_m N$ different preliminary alignments. Thus, the multiple chunkings also have the practical benefit of increasing the amount of training data available for the aligner.

Most previous work on neural network alignment has used feed-forward, recurrent, or convolutional neural networks to score source-target word pairs and then fed these scores to a traditional alignment model, such as an HMM or a greedy aligner (Yang et al. 2013; Tamura, Watanabe, and Sumita 2014; Legrand, Auli, and Collobert 2016), rather than using the neural network itself to predict the alignments. This is likely due to the difficulty of adapting a neural network to the alignment task directly: two input sequences of unknown and often different lengths, as well as an output set of unknown size, with no clear linear relationship among output word pairs.

Our neural network aligner is based on the pointer network and learns a distribution over an output dictionary of variable size. The flexibility of the output size makes the

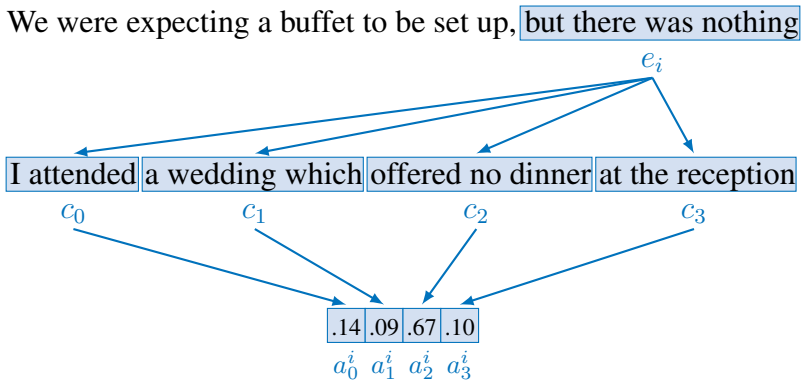

Figure 5.11: The pointer network aligner performing preliminary alignment between source chunk e_i and and target chunking $c_0c_1c_2c_3$.

pointer network well-suited to the task of aligning chunkings of variable length. For a given source chunk from the source chunking under consideration, we adapt the pointer network to predict a preliminary alignment over the entire target chunking:

$$a_j^i = v^T \tanh(W_1 e_i + W_2 c_j)$$

where e_i is the embedding for chunk i in the source chunking, c_j is the embedding for candidate chunk j in the target chunking, and v, W_1, and W_2 are learned parameters. (For convenience, in subsequent sections we use e_i and c_j to refer to both the chunk embeddings, which are vectors, and to the chunks themselves, which are sequences of tokens.) The chunk embeddings are generated by the LSTM language model described in the previous section, and are fixed at training time to allow for easier and faster batching. For each source chunk in the source chunking, the pointer network aligner produces a distribution over all candidate chunks in the target chunking. Figure 5.11 shows the pointer network aligner performing a preliminary alignment on a source/target chunking pair.

5.3.3 Voting and Final Alignment

For a fixed source chunking and a fixed source chunk, the pointer network aligner produces one preliminary alignment for each different chunking of the target sentence, and we perform a preliminary alignment for all source chunks in all chunkings of the source sentence. By performing preliminary alignments for all combinations of source and target chunkings, we are able to defer deciding the alignment unit size, instead allowing the voting procedure to discover it.

Because the standard evaluation metric for monolingual alignment is precision, recall, and F-measure for aligned token pairs – even for phrase-based models – the output of the voting procedure (i.e., the final output of our alignment system) is aligned token pairs. However, the longer aligned phrases that correspond to these aligned token pairs can be easily reconstructed: following MacCartney, Galley, and Manning (2008) and Yao et al. (2013b), two tokens are aligned if and only if the phrases containing them are aligned.

Token-level output is convenient for the voting procedure, which is described in pseudocode in Figure 5.12. Because the preliminary alignments are performed on chunkings of different granularities, we must vote at the level of the smallest possible chunks (the leaves in the constituent tree). Since it is not possible for the tokens within one of these smallest possible chunks to receive different amounts of votes (to do so would require the tokens to have been in two different chunks in some chunking) we simply vote on tokens. We convert the preliminary alignments to token-level alignments by assigning to each token pair the preliminary alignment score between the source and target chunks containing them.

Only one chunk e_{i_s} in a given source chunking s contains the token w, and only one

Inputs
- the source sentence W
- the target sentence U
- the set of source sentence chunkings S
- the set of target sentence chunkings T

Initialize
- set score$(w, u) = 0$ for tokens $w \in W$ and $u \in U$

Repeat for $e_i \in s$, for $(s, t) \in S \times T$
- predict preliminary alignment a^i
- add a^i_j to score(w, u) for tokens $w \in e_i$ and $u \in c_j$

Repeat for $w \in W$
- sum-to-one normalize score(w, u) for $u \in U$
- sort pairs (w, u) by score(w, u) in descending order: score$(w, u_1) > \ldots >$ score(w, u_m)
- select max k such that score$(w, u_k) > 1/(k + 1)$
- set $A_w = \{(w, u_1), \ldots, (w, u_k)\}$

Return $\bigcup_{w \in W} A_w$

Figure 5.12: Voting procedure for final output.

chunk c_{j_t} in a given target chunking t contains the token u. Here, i_s and j_t indicate the specific source and target chunks that contain the tokens w and u, respectively. The token-level scores are obtained by summing the preliminary alignment values for all source/target chunk pairs where the source chunk contains w and the target chunk contains u:

$$\text{score}(w, u) = \sum_{s \in S} \sum_{t \in T} a^{i_s}_{j_t}$$

where S is the set of all source chunkings of the source sentence, T is the set of all chunkings of the target sentence, and $a^{i_s}_{j_t}$ is the preliminary alignment value predicted by the pointer network aligner described in the previous section.

For a fixed source token w, we normalize its scores to produce a probability distribution over all target tokens. We select the k highest-scoring target tokens such that the score of each token is greater than $1/(k + 1)$. If we select four target tokens, for example, each has a score of at least 0.2, and the next-highest-scoring token has a score of less than 0.167.

Intuitively, we are looking for a large gap in the target token scores at which to cut off the selected tokens from the unselected tokens; the sum of the scores of all unselected tokens is less than the score of any selected token. We select the largest possible number of target tokens for which this requirement holds. This flexible threshold ensures that the selected tokens $u_1, \ldots u_k$ have much larger scores than the unselected tokens $u_{k+1}, \ldots, u_m$ while still allowing any number of tokens to be selected. The selected target tokens are then aligned to the source token w to produce aligned token pairs. The final output of our system is the union of the aligned token pairs for each source token in the source sentence.

5.4 Data

5.4.1 The Reddit Summarization Corpus

To train our model, we use the alignments in our Reddit summarization corpus; as we discussed in Section 5.1, our goal is to learn to align the longer, looser types of paraphrases likely to be found in human-written summaries. We randomly split the extractive/abstractive summary pairs in the corpus into 511 training, 108 validation, and 423 testing pairs, and within each subset further divide each summary pair into sentence pairs.

In order to train the pointer network aligner, we need target distributions for all of the preliminary alignments for source/target sentence pairs in the training and validation sets. While the Reddit summarization corpus already contained gold standard alignments between sentence pairs, those alignments had been produced by Amazon Mechanical Turk workers (Turkers), who had used a highlighting interface to align any substring from an

abstractive summary sentence to any substring from an extractive summary sentence – the Turkers were not constrained to obey the chunk boundaries that did constrain our preliminary alignments. In the example sentence shown in Figure 5.9, for instance, "dinner at the reception" would have been a perfectly valid span for the Turkers to align, but because there is no chunk boundary between "no" and "dinner," the pointer network aligner would never be able to predict that exact alignment; it would be forced to align either "no dinner at the reception" or "at the reception."

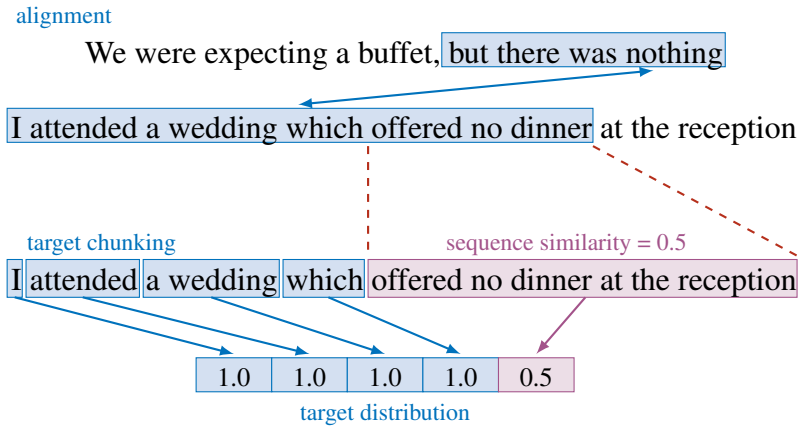

Figure 5.13: An training target distribution for preliminary alignment.

As a result, for a given source and target chunking, the training target distributions for preliminary alignment need to approximate as closely as possible the gold standard human alignment. To produce the training target distribution for the preliminary alignment between a given source sentence chunk and a given target sentence chunking, we first calculate, for each human alignment, the length of the longest common subsequence between the source chunk and the source alignment span, as well as the length of the longest common subsequences between each target chunk and the target alignment span, normalized by the lengths of the chunk and span. This normalized longest common subsequence score

represents our confidence that the chunk matches the alignment span, and we weight each target chunk's position in the training target distribution by its score (Figure 5.13). Finally, we weight the entire target training distribution by the source chunk's score. It is important to note that the target training distributions are not probability distributions; the maximum possible value at a position in the distribution is 1.0, and the absolute value of a position is less important than how it compares to the values of other positions in the distribution.

5.4.2 The MSR RTE Corpus

While we train solely on the Reddit summarization corpus, we use both the Reddit corpus and the MSR RTE corpus (described in detail in Section 5.2) for evaluation. The MSR RTE corpus is the most widely-used corpus for evaluating alignment systems (MacCartney, Galley, and Manning 2008; Thadani and McKeown 2011; Yao et al. 2013a; Yao et al. 2013b; Sultan, Bethard, and Sumner 2014a; Sultan, Bethard, and Sumner 2015), although it has previously been used for evaluation using *sure* alignments only[3].

In order to use this corpus to train their phrase-based system, Yao et al. (2013b) created longer alignments by merging consecutive, word-level *sure* alignments in the MSR RTE training set into larger, phrase-level alignments. They reported that doing so increased the percentage of multi-word alignments in the training set from 4% to 21%, but even after merging, alignments involving at least one phrase of four words or longer still made up less than 1% of the corpus.

If *possible* alignments are included, however, the percentage of alignments in the MSR

[3]Yao (2014) performed experiments using a different definition of "sure" and "possible": his "sure" alignments were those with perfect agreement among the MSR RTE annotators, and "possible" were those with some disagreement.

RTE training set involving phrases of four or more words increases to 27%, and if we restrict ourselves to sentence pairs that contain at least one *possible* alignment, that percentage increases to 61%. Thus, the set of MSR RTE sentence pairs that contain at least one *possible* alignment forms a subcorpus of precisely the type of long quasi-paraphrases that our system is designed to align. Unfortunately, the MSR RTE training set consists of only 800 sentence pairs, a very small amount of data for a neural network, and restricting the sentence pairs to those containing *possible* alignments reduces the amount of data even further, so we do not attempt to train our system on this subcorpus. Instead, we evaluate our system, as well as the state-of-the-art word- and phrase-based aligners, on the subset of 406 sentence pairs in the MSR RTE test set that contain *possible* alignments.

5.5 Experiments

Our chunk embedding model is implemented with Lasagne and trained for 25 epochs using the dictionary datasets and hyperparameter settings of Hill et al. (2016). Our pointer network aligner is implemented with PyTorch, using the pointer network settings of Vinyals, Fortunato, and Jaitly (2015) and cosine distance of the predicted preliminary alignment a^i from target preliminary alignment as the loss function; we train for 16 epochs using early stopping based on validation set performance.

We report the results of our experiments using the standard alignment evaluation metrics of precision, recall, and F-measure for aligned token pairs, where two tokens are considered aligned if and only the phrases containing them are aligned. As Yao et al. (2013b) argue, evaluating at the token level allows for alignment systems to receive partial credit

for phrases that are partially, but not fully, aligned correctly. We do not report the exact match percentage, a fourth metric commonly used in alignment evaluation, simply because that number was close to zero for all systems we tested – getting an exact match on a long alignment is extremely difficult.

5.5.1 Baselines

We compare our alignment system (hereafter *pointer-aligner* for simplicity) against three existing systems: Sultan, Bethard, and Sumner (2014a), a state-of-the-art word-level aligner[4]; JacanaAlign-phrase (Yao et al. 2013b), a state-of-the-art phrase-based aligner, and SemAligner (Maharjan et al. 2016). As discussed in Section 5.2, SemAligner has not previously been evaluated as a monolingual alignment system, as it was designed for textual similarity, but we include it as a baseline because its approach of aligning chunks is more similar to ours than the approaches of the two state-of-the-art systems. SemAligner assigns semantic relations to pairs of chunks, so in this evaluation, we treat chunk pairs assigned the *equivalent*, *specification*, and *related* relations as aligned and those assigned the *opposite* relation as not aligned. Because the evaluation is on phrase-level alignments, for fairness, we followed Yao et al. in converting word-level alignments into phrase-level ones by merging consecutive single-word alignments into larger phrase alignments.

We also evaluate a greedy baseline on the Reddit summarization corpus, which scores each candidate chunk in the target sentence based on the cosine similarity between its chunk embedding and that of the source chunk. We calculate the score using cosine distance as

[4]In private correspondence, Sultan recommended that we use this system, rather than Sultan, Bethard, and Sumner (2015), because it was "so much simpler with only a tiny difference in quality."

follows: let e and c be the chunk embeddings for the source and candidate target chunk, respectively;

$$\text{score} = 1 - \frac{ec}{\|e\|\|c\|} + 0.25m$$

where the constituent mismatch indicator m is a binary indicator that takes the value 0 if the source and candidate chunks are of the same constituent type, and 1 otherwise. This penalty encourages the greedy aligner to align constituents of the same type, but still allows, for example, a verb phrase to be aligned to its nominalized form. The mismatch penalty of 0.25 is tuned on the validation set.

The greedy baseline aligns each source chunk to the candidate target chunk with the lowest score (lower is better for cosine distance). If there are no target chunks with scores below a gap threshold of 0.6 (also tuned on the validation set), the source chunk remains unaligned. We convert the chunk-level alignments of the greedy baseline into token-pair alignments by considering two tokens to be aligned if and only if the chunks containing them are aligned. Finally, we take the union of all token alignments for all chunkings of the source and target sentences.

5.5.2 Reddit Summarization Corpus Evaluation

Table 5.1 shows the performance of the pointer-aligner on the Reddit summarization corpus test set, compared with the three other systems and greedy baseline. Our approach has an order of magnitude improvement in recall and F-measure over existing aligners. The greedy baseline also dramatically improves recall, demonstrating the importance of phrase-level similarity, but is significantly worse than the pointer-aligner that is key to success.

System	P%	R%	F_1%
Sultan et al.	**76.1**	1.4	2.8
SemAligner	65.7	2.5	5.4
JacanaAlign-phrase	59.5	3.9	7.3
greedy	51.4	27.5	35.8
pointer	54.3	**79.5**	**64.5**

Table 5.1: Performance on the Reddit summarization corpus test set.

Figure 5.14 shows an example of predicted alignments from the pointer-aligner; the greedy baseline; and the three baseline systems, SemAligner, Sultan et al., JacanaAlign-phrase. As expected, the state-of-the-art word-level aligner, Sultan et al., achieves the highest precision and aligns spans that are word-for-word identical (excepting minor grammatical differences) between the two sentences. Of the three baseline systems, JacanaAlign-phrase performs the best, although it does not outperform the greedy baseline; it produces one longer alignment, shown in green. The pointer-aligner and the greedy baseline align the longest spans, although they seem to have trouble with over-aligning and including extra words ("which I miraculously," shown in red in Figure 5.14b, and "and," shown in green in Figure 5.14c) while excluding others that should be aligned ("my boyfriend" in Figure 5.14b and "who I had been living with for two years" in Figure 5.14c).

One interesting note is that the greedy baseline aligns "and she repaid me by hooking up with my boyfriend" to "about a month later her and my boyfriend" and also separately aligns "hooking up" to "hooked up" (Figure 5.14c). While it is also possible for the pointer-aligner to align one source phrase to two non-consecutive target phrases, it does not do so in any of our experiments. Of the three other systems, only JacanaAlign-phrase is able to align a single source token or phrase to multiple target tokens, but those target tokens are

restricted to be a single consecutive phrase; SemAligner and Sultan et al. do not allow multiple alignments.

I saved my friend's life from a heroin overdose, and she repaid me by hooking up with my boyfriend.

I saved my (at the time) best friend from a heroin overdose by sticking suboxone under her tongue, which I miraculously had with me at the time. About a month later her and my boyfriend who I had been living with for two years hooked up.

(a) The Reddit summarization corpus's gold standard alignment.

I saved my friend's life from a heroin overdose, and she repaid me by hooking up with my boyfriend.

I saved my (at the time) best friend from a heroin overdose by sticking suboxone under her tongue, which I miraculously had with me at the time. About a month later her and my boyfriend who I had been living with for two years hooked up.

(b) The pointer-aligner's predicted alignment.

I saved my friend's life from a heroin overdose, and she repaid me by hooking up with my boyfriend.

I saved my (at the time) best friend from a heroin overdose by sticking suboxone under her tongue, which I miraculously had with me at the time. About a month later her and my boyfriend who I had been living with for two years hooked up.

(c) The greedy baseline's predicted alignment. The source phrase "hooking up" aligned to both the green and yellow target phrases.

I saved my friend's life from a heroin overdose, and she repaid me by hooking up with my boyfriend.

I saved my (at the time) best friend from a heroin overdose by sticking suboxone under her tongue, which I miraculously had with me at the time. About a month later her and my boyfriend who I had been living with for two years hooked up.

(d) SemAligner's predicted alignment.

I saved my friend's life from a heroin overdose, and she repaid me by hooking up with my boyfriend.

I saved my (at the time) best friend from a heroin overdose by sticking suboxone under her tongue, which I miraculously had with me at the time. About a month later her and my boyfriend who I had been living with for two years hooked up.

(e) JacanaAlign-phrase's predicted alignment.

I saved my friend's life from a heroin overdose, and she repaid me by hooking up with my boyfriend.

I saved my (at the time) best friend from a heroin overdose by sticking suboxone under her tongue, which I miraculously had with me at the time. About a month later her and my boyfriend who I had been living with for two years hooked up.

(f) Sultan, Bethard, and Sumner (2014a)'s predicted alignment. While this system aligns at the word level, consecutive alignments that we merged for evaluation are shown in the same color here.

Figure 5.14: Alignments on a sentence pair from the Reddit summarization corpus.

The pointer-aligner has difficulty with clean phrase boundaries, eg. omitting "my boyfriend" but including "which I miraculously" in Figure 5.14b. Because our system considers the alignment score of a token to be the sum of the preliminary alignment scores of the chunks that contain that token, it is possible for words within a constituent to receive different scores if there is a potential chunk boundary inside the constituent. In the first sentence of Figure 5.14, for example, there is a potential chunk boundary between "she repaid me by hooking up with" and "my boyfriend" (because "my boyfriend" is itself a constituent). Thus, there is a chunking where "my boyfriend" is its own, separate chunk, and in the preliminary alignment for that chunking, the pointer-network must have assigned "my boyfriend" a lower score than it did the rest of the chunks in that sentence. While other, coarser chunkings would have given "my boyfriend" some score, it was apparently not enough to make up the difference, and "my boyfriend" did not accumulate enough score to be included in the final alignment. The exclusion of "my boyfriend" is clearly an error on the part of our system; it may be worth considering to constraining the system not to break up certain types of constituents, such as prepositional phrases.

5.5.3 MSR RTE Evaluation

We evaluate on the MSR RTE corpus test set, using a majority vote among the three MSR RTE annotators: any alignments that at least two annotators marked as *sure* or *possible* are included. Because we are interested in evaluating the systems on long quasi-paraphrases, we remove from consideration the 394 sentence pairs in the MSR RTE test set that do not contain any *possible* alignments, leaving 406 sentence pairs. As discussed in Sec-

tion 5.4, Yao et al. (2013b) found that, even after merging consecutive single-word alignments, the *sure* alignments of the MSR RTE corpus consisted overwhelmingly of phrases fewer than four words in length. It is not until we add in the *possible* alignments that the percentage of four-word or longer phrases grows to 24% in the MSR RTE test set; when we look only at sentence pairs containing a least one *possible* alignment, the percentage of longer phrases grows to 44%. Thus, evaluating only on the 406 sentence pairs that contain at least one *possible* requires systems both to perform well on longer alignments, but also to avoid sacrificing performance on short alignments.

System	P%	R%	F_1%
Sultan et al.	6.7	3.4	4.4
SemAligner	4.1	6.8	5.1
JacanaAlign-phrase	5.2	6.7	5.8
pointer	**23.4**	**47.7**	**31.4**

Table 5.2: Performance on the restricted, 406-pair MSR RTE test set.

Figure 5.15 shows predicted alignments from the pointer-aligner and the three baseline systems. This particular sentence pair is very good for the pointer-aligner because the gold standard alignment, except for the word "50-50", is a single constituent neatly separated out from the rest of the sentence as a parenthetical. Of the three baseline systems, both SemAligner and Sultan et al. align more conservatively: the prepositions "of" and "with" for SemAligner (Figure 5.15c), and the semantically related words "business" and "production" for Sultan et al. (Figure 5.15e). JacanaAlign-phrase attempts to align longer spans, but its alignments "is a business" ⇔ "'s", shown in green, and "of" ⇔ "joint venture with", shown in yellow (Figure 5.15d), suffer from the same noisy, constituent-breaking boundaries as does the pointer-aligner on sentence pairs less well-suited to our approach.

Comparing the gold standard alignments of the MSR RTE corpus with those the Reddit summarization corpus, we find that it is often the case with the Reddit alignments that one sentence contains much more information than other. While some MSR RTE sentence pairs have this property (the modifier "prominent" in Figure 5.3c, for example), not all do. This is likely a side effect of the Reddit corpus's being designed for summarization – the sentence pairs are composed of an excerpt from a narrative and a human-written summary, which by definition compresses the content of the narrative. Further, the Reddit alignments were generated by Amazon Mechanical Turk workers, who were instructed to highlight aligned spans. In Figure 5.14a, the clause "who I had been living with for two years" should probably not have been aligned, but the workers may have found it bothersome to remove the clause (which would require splitting the alignment shown in green into two separate alignments), so the clause remains in gold standard alignments. Being trained on this data, the pointer-aligner seems to have learned this preference for retaining extra information contained within a larger, more strictly aligned span, such as the word "50-50" in Figure 5.15b. While it is possible for the pointer-aligner to to align a single source phrase to two non-consecutive target phrases, it did not encounter such examples in training and never does so in any of our experiments.

Botswana *is a* **business** *partner of* **De Beers.**

Production at mines operated by Debswana – **Botswana***'s* 50-50 *joint* **venture** *with* **De Beers** – reach 33 million carats**.**

(a) The MSR RTE corpus's gold standard annotation, with sure alignments in **bold** and possible alignments in *italics*.

Botswana is a business partner of De Beers.

Production at mines operated by Debswana – Botswana's 50-50 joint venture with De Beers – reach 33 million carats.

(b) The pointer-aligner's predicted alignment.

Botswana is a business partner of De Beers.

Production at mines operated by Debswana – Botswana's 50-50 joint venture with De Beers – reach 33 million carats.

(c) SemAligner's predicted alignment.

Botswana is a business partner of De Beers.

Production at mines operated by Debswana – Botswana's 50-50 joint venture with De Beers – reach 33 million carats.

(d) JacanaAlign-phrase's predicted alignment.

Botswana is a business partner of De Beers.

Production at mines operated by Debswana – Botswana's 50-50 joint venture with De Beers – reach 33 million carats.

(e) Sultan, Bethard, and Sumner (2014a)'s predicted alignment. While this system aligns at the word level, consecutive alignments that we merged for evaluation are shown in the same color here.

Figure 5.15: Alignments on a sentence pair from the MSR RTE test set.

5.6 Discussion

The main limitation of our approach is that it is still computationally expensive. Despite its independence from the length of the paraphrases it is to detect and align, the system is still very dependent on the lengths of the source and target sentence: we expand each pair of input sentences into multiple chunkings, and a preliminary alignment has to be run on each pairing of a source chunk and target chunking. The number of potential chunk boundaries in an input sentence varies roughly with sentence length: if the source sentence has length M, and the target sentence has length N, then there are roughly $M/2$ potential chunk boundaries in the source sentence and $N/2$ potential chunk boundaries in the target sentence. There are 2^{M-1} different chunkings of the source sentence and 2^{N-1} different chunkings of the target sentence, so the complexity of our system was thus

$$\mathcal{O}\left((M/2+1)\,2^{M-1}2^{N-1}\right) = \mathcal{O}\left(M2^{M+N-3}\right)$$

As a result, our approach in its original form is not an improvement in time complexity over the $\mathcal{O}\left(L_s L_t^2 M N^2\right)$ of Yao et al. (2013b). Unlike Yao et al., however, our system's complexity can easily be reduced. In the original system, there is a great deal of redundancy among chunkings. Each chunking of a given sentence is identical to one other chunking of that sentence except for a single merge/no merge decision at a single potential chunk boundary; thus, the preliminary alignments for those two chunkings are nearly identical and thus redundant.

To extract paraphrase pairs for training the lexical paraphrasing system that we describe

in the next chapter, we use a slightly worse-performing but much faster version of this alignment system. The faster system fixes a constant number of chunkings to align for every input sentence pair. Each sentence had only three different chunkings:

- The most fine-grained chunking i.e., the leaves of the constituent tree;
- The second coarsest chunking, i.e., the children of the root of the constituent tree (the coarsest chunking would simply be the full sentence, which was rarely the best unit of alignment);
- A chunking of intermediate granularity such that the average size of its chunks was halfway between the average chunk sizes of the most fine-grained and second coarsest chunkings.

This faster system sacrifices some flexibility in its ability to discover the best units of alignment for its input sentences but drastically reduces its time complexity to the much more manageable $\mathcal{O}(M)$.

5.7 Conclusion

The main contributions described in this chapter (based on work published in Ouyang and McKeown (2019)) are as follows:

- We design a monolingual alignment system that is capable of aligning paraphrases of arbitrary length, from single words to full sentences. By composing the semantics of the words in a text span into a single embedding representing the meaning of the entire span, our system is able to align longer spans as if they were single words.

Unlike previous work on phrase-based alignment, the time complexity of our system is independent of the length of the paraphrase spans.

- Our experiments on aligning long quasi-paraphrases in both the Reddit summarization corpus and in the MSR RTE corpus show that our system achieves significant increases in recall (over 75 points), while still maintaining a strong lead in F-measure on aligning long paraphrases of four or more words, compared with existing state-of-the-art word- and phrase-based aligners.

- Our pointer-network-based alignment system is the first use of a neural network to directly perform the alignment task, rather than using the neural network to generate features for a more traditional alignment algorithm.

6. Lexical Paraphrasing

With our completed paraphrase alignment system in hand, we now move on to the second stage of our two-stage narrative summarizer: an editing and rewriting system to transform the extractive summaries described in Chapter 4 into fluent and coherent abstractive summaries.

In their analysis of human-written summaries, Jing and McKeown (1999) identify six major sentence-level operations that expert human summarizers use to rewrite segments of text from the original document into more summary-appropriate forms: reduction, combination, syntactic transformation, lexical paraphrasing, generalization, and specification. Of these six, reduction and combination (more commonly called *compression* and *fusion*, respectively) have become well-established research areas of their own (Jing (2000), Knight and Marcu (2000), and Berg-Kirkpatrick, Gillick, and Klein (2011); and Barzilay, McKeown, and Elhadad (1999), Daumé III and Marcu (2004), Barzilay and McKeown (2005), and Filippova and Strube (2008), to name just a few).

We choose to focus on *lexical paraphrasing*; while there has been work on both generic paraphrase generation and paraphrase generation for tasks such as question-answering and machine translation evaluation, paraphrasing for summarization has not been explored. We hope to answer two key questions:

- Every sentence can be paraphrased somehow, simply by substituting a synonym for some word in the sentence, but not all sentences should be paraphrased – anyone

who has read the writing of a child who has just discovered the thesaurus knows this! When do human summarizers find it appropriate to paraphrase? Can we learn to distinguish sentences that human summarizers would paraphrase from those they would not?

- Can we learn to perform the type of paraphrasing found in the Reddit summarization corpus? As we discussed in the previous chapter, these paraphrases are long, often full sentences, and as we showed in Chapter 3, they are often mixed in with other rewriting operations in the same sentence. Unlike most prior work on paraphrase generation, we do not want to generate just any paraphrase; we want to generate the type of paraphrases that human summarizers generate – can it be done?

6.1 Related Work

In the previous chapter, we described several approaches for identifying paraphrases. In this section, we first focus on approaches for generating sentence-level paraphrases, and then discuss work on paraphrases, whether identifying or generating, as related to the task of summarization.

6.1.1 Paraphrasing Approaches

Early work on sentence-level paraphrasing used paraphrase pairs or patterns extracted from thesauruses or parallel corpora – in effect, paraphrase lexicons. If an input sentence contained one half of a paraphrase pair or matched a paraphrase pattern, the other half of the pair would be substituted in, or the pattern would be applied, to create a paraphrased output

sentence (Barzilay and Lee 2003; Pang, Knight, and Marcu 2003; Kauchak and Barzilay 2006; Zhao et al. 2008).

Techniques used to create paraphrase lexicons could also be used to generate new paraphrases directly for some downstream task. Duboue and Chu-Carroll (2006), for example, generated paraphrases of question-answering queries by translating them into some other language and back. As mentioned in the previous chapter, there are also several paraphrase lexicons created using machine translation: Mallinson, Sennrich, and Lapata (2017), Wieting, Mallinson, and Gimpel (2017), Wieting and Gimpel (2018), and Hu et al. (2019). While other lexicon-creation techniques are examples of paraphrase identification (because they extract paraphrases found "in the wild" in parallel corpora), the machine translation approach generates sentence-level paraphrases directly.

The work of Hu et al., in particular, addressed a major limitation of previous approaches: coverage. In the case of substitution-based approaches, not every input sentence can be paraphrased; only sentences containing one half of a known paraphrase pair, or matching a known pattern, can be paraphrased; in the case of a naive machine translation approach, there is no way of ensuring that a given input sentence will be modified during the translation process – it could very well be translated back to English exactly the same.

Paraphrase: It didn't mean anything, okay? ⇔ It was nothing, all right?
Constraints: disallow "okay," "mean," "means," "meaning," "meant"

Figure 6.1: A generated paraphrase from Hu et al. (2019).

While the earlier work of Duboue and Chu-Carroll (2006) addressed this limitation to some extent by using different machine translation systems and different languages to generate multiple candidate paraphrases, Hu et al.'s approach was more deliberate. They used

lexically-constrained decoding (Hokamp and Liu 2017), modifying the beam search procedure of their machine translation system to require that certain words be either included or excluded in the output. These constraints allowed Hu et al. both to perform lexicon-style paraphrasing (by requiring the inclusion of a word that was a known paraphrase of a word in the input) and also to force the system to discover new paraphrases for itself – by disallowing a given word in the input sentence but not providing the system with a replacement, they forced it to find a suitable paraphrase on its own. As a result, Hu et al.'s approach was capable of generating paraphrases for any input sentence. It is interesting to note that, while the constraints were single words, multiple constraints used together could produce surprisingly sophisticated paraphrases, such as that shown in Figure 6.1.

6.1.2 Paraphrasing for Summarization

Prior work on paraphrasing for summarization has mostly focused on identifying, rather than generating, paraphrases.

Identifying paraphrases is part of the *fusion* operation (McKeown et al. 1999; Barzilay, McKeown, and Elhadad 1999; Barzilay 2003). In *multi-document summarization*, the task is to summarize a cluster of documents all about the same topic or event. In such a situation, there are likely to be many sentences all roughly saying the same thing, and so one approach to multi-document summarization is to find groups of sentences that are similar to each other. However, sentences within a group will still differ in content, perhaps focusing on different aspects of the common topic or event (we saw an example of this in Chapter 3). It is often desirable to fuse the sentences within a group into a single sentence containing the

most important information in that group, and doing this requires the fusion system to be able to identify shared information among sentences that may be worded very differently from each other. Thus, being able to determine whether two phrases that contain different words are paraphrases, as opposed to truly different phrases representing different pieces information, is key.

Zhou et al. (2006) applied paraphrase identification to the task of evaluating summarization systems. They argued that metrics that use n-gram overlap unfairly penalize summaries that paraphrase the reference summary, rather than match it word-for-word. They extracted paraphrase pairs using *bilingual pivoting*, which they used to perform a three-stage summary evaluation, first counting phrase-level paraphrase matches between the generated and reference summaries, then counting word-level paraphrase matches, and finally counting exact word matches.

Ganitkevic (2018) recently applied paraphrase generation to the *compression* operation. He used bilingual pivoting to identify a set of syntactic paraphrase rules in the form of a featurized synchronous context-free grammar (SCFG). This rule, for example, addresses the use of the subordinating conjunction "that" in relative clauses: NP $\rightarrow$ NP that VP | NP VP. Each rule was assigned a set of features, such as the frequency of the paraphrase pair in the corpus used to construct the rules, and paraphrasing was performed by decoding the SCFG to "translate" the input sentence into a paraphrased output sentence.

Ganitkevic suggested that the feature set could be tailored to different rewriting operations based on how the generated paraphrases would be used; he argued that any rewriting operation could be achieved using paraphrasing. Compression, for example, could be achieved by replacing words or phrases in the input sentence with shorter paraphrases. He

performed compression by using features such as the difference in length between the two halves of a paraphrase pair, and his human evaluations showed that he was indeed able to shorten sentences without losing their original meaning,

Ganitkevic's claim is technically true in the sense that the overall length of the sentence can be reduced using paraphrasing, and Cohn and Lapata (2008) had previously argued that word substitution is an important part of sentence compression. However, compression using nothing but paraphrasing is only useful when the input sentence does not contain any extraneous information that should be gotten rid of – paraphrasing can only replace words or phrases with other words or phrases meaning the same thing; it cannot remove them entirely. In the case of summarization, it is often desirable to remove some of the information in the input document, so Ganitkevic's approach is less suitable.

6.2 Approach

Our approach to lexical paraphrasing for summarization combines ideas from two of the works described in the previous section: Hu et al. (2019)'s neural approach and Ganitkevic (2018)'s feature-based approach.

Hu et al. showed that a neural machine translation system consisting of an LSTM encoder-decoder with attention was capable of generating lexically diverse, sentence-length paraphrases, and further, that this ability could be controlled. However, their form of control, lexical constraints, required them to decide ahead of time which words were to be paraphrased; Figure 6.2 shows how strongly the choice of constraint affected the paraphrases generated by their system.

Input Sentence: It didn't mean anything, okay?

Constraints: disallow "anything"
Paraphrase: It didn't mean a thing, okay?

Constraints: disallow "okay"
Paraphrase: It didn't mean anything, all right?

Constraints: disallow "anything", "okay"
Paraphrase: It meant nothing, all right?

Figure 6.2: The effect of constraints on the paraphrases of Hu et al. (2019).

Since Hu et al.'s goal was to generate multiple, diverse paraphrases of a single input sentence, they were able to choose their constraints using heuristics, such as requiring paraphrase matches from the Paraphrase Database (Ganitkevitch, Van Durme, and Callison-Burch 2013) or disallowing relatively rare words from the input sentence. In our case, we want to perform a very specific type of paraphrasing; we want to learn to paraphrase when and how human summarizers do. Thus, we turn to Ganitkevic (2018)'s use of tailored features to guide the type of paraphrasing to be performed.

Our base model is a pointer-generator (See, Liu, and Manning 2017); it consists of an LSTM encoder-decoder with attention, like Hu et al.'s translation model, with the addition of a pointer-network (Vinyals, Fortunato, and Jaitly 2015) to allow rare words, such as names, to be copied directly from the input sentence. This encoder-decoder-with-copy-attention model has been widely and successfully used in general-purpose abstractive summarization research (Gu et al. 2016; Chopra, Auli, and Rush 2016; See, Liu, and Manning 2017; Paulus, Xiong, and Socher 2018); it is appealing for our task because it includes a learned probability switch that determines whether it copies tokens from its input (which we would like it to do on non-paraphrasing input sentences) or generates new tokens from

the vocabulary (which we would like it to do on paraphrasing input sentences).

We initialize the embedding matrices of both the encoder and decoder LSTMs with the pretrained 300-dimensional GloVE Common Crawl word vectors (Pennington, Socher, and Manning 2014), augmented with seven additional features: word length; activation, evaluation, and imagery scores from the Dictionary of Affect in Language (Whissell 1989) (augmented with Wordnet (Miller 1995) for better word coverage), as well as the activation-evaluation norm for subjectivity (Agarwal, Biadsy, and McKeown 2009); and word-level formality and complexity scores from the lexicons of Pavlick and Nenkova (2015). In Chapters 2 and 4, we used these features to predict, at the sentence level, what content should be included in a personal narrative summary; now we use them again to paraphrase those sentences by capturing aspects of word choice that a human summarizer might consider, but that might not be captured by the pre-trained embeddings.

Our goal is to train this model to learn not only how to perform lexical paraphrasing for summarization, but also when it was appropriate to do so – this is in constrast to previous work on paraphrase generation, which took for granted that their input sentences ought to be paraphrased somehow. Thus, we train our model on a mix of two different types of data:

- Document-summary sentence pairs where the summary sentence is a paraphrase of the document sentence. These are the "positive" training examples. On such pairs, the model should learn how to rewrite the document sentence into its paraphrased summary sentence.
- Document sentences that are not paraphrased by any of the sentences in their summaries. These are the "negative" training examples. On such an input sentence,

the model should learn to do nothing and return the sentence with no modifications, behaving as a simple autoencoder.

In the next section, we describe in greater detail the data sources and preprocessing steps that we use to create this mixed-type training set for our lexical paraphraser.

6.3 Data

In addition to our own Reddit summarization corpus, we use three other corpora in our lexical paraphrasing experiments. One is a paraphrase alignment corpus, the Microsoft Research Recognizing Textual Entailment (MSR RTE) corpus (Brockett 2007), which we used in the previous chapter. The other two are summarization corpora, both of which we discussed in Chapter 3: the New York Times annotated corpus (Sandhaus 2008) and the Webis-TLDR-17 corpus (Völske et al. 2017).

A quick note of terminology: because the input to our model is an extractive summary sentence and the output is also a sentence, we refer to our data as consisting of *sentence pairs*. However, in the case where the input sentence should not be paraphrased (i.e., its corresponding abstractive summary sentence is not a paraphrase of it), the desired output sentence is identical to the input sentence. The model is meant to target lexical paraphrasing specifically, and so it should leave non-paraphrasing input sentences alone; for these "negative" examples, there are not two different sentences, but rather two copies of the same sentence. For consistency, we still refer to these examples as *sentence pairs*.

6.3.1 The Reddit Summarization Corpus

The Reddit summarization corpus, which was the focus of Chapter 3, contains 6173 alignments between 1088 pairs of corresponding extractive and abstractive summaries. Of these alignments, 1218 are annotated with the lexical paraphrasing operation. However, the number of sentence pairs that we can use is actually lower. Because each pair of extractive and abstractive summaries was aligned by three separate Amazon Mechanical Turk workers (Turkers), alignments for given summary pair often overlap with or contain each other. Thus, when we convert the alignments (many of which are between phrases, and not full sentences) to sentence pairs (by expanding each alignment to include all sentences that contained part of the alignment, shown in Figure 6.3), we are left with 3589 sentence pairs, of which 1047 demonstrate the lexical paraphrasing operation, and 2542 do not.

Alignment:
Extractive: from the looks of the video tried to slip my friend a couple of "somethings" into his drink!
Abstractive: Security tapes revealed that she attempted to drug him,

Sentence Pair:
Extractive: It turns out the VICTIM girl was the one that pulled out the drugs and from the looks of the video tried to slip my friend a couple of "somethings" into his drink!
Abstractive: Security tapes revealed that she attempted to drug him, and intentionally drugged herself.

Figure 6.3: An alignment and its corresponding sentence pair.

As we described in Chapter 3, alignments could be annotated with multiple rewriting operations – of the 1218 alignments with lexical paraphrasing, 687 also contain at least one other rewriting operation. The presence of other rewriting operations in a sentence makes it more difficult for our model to learn to isolate and target lexical paraphrasing,

so we weight each sentence pair based on the number of rewriting operations in it. Since most sentence pairs represent multiple overlapping alignments, we take the union of the rewriting operations for each alignment belonging to a given sentence pair.

If the sentence pair contains only lexical paraphrasing and no other operations, it receives a weight of 7, while a pair that contains n different operations (including lexical paraphrasing) receives weight $7 - n$; a sentence pair containing all six rewriting operations[1], for example, would have weight 1.

	Mass	
Sentence Pair	Unweighted	Weighted
Lexical Paraphrasing	1047	4348
Non-Lexical-Paraphasing	2542	5084

Table 6.1: Mass of lexical paraphasing and non-paraphrasing sentence pairs before and after weighting.

Using this weighting scheme, the total mass of sentence pairs demonstrating the lexical paraphrasing operation is 4348; to balance the mass of lexical paraphrasing pairs with that of non-paraphrasing pairs, we assigned the latter 2542 pairs a weight of 2. Table 6.1 shows the mass of lexical paraphasing and non-lexical paraphasing sentence pairs with and without weighting; from the right-most column, we see that assigning weight 2 to non-paraphrasing pairs balances the two classes, paraphrasing and non-paraphrasing, relatively evenly. We randomly divide the sentence pairs into a training set of 1435 pairs and validation, development, and testing sets of 718 pairs each (a 40-20-20-20 split), following the natural distribution of lexical paraphrasing and non-paraphrasing pairs. We apply

[1]While it is not possible for an alignment to be annotated with all six rewriting operations, since generalization and specification are mutually exclusive, it is possible for a sentence pair to have both. Since a sentence pair is made up of multiple alignments, it could contain two non-overlapping alignments, one with generalization and one with specification.

the weights described above only to the training and validation sets by weighting the loss function.

6.3.2 The MSR RTE Corpus

The 1435 sentence pairs of the Reddit summarization corpus training set are nowhere near enough data to train a neural model, so we collect both positive and negative examples of lexical paraphrasing for summarization from three other corpora, none of which is a perfect match for our task, but each of which is imperfect in a different way.

The first is the MSR RTE corpus; as we showed in the previous chapter, it contains high-quality, human-annotated paraphrases in the form of the *possible* alignments. It was designed for entailment, not summarization, but on examining the MSR RTE sentence pairs, we find that, if a premise entails a hypothesis, the latter can be considered a (not necessarily good) summary of the former. Figure 6.4 shows two sentence pairs where the premise entails the hypothesis. In the first pair, the hypothesis misses the main point of the premise sentence, but it could be considered a very poor summary with bad content selection; in the second pair, the hypothesis is actually a fairly good summary of the premise, if a little light on the details.

Premise: Sunday's earthquake was felt in the southern Indian city of Madras on the mainland, as well as other parts of south India.
Hypothesis: The city of Madras is located in Southern India.

Premise: Police sources stated that during the bomb attack involving the Shining Path, two people were injured.
Hypothesis: Two people were wounded by a bomb.

Figure 6.4: Entailed sentence pairs from the MSR RTE.

Because the MSR RTE corpus is very small, only 1600 sentence pairs, we use both the training and testing sets of the corpus to collect training pairs for our lexical paraphraser. We filter the 1600 sentence pairs in the MSR RTE to remove those where the premise does not entail the hypothesis, as well as those that do not contain any *possible* alignments. We are left with 403 sentence pairs where the hypothesis is both a (possibly poor) summary and a paraphrase of the premise, and we use these as positive training examples. We randomly select an additional 403 pairs where the hypothesis is entailed by the premise, but there are no *possible* alignments, to use as negative examples.

6.3.3 The New York Times Annotated Corpus

Because the MSR RTE data provides poor summaries, we use the New York Times (NYT) annotated corpus as another data source. Unlike MSR RTE, the NYT corpus consists of human-written summaries of news articles, but it does not contain any paraphrase information. To identify document-summary sentence pairs that are paraphrases of each other, we use a faster version of the pointer-aligner we desigined in the previous chapter. This faster aligner differs from the full pointer-aligner in two ways:

- Rather than aligning every possible chunking of the source and target sentences, it aligns only three chunkings per sentence.
- It generates chunk embeddings by averaging the embeddings of the words in the chunk, rather than using the LSTM language model.

Its performance is somewhat worse than that of the full pointer-aligner (Table 6.2), likely due to a combination of less flexibility in chunk size, less training data resulting from fewer

chunkings, and the use of averaged chunk embeddings, which we had found to underperform the LSTM embeddings.

	Reddit Test Set			MSR RTE Test Set		
System	P%	R%	F_1%	P%	R%	F_1%
full pointer	54.3	79.5	64.5	23.4	47.7	31.4
fast pointer	51.6	53.1	52.3	21.6	28.8	24.7

Table 6.2: Performance of the fast pointer-aligner compared with the full pointer-aligner.

The faster version of the pointer-aligner still requires parsing all sentence pairs, and the news articles of the NYT corpus are much longer than the extractive summaries of the Reddit summarization corpus, which average four sentences long, or the premises of the MSR RTE corpus, which are at most two sentences long. To avoid having to parse every single sentence in the NYT articles, we first filter out article-summary sentence pairs that are less likely to be paraphrases of each other, using the Paraphrase Database (PPDB) (Ganitkevitch, Van Durme, and Callison-Burch 2013): for each word in a summary sentence that is part of a paraphrase pair in the PPDB, we filter out article sentences that do not contain the other half of that paraphrase pair. As we argued in the previous chapter, the paraphrase pairs in the PPDB are fairly strict in that they mostly consist of close synonyms; for an article-summary sentence pair not to contain any PPDB paraphrase pair at all would mean that the sentences did not contain any close synonym matches, and thus would be unlikely to be paraphrases of each other.

After filtering with the PPDB, we parse the remaining candidate sentence pairs and align them with the fast pointer-aligner. Any sentence pair whose aligned spans are at least half as long as the sentences themselves (such that the sentences are not identical to each

Article: in the house , bipartisan support is coalescing around a proposal that would create an affordable-housing fund by setting aside a small portion of profits from fannie mae and freddie mac , the federally sponsored mortgage finance giants .

Summary: proposal with bipartisan support in house would create affordable-housing fund by setting aside small portion of profits from fannie mae and freddie mac

Figure 6.5: An article-summary sentence pair selected from the NYT annotated corpus.

other) are considered paraphrases. Figure 6.5 shows one such selected article-summary sentence pair. One concern with this data is that the language used in news articles is very different from the much more colloquial personal narratives; to combat this issue, we select sentence pairs from only the Opinion section of the NYT corpus, where we are more likely to find subjective, emotional language. We find 8170 paraphrasing sentence pairs in the Opinion section and randomly select an additional 8170 negative pairs from among those that had been filtered out by the PPDB.

6.3.4 The Webis-TLDR-17 Corpus

Finally, we use the Webis-TLDR-17 corpus (hereafter called the Webis corpus for brevity). Like the NYT corpus, the Webis corpus is a summarization corpus. It is, however, drawn from Reddit posts rather than news articles, and so matches more closely the type of language found in our own Reddit summarization corpus (although Webis is not restricted to the genre of personal narrative). As we showed in Chapter 3, the Webis summaries are of lower quality than those in either the NYT corpus or our Reddit summarization corpus; they were extracted from Reddit posts using the keyword "TLDR," or "too long; didn't read," and were not always true summaries of their posts. Thus, the Webis data rounded out our three additional data sources:

- The MSR RTE data contains high-quality paraphrases but low-quality summaries and does not match the language of the Reddit summarization corpus.
- The NYT data contains high-quality summaries and paraphrases of unknown quality and does not match the language of the Reddit summarization corpus.
- The Webis data contains summaries and paraphrases of unknown quality but does match the language of the Reddit summarization corpus.

Post: more important than bindings are your boots .

Summary: bindings matter , but boots matter more .

Figure 6.6: A post-summary sentence pair selected from the Webis corpus.

We filter and align sentence pairs in the Webis corpus following the same procedure we used on the NYT corpus, resulting in 26,591 paraphrasing pairs and an equal number of randomly selected non-paraphrase pairs. Webis is far and away the largest of the four data sources we use in our experiments, and as we shall see, this turned out to be both a blessing and a curse. Figure 6.6 shows a paraphrasing pair selected from the Webis corpus.

6.4 Experiments

The base model described in Section 6.2 is designed to be an end-to-end lexical paraphrasing system: it is meant to learn both to determine whether or not a given input sentence should be paraphrased, and if it should, to perform the paraphrasing. This is a tall order, and so while we hoped that the end-to-end system would be able to learn both when and how to paraphrase, we wanted to be able to iterate on our approach to address any weaknesses

we might discover in it. For this reason, the evaluations in this section are performed on a development set of 718 sentence pairs from the Reddit summarization corpus. We show the final evaluation on the Reddit test set, along with discussion of the results and comparisons to baseline approaches from prior work on paraphrasing, in the next section; this section compares different iterations of our own approach.

6.4.1 End-to-End Lexical Paraphrasing

The base model consists of a bidirectional single-layer LSTM encoder with a hidden size of 512 and an embedding size of 307, a single-layer LSTM decoder with hidden size 1024 and embedding size 307, and an additive attention layer that doubles as the pointer-network for copying directly from the input. The first three hundred dimensions of the embedding layers of both the encoder and decoder are initialized with the GloVE 300-dimensional Common Crawl embeddings (Pennington, Socher, and Manning 2014); we limit the vocabulary to the 50,000 most common words in the Common Crawl, along with special sentence start, sentence end, and out-of-vocabulary tokens.

The last seven dimensions of the embedding layer are intialized with word length; activation, evaluation, and imagery scores from the Dictionary of Affect in Language (DAL) (Whissell 1989); the activation-evaluation norm for subjectivity (Agarwal, Biadsy, and McKeown 2009); and word-level formality and complexity scores (Pavlick and Nenkova 2015). We increase the coverage of the DAL by initializing missing words with the scores of their closest synonym in WordNet (Miller 1995), or inverting the score of their top antonym, if no synonyms are available. Any words in the vocabulary still missing DAL

scores and all words missing formality or complexity scores are initialized with "neutral" values equal to the average of all scores in their respective lexicons.

We train the model on both "positive" examples, where the input is an extractive summary or document sentence and the target is an abstractive summary sentence paraphrasing the input, and "negative" examples, where the input and target are both the same sentence, an extractive summary or document sentence that does not have a corresponding abstractive summary sentence paraphrasing it. Thus, on non-paraphrasing inputs, the model is trained to behave as an autoencoder, while on paraphrasing inputs, it is trained to perform paraphrasing.

We use several different combinations of data sources:

- The Reddit summarization data only; 1435 weighted sentence pairs.
- The Reddit data and the MSR RTE data; an additional 806 sentence pairs.
- The Reddit data, MSR RTE data, and NYT data; an additional 16,340 sentence pairs.
- The Reddit data, MSR RTE data, NYT data, and Webis data; an additional 53,182 sentence pairs.

We train for 20-31 epochs, depending on the training data sources, using early stopping based on the validation set error (regardless of which training data sources we use, we always validate on the 718 weighted sentence pairs in the Reddit summarization corpus validation set). We train using scheduled sampling with linear decay.

Tables 6.3 and 6.4 show the performance of the base, end-to-end lexical paraphrasing model on the Reddit summarization corpus development set. We evaluate performance using two metrics, reflecting the two different facets of the task: word error rate, for the

autoencoder behavior desired on non-paraphrasing inputs (Table 6.3), and ROUGE, for the lexical paraphrasing and summarizing behavior desired on paraphrasing inputs (Table 6.4).

	Word Error Rate		
		Paraphrasing	
Training Data Source	Non-Paraphrasing	Input	Target
Reddit	9.44	57.28	46.38
+RTE	7.04	52.99	41.30
+NYT	**6.47**	**51.72**	**38.61**
+Webis	11.14	62.06	46.45

Table 6.3: Word error rate for the end-to-end lexical paraphrasing system on the Reddit summarization corpus development set.

The three right columns in Table 6.3 show the word error rate (WER) between the system output and non-paraphasing inputs (i.e., when the system should behave as an autoencoder), as well as between the system output and the input and target for paraphrasing inputs (i.e., when the system should rewrite the input into the target). Ideally, we would like for WER of the system output to be low for the non-paraphrasing inputs (i.e., the system successfully returns the input sentence unchanged); it should be higher on paraphrasing inputs and low again for the paraphrasing target (i.e., the system rewrites the input into the target). What we find instead was that the system performs well on non-paraphrasing inputs and goes crazy on paraphrasing inputs; the WER is high for both the input and the target.

Table 6.4 shows ROUGE-1, -2, and -L scores for the system output relative to both non-paraphrasing and paraphrasing inputs. Comparing the scores on non-paraphrasing inputs to those on paraphrasing inputs, there is noticeable drop in ROUGE on paraphrasing inputs. That is, ROUGE is high on non-paraphrasing inputs because the system reproduces them

	ROUGE-1		
		Paraphrasing	
Training Data Source	Non-Paraphrasing	Input	Target
Reddit	55.52	50.56	18.62
+RTE	64.21	58.97	19.91
+NYT	**68.06**	**59.84**	**22.50**
+Webis	58.13	34.45	20.28
	ROUGE-2		
		Paraphrasing	
Training Data Source	Non-Paraphrasing	Input	Target
Reddit	31.84	30.45	3.29
+RTE	44.65	40.29	4.29
+NYT	**51.18**	**38.77**	**5.50**
+Webis	38.30	13.04	4.43
	ROUGE-L		
		Paraphrasing	
Training Data Source	Non-Paraphrasing	Input	Target
Reddit	52.46	50.09	14.79
+RTE	63.87	58.69	16.51
+NYT	**67.96**	**59.49**	**17.91**
+Webis	57.68	39.11	16.50

Table 6.4: ROUGE for the end-to-end lexical paraphrasing system on the Reddit summarization corpus development set.

without modification, but ROUGE is lower on paraphrasing inputs because the system does modify them; however, ROUGE is very low for the paraphrasing target, so whatever modifications the system is making, it is not producing the target. What seems to be happening is that the system can distinguish between paraphrasing and non-paraphrasing inputs, but it is not at all successful in rewriting the paraphrasing inputs into their target sentences. This came as a huge surprise; we expected that performing lexical paraphrasing would be the easier of the two tasks, and determining which sentences ought to be paraphrased would be more difficult – but instead, it was the reverse.

Input: I found an iPhone 3g at the Disneyland parking lot .
Output: i found an iphone shop at the disneyland parking lot .

Input: There is a woman standing next to me when a huge piece of dirt comes flying straight at her face .
Output: there is a woman standing sitting to me when a huge piece of dirt comes blowing straight at her face .

(a) System output on non-paraphrasing inputs.

Input: I once flew into a foreign country for this woman 's wedding since I was a bridesmaid .
Target: I flew to a foreign country to be my friend 's bridesmaid .
Output: another guy turns into any foreign security for this guy 's wedding because she was a hitchhiker

Input: This damn caterer , who from day one went on and on about no hidden charges and his price was all inclusive , called me the day before my wedding to ask if I had found a rental company to supply the glassware and chairs for the reception ...
Target: The caterer for my wedding who insisted that his services were all inclusive , called me a week before my wedding to ask if I had found a rental company to supply glassware and chairs .
Output: another monkey comes out from the day one night on and other aunts called the whole victim looks all inside , called me the day before my wedding to ask if i had found a ladder company to the oilfield and chairs for the whole ...

(b) System output on paraphrasing inputs.

Figure 6.7: End-to-end lexical paraphrasing system output on the Reddit summarization corpus development set.

Figure 6.7 shows some outputs of the best end-to-end lexical paraphrasing system, trained on the Reddit+RTE+NYT data, on the Reddit summarization corpus development set. The system does fairly well on non-paraphrasing inputs, making some minor errors and occasionally leaving off sentence-final punctuation. On paraphrasing inputs, its behavior is bizarre; it seems to be trying to rewrite the input, but not very well, producing an output sentence that roughly matches the input in terms of sentence structure and punctuation, but with some words deleted and others replaced with nonsense. We conclude that asking a single system to behave as an autoencoder on some inputs and as a paraphraser on other inputs is too much. Thus, our next experiment is to split the task into two subtasks: first

identifying which sentences should be paraphrased, and then separately performing the paraphrasing.

6.4.2 Two-Stage Paraphrasing: Sentence Classification

In this experiment, we address the lexical paraphrasing task in two stages. First, we train a model to classify input sentences as either paraphrasing or non-paraphrasing; then, we train a second model to perform lexical paraphrasing using paraphrasing sentence pairs only, without any non-paraphrasing pairs that might distract it. Since the end-to-end system is already able to distinguish between paraphrasing and non-paraphrasing inputs to some extent, we modify it to to perform the classification stage.

6.4.2.1 Sentence Classification

The base model in the previous experiment did fairly well on non-paraphrasing inputs but generated nonsense on paraphrasing inputs, so we knew that it had somehow learned the difference between the two. However, we did not know where that difference was represented in the base model, so we tested three different methods for using it to classify input sentences:

1. Encode the base model's output using its own encoder, then pass both the encoded input sentence and the encoded output sentence through a bilinear layer and five fully-connected layers to gradually reduce the hidden size of 1024 to the classification output size of 2.

2. At each time step in the decoder, pass the decoder state and attention context vector through a bilinear layer, then run the whole sequence through a recurrent network and five fully-connected layers, as above.

3. Concatenate the two final hidden states for each direction of the bidirectional encoder (i.e., the decoder's initial hidden state) and run them directly through five fully-connected layers, as above.

We train these three classifiers using the same four training sets as in the previous section, with the base model's encoder, decoder, and attention layer fixed during this training. We also tune four baseline classifiers using the WER and ROUGE-1, -2, and -L scores: each baseline takes the metric score between the base model's output and the input sentence and compares it against a threshold tuned on the validation set. For example, the WER baseline classifies an input sentence as paraphrasing if the the word error rate between the input sentence and the base model output on that sentence is greater than 39.4.

Table 6.5 shows the performance of each classifier on the Reddit development set. We find that the WER baseline was very strong and is overtaken by the decoder state and attention context vector classifier only after the addition of the larger NYT and Webis training sets. The ROUGE baselines are weaker overall because there is more overlap between the ranges of scores for paraphrasing and non-paraphrasing inputs, although they also improve significantly after adding the NYT and Webis data.

Of the three classification methods, re-encoding the base model's output and comparing it to the encoded input performs the worst, despite the baselines showing that directly

comparing the model's output sequence to the input sequence is a strong signal for paraphrasing versus non-paraphrasing inputs; likely re-encoding the output dulls that signal. Classifying based on the encoder's final state performs better, but comparing the decoder state to the attention context vector, which is a weighted average of the encoder output at each time step, performs the best, suggesting that the base model's ability to distinguish between paraphrasing and non-paraphrasing input sentences is a product of the entire model, and not solely of the encoder.

	Training Data Source											
	Reddit			+RTE			+NYT			+Webis		
System	P%	R%	F_1%	P%	R%	F_1%	P%	R%	F_1%	P%	R%	F_1%
Encode and Re-encode	44.8	69.3	54.4	50.8	59.4	54.8	58.0	68.7	62.9	65.1	73.8	69.2
Decoder State and Context	52.3	62.2	56.8	58.2	66.5	62.1	**65.8**	**77.5**	**71.2**	**74.0**	**79.4**	**76.6**
Encoder Final State	55.9	56.9	56.4	**61.2**	63.0	62.1	63.9	65.7	64.8	71.3	73.3	72.3
WER Baseline	**62.1**	**74.7**	**67.8**	59.6	**68.4**	**63.7**	58.1	68.3	62.8	69.5	78.4	73.7
ROUGE-1 Baseline	45.1	54.3	49.4	44.9	56.7	50.1	49.8	59.7	54.3	59.1	68.2	63.3
ROUGE-2 Baseline	29.4	40.9	34.2	52.6	42.3	46.9	59.5	71.1	64.8	64.0	69.6	66.7
ROUGE-L Baseline	30.6	41.2	35.1	45.3	52.7	48.7	48.9	54.9	51.7	57.3	65.7	61.2

Table 6.5: Paraphrasing classifier performance on the Reddit summarization corpus development set.

6.4.2.2 Sentence Paraphrasing

After an input sentence is classified as a paraphrasing input, we need to perform paraphrasing on that sentence. We retrain the base model on only the paraphrasing sentence pairs from our four data sources. Tables 6.6 and 6.7 show the WER and ROUGE scores for this second-stage, paraphrasing-only system on the Reddit summarization corpus development set.

	Word Error Rate	
Training Data Source	Input	Target
Reddit	35.57	38.73
+RTE	32.09	34.12
+NYT	**27.75**	**31.62**
+Webis	33.38	34.20

Table 6.6: Word error rate for the second-stage paraphrasing system on the Reddit summarization corpus development set.

It is interesting to note that while the addition of the Webis data improved the performances of the paraphrasing/non-paraphrasing classifiers, it hurts the performance of the paraphraser. The Webis data also hurt the performance of the end-to-end lexical paraphraser in the previous section. Examining the sentence pairs in the Webis data, we find that the sentences had a high proportion of out-of-vocabulary words, such as character names from movies, video game titles, and unexplained abbreviations and acronyms; this may have made it difficult for the paraphraser to learn from these examples.

In Table 6.6, the WER between the second-stage paraphrasing system's output and its input/target is lower than it had been for the end-to-end system (Table 6.3); in Table 6.7 the ROUGE scores comparing this system's input and output is lower than that of the end-to-end system (Table 6.4), and the ROUGE scores between the output and target are

	ROUGE-1		ROUGE-2		ROUGE-L	
Training Data Source	Input	Target	Input	Target	Input	Target
Reddit	33.94	33.84	15.67	13.41	33.48	28.64
+RTE	40.26	38.51	19.71	16.69	40.04	35.93
+NYT	**47.09**	**47.07**	**28.17**	**27.38**	**46.08**	**40.88**
+Webis	37.08	34.49	17.19	16.48	39.26	36.18

Table 6.7: ROUGE for the second-stage paraphrasing system on the Reddit summarization corpus development set.

higher. The overall lower WER suggests that the second-stage paraphrasing system is not introducing as many nonsense words as the end-to-end system, instead producing output that is more similar to both the input and the target. Similarly, the ROUGE scores between the output and the input and between the output and the target are roughly equal, with the input score slightly higher, suggesting that the second-stage paraphrasing system rewrites more conservatively than the end-to-end system. Figure 6.8 shows this system's output on the same two paraphrasing inputs from the previous section. (For comparisons to baselines based on prior work on paraphrase generation, and for more examples to help interpret these scores, see Section 6.5.)

Input: I once flew into a foreign country for this woman 's wedding since I was a bridesmaid .
Target: I flew to a foreign country to be my friend 's bridesmaid .
Output: i flew to a foreign country for funny friend 's bridesmaid .

Input: This damn caterer , who from day one went on and on about no hidden charges and his price was all inclusive , called me the day before my wedding to ask if I had found a rental company to supply the glassware and chairs for the reception ...
Target: The caterer for my wedding who insisted that his services were all inclusive , called me a week before my wedding to ask if I had found a rental company to supply glassware and chairs .
Output: the caterer who from day one called the whole price all inclusive , called me the day before my wedding to ask if i had found a furniture company to hire lining and chairs

Figure 6.8: Second-stage paraphrasing system output on the Reddit summarization corpus development set.

The paraphrasing-only system is more likely to break free from the input sentence structure and punctuation than the end-to-end system was; the latter needed to balance both paraphrasing and non-paraphrasing performance. It is also more successful at generating words and phrases that do not appear in the input, without devolving into complete nonsense. The first example in Figure 6.8 is a fairly good paraphrase (not by human standards, but very good for an automatic system, as we show in the final evaluation); the second is not bad, either, except that, in addition to rewriting "who from day one went on and on about no hidden charges and his price was all inclusive" into the fairly reasonable "who from day one called the whole price all inclusive," it also rewrites the end of the sentence, which is identical between the input and the target and should not be paraphrased ("furniture," "hire," and "lining" versus "rental," "supply," and "glassware").

6.4.3 Three-Stage Paraphrasing: Intra-Sentence Tagging

The two-stage classify-and-paraphrase approach worked reasonably well, except that the paraphraser sometimes rewrote parts of the input sentences that should not be rewritten. To address this problem, we divide the task into three stages: first classifying which sentences ought to be paraphrased, then tagging which words within those sentences ought to be paraphrased, and finally performing the paraphrasing. We already had a working sentence-level paraphrase classifier, so in this section we focus on the tagging and paraphrasing subtasks.

6.4.3.1 Intra-Sentence Tagging

To identify the words or phrases in an input sentence that ought to be paraphrased, we modify the base model to perform tagging. At each time step, instead of having the decoder predict a distribution over the vocabulary, we modify the decoder to predict a distribution over two word-level tags, *paraphrase* and *not-paraphrase* (as with the sentence-level classifier, we use five fully-connected layers to gradually reduce the decoder's hidden size to the output size). The input to the decoder at each time step is the concatenation of the embedding for the current input sentence word to be tagged and the tag for the preceding word.

All four of our data sources can be converted into tagging training data:

- The Reddit summarization corpus alignments are expanded into sentence pairs for the previous experiments. For tagging, we assign all words in an input sentence (from an extractive summary) that are contained in at least one lexical paraphrasing alignment the *paraphrase* tag, and all other words the *not-paraphrase* tag. We weight each alignment based on the number of rewriting operations contained within it: an alignment is assigned weight $1/n$, where n is the number of rewriting operations it contains, including lexical paraphrasing. Each *paraphrase*-tagged word receives the sum of the weights of all alignments containing it; *not-paraphrase* words receive weight 1.

- The MSR RTE corpus also contains human-annotated alignments between sentences. All words in an input (i.e., premise) sentence that are part of a *possible* alignment

are assigned the *paraphrase* tag, along with any *sure* paraphrase words immediately adjacent to a *possible* alignment.

- The NYT and Webis corpora are filtered using the Paraphrase Database (PPDB) and aligned using the pointer-aligner described in the previous chapter. All words in an input (i.e., document) sentence that are part of an alignment such that the two aligned phrases 1) are not identical and 2) contain a PPDB paraphrase pair are assigned the *paraphrase* tag.

Table 6.8 shows the performance of the intra-sentence paraphrase tagger on the paraphrasing input sentences of the Reddit summarization corpus development set. This evaluation, unlike the other experimental evaluations, is weighted. The sentence-level weights described earlier in this chapter were meant to balance the Reddit summarization corpus training data, since paraphrasing sentence pairs were much rarer than non-paraphrasing pairs. However, in the tagging subtask, the word-level weights represent confidence. The Reddit summarization corpus allowed alignments of any size, and each alignment could be annotated with multiple rewriting operations, so there was no way to determine precisely which words in an alignment belonged to which rewriting operation. Thus, the weight of each word-level tag indicates our confidence that a given word ought to be paraphrased.

Training Data Source	P%	R%	F_1%
Reddit	46.8	54.0	50.1
+RTE	**53.1**	52.2	**52.6**
+NYT	32.4	85.6	47.0
+Webis	25.3	**100.0**	40.4

Table 6.8: Intra-sentence paraphrase tagging performance on the Reddit summarization corpus development set.

Unfortunately, the tagging approach was not successful. Training only on the Reddit and RTE data seemed promising, but those training sets were small. With the addition of the larger NYT and Webis training sets, the system learns to take the easy way out and simply tags every single word as *paraphrase*. In both the NYT and Webis data, the input and target sentences are often very similar to each other; in the Webis data especially, the sentences are very short, and are often only one word or punctuation mark away from being identical to each other. As a result, most of each input sentence (or all of it, in the case of the Webis sentences) is tagged as *paraphrase* in the training data. The high degree of similarity between input and target sentences in the NYT and Webis data is due to the high degree of similarity between the summaries and their documents in those corpora; many NYT and Webis summary sentences are only lightly rewritten versions of some sentence in their documents; heavily rewritten summary sentences, like those found in the Reddit and MSR RTE corpora, are rare.

6.4.3.2 Intra-Sentence Paraphrasing

For curiosity's sake, we continue on to the intra-sentence paraphrasing stage of the three-stage approach, despite our lack of success with the tagger. If we had an oracle tagger, how would the three-stage approach perform? We retrain the base model once again, using only the *paraphrase*-tagged portion of each input sentence as input and the corresponding aligned portion of the target sentence as the target; for the sentence pair shown in Figure 6.3 (reproduced here as Figure 6.9), for example, we used the top input and target, rather than the full sentence input and target at the bottom.

Tables 6.9 and 6.10 show the performance of this intra-sentence paraphraser, trained

Alignment:
Extractive: from the looks of the video tried to slip my friend a couple of "somethings" into his drink!
Abstractive: Security tapes revealed that she attempted to drug him,

Sentence Pair:
Extractive: It turns out the VICTIM girl was the one that pulled out the drugs and from the looks of the video tried to slip my friend a couple of "somethings" into his drink!
Abstractive: Security tapes revealed that she attempted to drug him, and intentionally drugged herself.

Figure 6.9: An alignment and its corresponding sentence pair.

on each of our four data sources, on the Reddit summarization corpus development set, evaluated against aligned target sentence portions, rather than full target sentences.

	Word Error Rate	
Training Data Source	Input	Target
Reddit	42.43	46.88
+RTE	41.15	41.17
+NYT	**34.35**	**33.88**
+Webis	34.75	37.92

Table 6.9: Word error rate for the intra-sentence paraphrasing system on the Reddit summarization corpus development set.

The intra-sentence paraphraser underperforms the sentence-level, second-stage paraphraser from the previous section (Tables 6.6 and 6.7) by about 1-7 WER and 1-8 ROUGE. The difference in performance is most pronounced when training on only the Reddit and RTE data; the addition of the NYT and Webis data somewhat mitigates the loss in performance.

The input and target sequences in the Reddit and RTE data are derived from paraphrase alignments, but there is nothing that constrains those alignments to be between constituents of the same type. In Figure 6.9, for example, the input sequence is the predicate "from the

	ROUGE-1		ROUGE-2		ROUGE-L	
Training Data Source	Input	Target	Input	Target	Input	Target
Reddit	33.94	30.20	9.66	9.01	29.11	26.16
+RTE	38.51	32.11	14.12	12.48	35.80	30.44
+NYT	**45.49**	**44.10**	**24.94**	**24.48**	**42.93**	**39.27**
+Webis	35.97	34.09	15.63	15.42	34.96	28.51

Table 6.10: ROUGE for the intra-sentence paraphrasing system on the Reddit summarization corpus development set.

looks of the video tried to slip my friend a couple of "somethings" into his drink!", while the target is the clause "Security tapes revealed that she attempted to drug him." Where the sentence-level paraphraser was guaranteed complete sentences for input and target, there is no rhyme or reason governing the types of constituents that the intra-sentence paraphraser might receive as input and target. As a result, the Reddit- and RTE-trained intra-sentence paraphraser never learns proper English sentence structure and produced mostly topically-related gibberish.

Input: This damn caterer , who from day one went on and on about no hidden charges and his price was all inclusive , called me the day before my wedding to ask if I had found a rental company to supply the glassware and chairs for the reception ...
Target: The caterer for my wedding who insisted that his services were all inclusive , called me a week before my wedding to ask if I had found a rental company to supply glassware and chairs .

Reddit Only: towards flight 's wedding
Full Training: the caterer for my wedding who chose that his divorce was all inclusive , called the day before my wedding to ask if i had found a blizzard company to supply leg and chairs
Sentence-level: the caterer who from day one called the whole price all inclusive , called me the day before my wedding to ask if i had found a furniture company to hire lining and chairs

Figure 6.10: Reddit-only and full training set intra-sentence paraphrasing systems' output on the Reddit summarization corpus development set.

Figure 6.10 shows the output of the Reddit-only intra-sentence paraphraser and the

output of the full intra-sentence paraphraser trained on all four data sources. The Reddit-only system produces (a very small) word salad, while the system trained on all four data sources performs comparably to the sentence-level paraphraser in the previous section. As we discussed earlier in this section, the NYT and Webis alignments are mostly full sentences; being much larger than the Reddit and RTE training sets, the NYT and Webis training sets pull the intra-sentence paraphraser back towards the behavior of the sentence-level paraphraser.

6.5 Final Results and Discussion

In the previous section, we described how we experimented with different approaches to lexical paraphrasing for summarization. Of the three approaches we tested – end-to-end paraphrasing; two-stage input classification and paraphrasing; and three-stage input classification, paraphrase tagging, and intra-sentence paraphrasing – the two-stage approach was the most successful. In this section, we discuss our final evaluation of this approach on the Reddit summarization corpus test set, compare it with previous approaches to paraphrase generation, and analyze its performance.

6.5.1 Final Evaluation

We evaluate the two-stage, classify-then-paraphrase approach on the 718 sentence pairs of the held-out Reddit summarization corpus test set – of the test pairs, 181 are paraphrasing and 537 are non-paraphrasing. For the paraphrasing/non-paraphrasing input sentence classifier, we use the decoder state and attention context vector model trained on the full

training set of all four data sources; for the sentence-level paraphraser, we use the model trained only on the Reddit, RTE, and NYT training sets, since excluding the Webis training data improved performance for this model on the development set.

Table 6.11 shows the performance of the paraphrasing/non-paraphrasing input sentence classifier, along with two baselines: a Paraphrase Database (PPDB) baseline and a neural machine translation (MT) baseline. The two baselines represent the two main previous approaches to paraphrase generation, lexicon-based paraphrasing and translation-based paraphrasing. The PPDB baseline classifies an input sentence as paraphrasing if it contains at least one word or phrase that was part of a known paraphrase pair in the PPDB, and as non-paraphrasing otherwise. The NMT baseline translates each input sentence into a pivot language and back to English; it classifies a sentence as paraphrasing if its translation is different from the original sentence (excluding pronoun gender and verb tense errors), and as non-paraphrasing otherwise. We use Google Translate as our NMT system, with Czech as the pivot language, following Mallinson, Sennrich, and Lapata (2017), Wieting, Mallinson, and Gimpel (2017), Wieting and Gimpel (2018), and Hu et al. (2019).

System	P%	R%	F_1%
Decoder State and Context	**73.2**	**78.2**	**75.6**
PPDB Baseline	25.5	100.0	40.6
NMT Baseline	27.2	100.0	42.8

Table 6.11: Paraphrasing classifier performance on the Reddit summarization corpus test set.

The decoder state and attention context vector model performs about the same on this test set as it did on the development set (Table 6.5), scoring in the 70s for precision, recall, and f-measure. The PPDB and NMT baselines perform comparably to each other. Both

achieve perfect recall at the cost of very low precision – paraphrasing sentences comprise just over one quarter of the input sentences in the test set, so it is clear that both baselines consider nearly every single sentence to be a paraphrasing sentence. This illustrates a major limitation in previous work on paraphrasing: prior research focused entirely on how to produce paraphrases, without giving any thought to whether and when paraphrasing should be performed.

	Word Error Rate	
System	Input	Target
Sentence Paraphraser	32.34	**36.07**
PPDB-oracle Baseline	**6.04**	121.98
PPDB-random Baseline	213.29	274.98
NMT Baseline	44.00	101.80

Table 6.12: Word error rate for sentence-level paraphrasers on the Reddit summarization corpus test set.

Tables 6.12 and 6.13 show the WER and ROUGE scores of the sentence-level paraphraser and the PPDB and NMT baselines. We use two different PPDB baselines in this evaluation. Both first search the input sentence for candidate words and phrases that are part of at least one known paraphrase pair in the PPDB. The *PPDB-random* baseline then selects one paraphrase pair at random, for each candidate word or phrase, and substitutes the paraphrase into the sentence. The *PPDB-oracle* baseline cheats in order to more closely emulate our goal of learning when and how humans paraphrase for summarization; given a candidate word or phrase in the source sentence, it searches the target sentence for the correct paraphrase pair to use and does nothing if no matching paraphrase is found. The NMT baseline cannot be focused in this way (if the paraphrase tagger were more successful, we might be able to target paraphrase words with lexically-constrained decoding), so it does

not have an oracle version.

	ROUGE-1		ROUGE-2		ROUGE-L	
System	Input	Target	Input	Target	Input	Target
Sentence Paraphraser	46.40	**45.38**	27.32	**25.51**	45.53	**42.53**
PPDB-oracle Baseline	**88.67**	28.64	**81.04**	9.18	**88.67**	21.80
PPDB-random Baseline	31.15	13.44	4.57	0.50	28.57	10.17
NMT Baseline	64.62	29.03	39.69	11.37	61.11	23.53

Table 6.13: ROUGE for sentence-level paraphrasers on the Reddit summarization corpus test set.

As in the classification stage, our second-stage, sentence-level paraphraser performs similarly on the test set as it did on the development set (Tables 6.6 and 6.7). For the PPDB-oracle and NMT baselines, their WER with respect to the input is very low, while their WER with respect to the target is very high; for comparison, the average WER between an input sentence and its target sentence in this test set is 119.04. Further, their ROUGE scores with respect to the input are very high – they do not rewrite the input sentence very much. The PPDB-random baseline, on the other hand, goes wild rewriting the input into barely recognizable output sentences. Figure 6.11 shows output from our sentence-level paraphraser and the three baselines.

On these sentence pairs, the PPDB-oracle baseline is unable to find a paraphrase pair that matches both the input and target sentences, so it does nothing. The PPDB-random baseline finds many paraphrasing candidates and randomly selects a paraphrase for each one; the resulting sentences are confusing and badly overwritten – from the language, it is quite apparent that the PPDB's main source of paraphrase pairs is EU Parliament proceedings. The sentence-level paraphraser and NMT baseline fall somewhere in between: both produce fairly good sentences, with the paraphraser leaning more on the side of rewriting

Input: She then asked to speak to my manager and shortly thereafter faxed over a written complaint about me and how horribly she had been misled and abused by me .
Target: A potential customer filed a written complaint about me , falsely saying that I verbally abused and misled her .

Sentence Paraphraser: she updated a written complaint about how she had been frustrated and abused by me .

PPDB-oracle Baseline: She then asked to saying that to my manager and shortly thereafter faxed over a written filed a me and how horribly she had been misled and abused by me .

PPDB-random Baseline: She asked to see my country point agent and , subsequently , more recently in the future faxed writing an international complainant around how so terribly she has therefore been misled and , subsequently , maltreatment by way of love by me .

NMT Baseline: She then asked to speak to my manager and then fax shortly after through a written complaint about me and how terribly she was deceived and abused.

(a)

Input: She then proceeded to screw up every single part of her job .
Target: My wedding coordinator criticized my preferences during the planning phase of the wedding , then on the day of the wedding she failed to do her job .

Sentence Paraphraser: she refused to do her job .

PPDB-oracle Baseline: She then proceeded to screw up every single of her job .

PPDB-random Baseline: She shall now proceed to fucked up everyone , and concerning his father was working .

NMT Baseline: Then she went to every single piece.

(b)

Figure 6.11: Sentence-level paraphraser output on the Reddit summarization corpus test set.

(perhaps a bit too much; its claim that the wedding coordinator "refused" to do her job is not supported by either the input or the target in Figure 6.11b) and the NMT baseline leaning more on the side of preserving the input, although it seems to have trouble with the more colloquial phrase "screw up."

These sentence pairs also illustrate the challenge of paraphrasing for summarization.

Both examples contain several other rewriting operations, besides paraphrasing, which make it very difficult for a paraphrase system to produce the target sentence exactly because it must somehow deal with the other rewriting operations as well. In Figure 6.11a, the human summarizer's use of *specification* makes it impossible for a paraphrasing system to generate "a potential customer" without inventing it from whole cloth. In Figure 6.11b, the human summarizer must have used *fusion* to combine some other sentence with the input to produce the target; the entire target clause "my wedding coordinator criticized my preferences during the planning phase of the wedding" is impossible to generate for a paraphrasing system that was given only one input sentence.

Input: When the family asked me why I was beaming , I told them what happened and they proceeded to ridicule me for the rest of the day calling me " hero " sarcastically .
Target: *I saved my friend 's life* and his family ridiculed me .
Output: later family noticed beaming , told them what happened and they proceeded to laughing me

Input: As she hung up , she screamed into the room with the party " Who the fuck abandons their bride on their wedding night ? !
Target: A groom left his bride on their wedding night *to go to the bar with some friends* .
Output: she called came across the room " who the fuck abandons their wife on her wedding night ? "

Input: My step mother saw part of what I was doing but to her I was taking money .
Target: *I attempted to increase my sister 's chances of winning a board game* , however my step-mother misinterpreted my actions as cheating to beat my sister .
Output: step mother noticed what i was doing but to her was about money

Input: Saved a girl on an escalator when she got her heel stuck and fell backwards .
Target: I saved a woman from falling backwards down an escalator , and *she accused me of sexually assaulting her* .
Output: saved a lady from being on an escalator when she got her heel stuck

Figure 6.12: Sentence-level paraphraser output on impossible sentence pairs, with unrecoverable information in the target sentence shown in *italics*.

Examining the Reddit summarization corpus test set, we find that 58.8% of target sen-

tences are impossible for the paraphrasing system to generate exactly, even if we exclude target sentences like that in Figure 6.11a, which contain only a small number of impossible-to-generate words; Figure 6.12 shows examples of impossible input/target sentence pairs, along with the sentence-level paraphraser's output. Further, out of the 1218 total lexical paraphrasing alignments in the Reddit summarization corpus, only three contain no other rewriting operations – paraphrasing in the context of summarization never occurs in isolation. Any one of those other rewriting operations could be its own area of research, and their presence in paraphrasing sentence pairs complicates the task far beyond the capabilities of previous, naive approaches to paraphrase generation; even our two-stage model, which outperforms previous approaches by a large margin, cannot fully solve the problem of other operations.

6.5.2 Feature Ablation

Given how difficult paraphrasing for summarization turned out to be, it is somewhat amazing that we are able to do as well as we do. Our paraphrasing/non-paraphrasing classifier works fairly well, and our sentence-level paraphraser produces reasonable paraphrases, even if it rarely exactly matches the target sentence. How do they do it?

While it is difficult to interpret exactly what is going on inside a neural network, a reasonable question to ask would be, do the seven extra features in our augmented word embeddings help? We initialized the embedding layers of our encoders and decoders with word length; activation, pleasantness, evaluation, and subjectivity scores; and formality and complexity scores. Do the models actually use these scores, or do they simply use the extra

seven dimensions to store some other information that they learn to be important during trianing? To answer these questions, we retrain both the paraphrasing/non-paraphrasing input sentence classifier and the sentence-level paraphraser several times, initializing only six of the seven extra features each time – in effect, ablating the seventh feature; we also train the models one final time with none of the seven extra features initialized.

System	P%	R%	F_1%
Full Feature Set	**73.2**	**78.2**	**75.6**
- Activeness*	66.7	71.2	68.9
- Pleasantness*	69.6	74.1	71.8
- Subjectivity*	64.8	72.0	68.2
No Extra Features*	56.7	66.5	61.2

Table 6.14: Paraphrasing classifier feature ablation results. * indicates significant difference from the full feature set ($p < 0.024$).

For the paraphrasing/non-paraphrasing input sentence classifier, we find that removing the activeness, pleasantness, and subjectivity features significantly lowers classification performance, while removing the other four features individually has no significant effect. Table 6.14 shows the performance of the classifier with each of those features ablated. Note that the middle three rows of the table correspond to experiments where a single feature is not initialized; the "removal" of features does not accumulate across rows.

For the sentence-level paraphraser, we find that removing any feature other than word length and pleasantness has a significant impact on both WER and ROUGE (Tables 6.15 and 6.16). Removing features has a larger effect on the sentence-level paraphraser than it does on the paraphrasing/non-paraphrasing classifier.

It is interesting to note that removing features from the sentence-level paraphraser reduces its performance to below that of the full-featured paraphraser trained only on the

System	Word Error Rate	
	Input	Target
Full Feature Set	**32.34**	**36.07**
- Activeness*	66.76	86.98
- Imagery*	71.78	89.67
- Subjectivity*	64.18	82.61
- Formality*	63.37	82.59
- Complexity*	53.49	77.22
No Extra Features*	77.45	89.67

Table 6.15: Word error rate for sentence-level paraphraser feature ablation. * indicates significant difference from the full feature set ($p < 0.019$).

System	ROUGE-1		ROUGE-2		ROUGE-L	
	Input	Target	Input	Target	Input	Target
Full Feature Set	**46.40**	**45.38**	**27.32**	**25.51**	**45.53**	**42.53**
- Activeness*	27.82	21.87	5.78	5.06	21.56	16.63
- Imagery*	18.86	17.78	4.77	3.59	18.38	16.40
- Subjectivity*	23.96	21.08	8.71	7.02	23.03	19.29
- Formality*	32.02	27.78	10.49	8.39	29.52	26.13
- Complexity*	41.73	35.85	19.79	12.04	42.57	40.93
No Extra Features*	14.94	11.65	1.13	0.90	14.33	10.8

Table 6.16: ROUGE for sentence-level paraphraser feature ablation. * indicates significant difference from the full feature set ($p < 0.041$).

Reddit data, suggesting that the extra features we use to augment the word embeddings are doing most of the heavy lifting in terms of learning to perform paraphrasing.

Figure 6.13 shows the output of the sentence-level paraphraser trained without initializing any of the seven extra features: it is basically gibberish. That the seven extra features have such a tremendous impact on the performance of the paraphraser is very surprising; while we expected them to have some impact, we did not at all expect the paraphraser to be unable to function without them.

Input: She then asked to speak to my manager and shortly thereafter faxed over a written complaint about me and how horribly she had been misled and abused by me .
Target: A potential customer filed a written complaint about me , falsely saying that I verbally abused and misled her .
Sentence Paraphraser: another personal account received written complaint about mr.

Input: She then proceeded to screw up every single part of her job .
Target: My wedding coordinator criticized my preferences during the planning phase of the wedding , then on the day of the wedding she failed to do her job .
Sentence Paraphraser: she want need every day of school on her first grade

Figure 6.13: Sentence-level paraphraser trained without initializing extra features.

6.6 Conclusion

The main contributions described in this chapter are:

- We perform the first investigation into the task of lexical paraphrasing specifically in the context of summarization. Previous work on paraphrasing for summarization focused on identifying paraphrases in the input document(s) to assist with either content selection or sentence fusion; performing paraphrasing as part of summarization had not been explored before us.

- The success of our paraphrasing/non-paraphrasing input sentence classifier demonstrates that it is possible to predict which document sentences a human summarizer would paraphrase when writing an abstractive summary. This is in contrast with previous work on paraphrase generation, where one always wanted to paraphrase when possible (e.g., to generate multiple queries for better coverage in question-answering). The relatively poor performance of the Paraphrase Database (PPDB) and neural machine translation (NMT) baselines show that, while it is almost always possible to paraphrase any particular sentence, it is not always desirable to do so.

- Our sentence-level paraphraser is able to learn to rewrite a given input sentence into something closer to its target paraphrase, producing an output sentence that is somewhere in between the two. Our comparisons with the PPDB-oracle, PPDB-random, and NMT baselines show the limitations of lexicon- and translation-based approaches to paraphrase generation. On the one hand, naively and indiscriminately applying machine translation or a paraphrase lexicon results in, at best, a slightly rewritten sentence that still closely resembles the original input, or at worst, an output sentence so heavily and unnaturally rewritten that it is difficult to understand. On the other hand, using a paraphrase selection oracle to determine which words to rewrite, and which other words to rewrite them into, reveals that the PPDB, the largest word- and phrase-level paraphrase lexicon in existence, still has insufficient coverage.

- We demonstrate in our feature ablation experiments that both the paraphrasing/non-paraphrasing sentence classification task and the sentence-level paraphrasing task benefit from the addition of affectual and stylistic word-level features. Just seven features capture the bulk of the important information for the paraphrasing task and play a significant role in the classification task, suggesting that style and affect, more than meaning, drive human summarizers' paraphrasing decisions.

Paraphrasing for summarization is a hard task – harder than we realized when we set out to work on it. The work described in this chapter is full of surprises: the unexpectedly great success of the paraphrasing/non-paraphrasing sentence classifier, the extremely low coverage of the PPDB-oracle baseline, and the outsize importance of the affectual and

stylistic features. There are also some disappointments: that it is possible to determine whether or not a sentence ought to be paraphrased, but not which words within it ought to be paraphrased. In this last chapter on summarization for personal narrative, we conduct a thorough and deliberate series of experiments, probing the boundaries of what is possible in realm of lexical paraphrasing for summarization.

7. Cross-Lingual Summarization

In the preceding chapters, we addressed automatic summarization of personal narratives, a text genre that is commonly found on the Web and that presents a unique set of challenges: the difficulty of identifying important information in a document and the necessity of significant editing and rewriting to generate a well-written summary for a highly informal and colloquial document. In this final chapter, we turn to summarization of a different domain – one that requires even heavier editing and rewriting than personal narrative: cross-lingual summarization for low-resource languages.

7.1 Motivation

Cross-lingual summarization is a little-explored task that combines the challenges of automatic summarization with those of machine translation. The goal is to summarize in one language – English, in our case – a document that is written in another language. Just as personal narratives posted on the Web are an important new source of information that complements news, so foreign-language documents, whether news articles or social media posts, are an important new source of information that complements *English* news. An English speaker with friends or family living in a non-English-speaking country might be interested in that country's local news, but unless he or she speaks the local language, only the most dramatic and eye-catching of events – those that are picked up by international news agencies – are accessible to him or her (Orăsan and Chiorean 2008). Even in the case

of world news events, an English speaker might like to know how the event is viewed in other, non-English-speaking countries; for example, what do the op-eds in African newspapers say about the flood of Chinese investments coming into their countries?

There are two possible approaches to producing an English summary of a non-English document: summarize the document in the source language and then translate the summary into English, or translate the document into English and then summarize the translation. Of course, the availability of human annotators can greatly simplify the task. If human summarizers are available, who speak either English or the source language, the task is reduced to machine translation alone; if human translators are available, the task is reduced to summarization alone. But if there are no human annotators, then it is preferable to summarize in the source language and translate the summary (Wan, Li, and Xiao 2010).

There are two main reasons for preferring this approach. First, machine translation is computationally expensive, and so summarizing and then translating means that fewer sentences need to be translated. Second, errors on the part of an automatic summarization system are less likely to impact the performance of a machine translation system than vice versa. While summarization errors may result in an irrelevant, misleading, or even factually incorrect summary, when compared to the input document, a machine translation system that takes the summary as input is agnostic to its correctness. An extractive summary, regardless of its correctness, will always contain sentences as fluent and well-written as in the original document, and so it will be no harder to translate than the original document; an abstractive summary could produce difficult to translate gibberish, but even then, a two-stage approach, such as the one we used for personal narrative, could be applied to first extract some small number of sentences to be translated.

In contrast, translation errors can drastically affect the performance of a summarization system. A single mistranslated word can change the meaning of a sentence – the French word "l'avocat," for example, can be translated as either "the (male) lawyer" or "the avocado." A summarization system that receives such a mistranslation in its input can only compound the error: it might omit an otherwise desirable sentence because it contains a mistranslation, or it might include the mistranslation but without enough context for a user to realize what has happened, or it might be derailed into an unrelated part of the semantic space. The only way a summarizer could recover from a translation error would be to recognize and correct the error, which is far beyond the purview of existing summarization systems. The summarize-then-translate approach avoids this scenario; if a translation error must be made, better that it should be made at the end of the process than early on.

However, the summarize-then-translate approach can only be used when summarization systems are available in the source language. If the source language is a high-resource language, this is not a problem, but if the source language is one of the thousands of low-resource languages in the world, there are no summarization corpora, much less summarization systems, available. From our discussion of corpora in Chapter 3, it is clear that building a summarization corpus even in English, the highest-resource language of them all, is expensive and time-consuming – it would be downright impractical for very low-resource languages. Language-independent extractive summarization techniques, such as TextRank (Mihalcea 2005), might be used, but there could be serious difficulties in their application, such as morphologically rich languages that require detailed morphological analysis and more sophisticated similarity measures than the word-based measures that work well in English. In such a case, translate-then-summarize is the only possible ap-

proach.

Another problem is that any machine translation systems that exist for a low-resource language are unlikely to have had much parallel corpora on which to train; their translations will be of much lower quality than translations for well-studied, high-resource languages, like French or German. Thus, regardless of whether summarization or translation is performed first, the poor quality of the translations would have a significant effect: a summarization system taking the translations as input would have to contend with errors as discussed above, and even if a language-independent summarization technique could be used first, and its output translated into English, the translation would still be disfluent, difficult to read, and confusing for a human user.

What is needed, then, is for the disfluent translation output to be corrected into more fluent English, either by rewriting or eliding translation errors. In this chapter, we describe how we address this need by building an neural abstractive system for cross-lingual summarization for low-resource languages that learns to generate short, simple phrases to replace awkwardly-phrased input spans resulting from errorful translations.

7.2 Related Work

7.2.1 Cross-Lingual Summarization

Orăsan and Chiorean (2008) used Romanian as the source language and English as the target. They used the summarize-then-translate approach, performing extractive multi-document summarization on Romanian news articles and automatically translating the

summaries into English. Their evaluations showed that the poor quality of the translations transformed reasonable Romanian summaries into barely legible English ones.

The most extensively investigated source-target language pair is English-to-Chinese. Wan, Li, and Xiao (2010) addressed the problem of low-quality translations by training a regressor to predict, given an English sentence, the quality of its Chinese translation. They used this predicted translation quality score as a feature in selecting which English sentences to extract for their summaries, with the goal of selecting sentences that were both informative and easy to translate into Chinese. Wan (2011) explored the problem of unintended information loss or modification between an English sentence and its Chinese translation by first translating all the English sentences into Chinese. He then extracted sentences based on both the original English and the Chinese translation. Yao, Wan, and Xiao (2015) extended both these works by scoring aligned phrases from the original English documents and their Chinese translations to perform joint sentence extraction and compression based on both information and translation quality – their work on compression was the first attempt at abstractive cross-lingual summarization.

Zhang, Zhou, and Zong (2016) further explored abstractive approaches for cross-lingual summarization. They first parsed the original English documents into predicate-argument structures and then created matching structures in Chinese by aligning the original English sentences with their Chinese translations. They then selected an informative and non-redundant set of these bilingual predicate-argument structures and fused the Chinese words and phrases associated with them to create abstractive summary sentences.

Finally, Wan et al. (2018) experimented with both the summarize-then-translate and translate-then-summarize approaches; they extracted and ranked multiple candidate sum-

maries, both English summaries translated into Chinese and Chinese summaries extracted from translated documents. They were the first to suggest applying neural methods to cross-lingual summarization (in their case, to perform ranking).

7.2.2 Multilingual Summarization

It is important to draw a distinction between *cross-lingual* and *multilingual* summarization. While these two terms were originally both used to refer to the task of producing a summary in a different language from its document, multilingual summarization has come to mean something else: a single system that is capable of summarizing documents written in multiple languages, but whose summaries are written in the same language as their documents. Multilingual systems generally make use of language-independent extractive summarization techniques, such as topic-based or graph-based approaches, with some language-dependent preprocessing, such as word segmentation for Chinese. Some notable multilingual summarization systems include Radev et al. (2002), Mihalcea (2005), and Litvak, Last, and Friedman (2010). Multilingual summarization is the focus of the MultiLing shared task (Giannakopoulos et al. 2011; Giannakopoulos 2013; Giannakopoulos et al. 2015; Giannakopoulos et al. 2017).

Two notable systems that used the term "multilingual summarization" but would be considered *cross-lingual* systems based on this new distinction between the two names are Chen and Lin (2000) and Evans, Klavans, and McKeown (2004); both were also multi-document systems that used a clustering approach. Chen and Lin separately clustered English sentences and Chinese sentences, producing two sets of monolingual sentence

clusters. They then translated the nouns and verbs in the Chinese clusters into English and used word overlap to identify which English cluster was most similar to each Chinese cluster. The clusters were merged to create bilingual clusters, from which either an English or Chinese summary could be produced.

Evans et al. produced English summaries for non-English documents by translating non-English documents into English. These translated sentences were clustered alongside the originally English sentences; because clustering is based on the words that are present in a sentence, disfluencies do not pose a problem in forming the clusters. Evans et al. addressed the possibility of disfluent translations by preferring to extract sentences from originally English documents, rather than translated sentences.

7.2.3 Abstractive Summarization

Neural abstractive summarization is a booming area of research. In this section, we do not give an exhaustive list of work in this area, but briefly review approaches that inspired or are otherwise related to the work described in this chapter.

Rush, Chopra, and Weston (2015) presented the first neural abstractive summarization model, a convolutional neural network encoder and feed-forward network decoder with attention, which learned to generate news headlines from the lead sentences of their articles; Chopra, Auli, and Rush (2016) extended their work using a recurrent network for the decoder. Gu et al. (2016) used an RNN for both the encoder and decoder and added a pointer-network (Vinyals, Fortunato, and Jaitly 2015) to allow copying of rare or out-of-vocabulary words from the input document. Nallapati et al. (2016) improved on Gu et

al.'s "copy net" by adding linguistically-motivated part of speech and named entity type embeddings and document-level attention, and See, Liu, and Manning (2017) added a coverage vector and coverage penalty to prevent repetition in generated words. In the work described in this chapter, we used See et al.'s definition of the "pointer-generator" model.

The New York Times annotated corpus (Sandhaus 2008), which we use both in this chapter and in Chapter 6, had previously been used for neural abstractive summarization by Paulus, Xiong, and Socher (2018) and Celikyilmaz et al. (2018). Both used more complex models than we do in this chapter: Paulus et al. added attention over the decoder's previous predictions to both prevent repetition and maintain coherence in longer generated summaries, while Celikyilmaz et al. trained multiple, collaborating encoders, one for each paragraph in a document, to encode long documents in their entirety, rather than truncating them.

7.3 Data

To train a cross-lingual summarization system for low-resource languages, following the translate-then-summarize approach, we need data consisting of input documents written in a non-English language, and their English translations, paired with reference summaries written in English. Unfortunately, only one such dataset exists: the DUC 2004 Task 3 dataset. The DUC 2004 test set provides both high-quality, human translations and lower-quality, automatic translations of Arabic news articles, paired with human-written, abstractive English summaries. However, it is too small to train an abstractive summarization system – there are only 480 document-summary pairs, and the summaries are very short,

with a maximum length of ten words. Thus, while we can and do use the DUC 2004 test set to evaluate our cross-lingual summarization system, we need another source of data on which to train.

Fortunately, we do not need genuine translations of actual non-English input documents for training. Following the translate-then-summarize approach, our summarization system would never encounter the non-English documents; we would have translated them into English first. Thus, to train the summarization system, all we need are the sort of disfluent, errorful documents that are likely to be the output of a machine translation system for a low-resource language.

We create simulated errorful translations by automatically translating English input documents into three low-resource languages, Somali, Swahili, and Tagalog, and then automatically translating them back into English, accumulating translation errors and disfluencies characteristic of the particular language pair in both directions. The new, noisy English input documents remain paired with their original, clean English reference summaries, producing a training corpus of synthetic "translations."

We use the New York Times (NYT) annotated corpus (Sandhaus 2008), which consists of 654,759 articles paired with human-written abstractive summaries. We follow the train/validation/test split used by Paulus, Xiong, and Socher (2018) and Celikyilmaz et al. (2018): we sort the article/summary pairs in chronological order and use the first 90% (589,284 pairs) for training, the next 5% (32,736 pairs) for validation, and the final 5% (32,739) for testing. We also follow some of Paulus et al.'s preprocessing steps:

- Convert tokens to lower case;

- Replace numbers with “0”;
- Remove “(s)” and “(m)” marks from summaries;
- Remove the words “photo,” “graph,” “chart,” “map,” “table,” and “drawing” at the ends of summaries;
- Replace semicolons with periods in summaries.

We differ from Paulus et al.’s preprocessing steps in the following ways:

- We do not anonymize named entities. Named entity anonymization consists of replacing proper nouns, which tend to be relatively rare vocabulary items, with a special entity token. However, it is difficult to implement in our case. Anonymization before the round-trip translation process could result in the machine translation system mistranslating or deleting the special entity token, while waiting until after the translation process could result in named entities no longer being automatically identifiable due to the added noise.

- We do not include headlines and bylines. Paulus et al. used special separator tokens to indicate which sentences were the article’s headline and byline, but we simply delete the headline and byline from our articles because we do not intend to evaluate our system solely on the NYT test set. To do so would mean testing solely on synthetic data; we intend to test on real-world Somali, Swahili, and Tagalog documents, in addition to the DUC 2004 Arabic articles, which do not necessarily follow Paulus et al.’s headline/byline formatting.

Due to the computational expense of translating the NYT corpus into three different

languages and back to English, we focus on a subset of 112k articles. Of these, 100k are taken from the training set, 6k from the validation set, and 6k from the test set. Half of the articles are taken from the NYT Opinion section, and the other half are selected randomly from the other sections. We target the Opinion section in particular because the reference summaries tend to be more abstracted than those of other sections; many are *indicative* summaries, which tell the reader what a document is generally about, rather than the *informative* summaries more commonly studied in automatic summarization research (including in the preceding chapters of this thesis). Emulating these highly abstractive, indicative summaries seems to us a good strategy for handling errorful input documents from which we cannot copy directly.

We translate the 112k articles into each of our three low-resource languages, Somali, Swahili, and Tagalog[1]; we then translate the articles back into errorful English. Figure 7.1 shows an example of a synthetic noisy English "translation." While this synthetic "translation" is not as garbled as the real-world translation that we show in Section 7.5.2, it is noticeably disfluent.

in the editor: why did president clinton continue to praise a program on welfare-to-work that failed in half of those assigned? in his comments, he praised the consultation of the community of kansas city, but was advised by gary j. stangler, director of the department of social service of missouri, which half of the participants failed. where are these people helping each other when the government cut them? back to the pantry of food. bad news, mr. president. the charity of the community will not help everyone who will come to us for help. glenn classic valley park, mo.

Figure 7.1: A synthetic errorful English article created by translating the original article into Tagalog and back.

We pair each noisy English "translation" with the clean English reference summary

[1]The neural machine translation system we use is described in the next section.

corresponding to the original, clean English article. For simplicity, we refer to the corpus created by translating into Somali and back as the *Somali NYT corpus*, and similarly with Swahili and Tagalog, but all three corpora consist of errorful English input documents, not Somali, Swahili, or Tagalog documents. Each of the three language-specific corpora, Somali NYT, Swahili NYT, and Tagalog NYT, is generated from the same 112k original articles.

7.4 Models

7.4.1 Machine Translation

We use neural machine translation systems developed at the University of Edinburgh to translate the NYT corpus into Somali, Swahili, and Tagalog, and back to English. The systems are built on the Marian framework (Junczys-Dowmunt et al. 2018). Table 7.1 shows the performance of the machine translation systems for each of the three language pairs on human-curated test sets of 500 parallel sentences from the IARPA MATERIAL project.

	BLEU	
Language	from English	to English
Somali	21.8	29.4
Swahili	44.5	37.8
Tagalog	37.2	36.2

Table 7.1: Neural machine translation performance.

The Somali systems (Somali-to-English and English-to-Somali) are transformer networks (Vaswani et al. 2017) trained on a mix of real-world parallel data, some human-

curated (23k sentences from the IARPA MATERIAL project and 46k from the DARPA LORELEI project) and some web-crawled (354k sentences from the ParaCrawl corpus), as well as just under two million sentences of synthetic parallel data generated by automatically translating CommonCrawl, NewsCrawl, and Wikipedia data from English into Somali and vice versa. The synthetic sentences, which were generated from source sentences by one neural machine translation model, were rescored using the word-normalized conditional cross-entropy of their source sentences under a neural machine translation model for the other direction (i.e., synthetic Somali sentences, which were generated from English source sentences using an English-to-Somali model, were rescored under a Somali-to-English model). Synthetic sentences with low source sentence cross-entropy scores were filtered out and not used for training. The real-world parallel data was oversampled three or five times (Somali-to-English and English-to-Somali, respectively).

The Swahili and Tagalog systems are ensembles of four BiDeep recurrent neural networks (Barone et al. 2017). The systems were trained on 24k and 51k sentences, respectively, of real-world parallel data from IARPA MATERIAL, as well as one million synthetic Swahili or Tagalog sentences automatically translated from English NewsCrawl data and 600k synthetic English sentences automatically translated from Swahili and Tagalog CommonCrawl data. The real-world parallel data was oversampled four or six times (to and from English, respectively).

7.4.2 Abstractive Summarization

For our abstractive summarization systems, we implement See, Liu, and Manning (2017)'s pointer-generator network in PyTorch (Paszke et al. 2017). We pre-train for 12 epochs on the full NYT training set (589,284 clean English document/summary pairs, as opposed to the 100k pairs in our synthetic "translation" training sets) to obtain a baseline system, using early stopping based on performance on the validation set; Table 7.2 shows the performance of this baseline on the NYT test set of 32,739 documents.

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT Baseline	**48.26**	29.30	36.81
Paulus et al.	47.03	30.51	**43.27**
Celikyilmaz et al.	48.08	**31.19**	42.33

Table 7.2: Baseline abstractive summarizer performance.

The baseline underperforms the more complex systems of Paulus, Xiong, and Socher (2018) and Celikyilmaz et al. (2018) on ROUGE-2 and ROUGE-L, but not on ROUGE-1. This is expected – although the pointer-generator has not previously been evaluated on the NYT corpus, it was outperformed by Paulus et al. and Celikyilmaz et al. on the CNN/Daily Mail corpus. The advantage of using the pointer-generator is that the simpler network is faster to train; See et al.'s reported training time on the CNN/Daily Mail corpus was roughly 60% of Celikyilmaz et al.'s reported training time. Having already invested a great deal of time creating the synthetic "translation" training sets, we choose to use the simpler, faster system; we are more interested in the improvements our cross-lingual approach can make over a baseline system than in the baseline's performance compared to state-of-the-art systems.

We use each of the three synthetic "translation" corpora to train a copy of the baseline system for another eight (Somali) or ten (Swahili and Tagalog) epochs, using early stopping based on performance on each language's validation set of 6k document/summary pairs. We train three language-specific cross-lingual abstractive summarization systems, and also a fourth, mixed-language system using 100k documents evenly and randomly selected from the Somali, Swahili, and Tagalog training sets. Thus, all four cross-lingual systems are trained on the same 100k document/summary pairs, with the only difference among systems being the language-specific noise added to the "translated" documents; all 100k training documents, in their original, clean English forms, had also already been used to train the baseline system.

7.5 Evaluation

7.5.1 Synthetic NYT "Translations"

Evaluating on the synthetic "translated" NYT test sets is more of a sanity check than a true evaluation; evaluating on test data created in the same way that our training and validation data were created can only tell us that our cross-lingual summarizers have learned to handle our synthetic "translated" data, but not necessarily that they can handle real-world translations. Despite this limitation, the sanity check evaluation results are promising.

Table 7.3 shows the performance of our cross-lingual abstractive summarizers on the synthetic Somali, Swahili, and Tagalog NYT test sets. Differences among the three language-specific models are not statistically significant, and the more general mixed model achieves

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT Baseline	32.94	10.36	22.51
Cross-lingual Somali*	37.72	15.39	26.56
Cross-lingual Swahili* †	37.26	14.94	25.92
Cross-lingual Tagalog* †	36.89	14.41	25.53
Cross-lingual Mixed*	**38.07**	**15.76**	**26.82**

(a) Performance on Somali NYT test set.

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT Baseline	35.28	12.96	25.64
Cross-lingual Somali* †	38.42	16.34	29.06
Cross-lingual Swahili*	39.24	17.01	29.88
Cross-lingual Tagalog* †	38.24	16.02	28.79
Cross-lingual Mixed*	**39.96**	**17.56**	**30.24**

(b) Performance on Swahili NYT test set.

Model	ROUGE-1	ROUGE-2	ROUGE-L
NYT Baseline	37.17	14.67	27.26
Cross-lingual Somali* †	38.97	17.01	29.16
Cross-lingual Swahili* †	39.14	17.28	29.43
Cross-lingual Tagalog*	**40.96**	18.72	31.06
Cross-lingual Mixed*	40.87	**18.91**	**31.14**

(c) Performance on Tagalog NYT test set.

Table 7.3: * indicates significant improvement over NYT Baseline ($p < 1.16 \times 10^{-19}$); † indicates significant difference between the mixed model and the language-specific models ($p < 0.05$).

the best scores overall. However, we find that summarizers trained solely on one language and tested on another language significantly underperform the mixed model ($p < 0.05$), which is trained on all three languages. That is, the differences between the mixed model and the two language-specific models *not* trained on a given test language are significant; for example, the difference between Cross-lingual Mixed and Cross-lingual Swahili and the difference between Cross-lingual Mixed and Cross-lingual Tagalog are significant on the Somali test set. The mixed model is trained on the same amount of data as each of the language-specific models, and differs from them only in that it has access to multiple

types of language-specific noise. This suggests that training on different types of language-specific noise generally improves performance, but training on some same-language data is still important.

	Perplexity		
Model	Somali NYT	Swahili NYT	Tagalog NYT
NYT Baseline	4986	4428	4707
Cross-lingual Somali	**3357**	3429	3528
Cross-lingual Swahili	3384	**3247**	**3312**
Cross-lingual Tagalog	3501	3476	3457
Cross-lingual Mixed	3464	3285	3402

Table 7.4: Language model perplexity of generated summaries on Somali, Swahili, and Tagalog NYT test sets.

We also train a bigram language model on the entire set of NYT reference summaries and calculate the average perplexity of our cross-lingual summarizers' output on the Somali, Swahili, and Tagalog test sets as a proxy for fluency (Table 7.4). We find that Somali is the most difficult source language overall, which is unsurprising given the relatively poor performance of the Somali neural machine translation systems, compared with those for Swahili and Tagalog (Table 7.1). However, all three language-specific models and the mixed model produce more fluent English, regardless of source language, than does the base model.

7.5.2 Real-World Weblog Evaluation

In order to evaluate our cross-lingual summarization systems on real-world Somali, Swahili, and Tagalog documents, we perform a human evaluation on twenty Somali, twenty Swahili,

and twenty Tagalog weblog entries from the IARPA MATERIAL project[2]. We translate the documents into English using the same neural machine translation systems we used to create our noisy NYT "translation" corpora, but unlike our synthetic NYT data, which we translated from English into the low-resource languages, these weblog entries are real-world, naturally occurring Somali, Swahili, and Tagalog documents. Thus, this evaluation demonstrates the performance of our cross-lingual summarizers in a real use-case.

Document: mange kimambi 'i pray for the parliamentary seat for kinondoni constituency for ticket of ccm. not special seats' kinondoni without drugs is possible i pray for the parliamentary seat for kinondoni constituency on the ticket of ccm. yes, it's not a special seats, khuini kinondoni, what will i do for kinondoni? tension is many i get but we must remember no good that is available easily. kinondoni without drugs is possible. as a friend, fan or patriotism i urge you to grant your contribution to the situation and propert. you can use western union or money to go to mange john kimambi. account of crdb bank is on blog. reduce my profile in my blog understand why i have decided to vie for kinondoni constituency. you will understand more.

NYT Baseline: mange kimambi, who pray for parliamentary seat for kinondoni constituency for ticket of ccm , is on blog, and not special seats' kinondoni without drugs.

Cross-lingual Mixed: mange kimambi, who pray for parliamentary seat for kinondoni constituency for ticket of ccm, comments on his plans to vie for 'kinondoni' without drugs.

Figure 7.2: A Swahili weblog entry, automatically translated into English, and its baseline and mixed model summaries.

Figure 7.2 shows a translated Swahili weblog entry and its NYT Baseline and Cross-lingual Mixed summaries[3]; phrases that are copied from the input document are color-coded. This example shows the advantage of our approach: unlike a machine translation system, which must translate every part of its input, our summarizer is able to delete most of the long, rambling, and disfluent weblog entry, summing it up fluently with the purely

[2]These documents are taken from one of the held-out test sets for the project.

[3]All four cross-lingual summarizers produce very similar summaries on this document.

generated phrase “comments on his plans” and the completely repurposed phrase “to vie for”.

Compared with the baseline system, which is trained on the full NYT corpus, the cross-lingual system, which is trained on a subset of the NYT corpus that is balanced between the Opinion section and the other sections, produces a more *indicative* summary: it describes what the document is about (Mange Kimambi’s campaign platform), rather than attempting to compress the important information in the document. While the baseline summarizer was also trained on some Opinion document/summary pairs, the Opinion section contains only 70,995 total document/summary pairs, compared to 583,764 pairs from other sections, and so it behaves more like an informative news summarizer than an indicative editorial summarizer: it mainly copies from the beginning of the document, with only a small amount of reordering and rephrasing.

We used five human judges, all native English speakers. The judges were shown a translated document and one of the summaries generated for that document and asked to rate the summary’s *content* (i.e., informational correctness and faithfulness to the document) and *fluency* (i.e., grammatical correctness and pleasantness to read) on a scale of 1–3 (Table 7.5). Each judge was shown two different summaries for each of the sixty documents in this evaluation, for a total of twenty-four summaries from each of the four cross-lingual models and the baseline; each document/summary pair was shown to two different judges, and each judge overlapped with every other judge on thirty document/summary pairs.

Our human judges rated the cross-lingual summarization systems higher in both fluency and content than the baseline system, and we find again that while the language-specific systems are more fluent on their own languages than are the language-specific systems for

Somali Weblogs		
Model	Content	Fluency
NYT Baseline	1.66	1.62
Cross-lingual Somali	1.92	1.90
Cross-lingual Swahili	1.94	1.88
Cross-lingual Tagalog	1.86	1.82
Cross-lingual Mixed	**2.08**	**2.04**
Swahili Weblogs		
Model	Content	Fluency
NYT Baseline	1.88	1.76
Cross-lingual Somali	2.14	1.90
Cross-lingual Swahili	2.22	**2.08**
Cross-lingual Tagalog	2.18	1.86
Cross-lingual Mixed	**2.36**	**2.08**
Tagalog Weblogs		
Model	Content	Fluency
NYT Baseline	1.72	1.76
Cross-lingual Somali	1.76	1.88
Cross-lingual Swahili	1.94	1.92
Cross-lingual Tagalog	1.80	2.08
Cross-lingual Mixed	**2.08**	**2.16**

Table 7.5: Average human-rated content and fluency scores on Somali, Swahili, and Tagalog weblog entries.

the other languages, the mixed model still performs the best. We also find that, while our improvement in content is more modest, our improvement in fluency is large. The judges achieved substantial agreement on their ratings (Fleiss's $\kappa = 0.72$).

7.5.3 DUC 2004 Arabic Evaluation

Given the computational expense of creating a training corpus of errorful "translations" for a low-resource language, and given the mixed-language model's strong performance, a natural question that arises from this work is whether or not the cross-lingual summarizers

can generalize across languages. To answer this question, we perform a final evaluation on a new, unseen language: Arabic. We use the DUC 2004 Task 3 test set, which we briefly described in Section 7.3.

The DUC 2004 Task 3 test set consists of 25 clusters of topically-related Arabic news articles taken from the Agence France Presse and the Arabic Newswire corpus, totaling of 480 articles; the clusters were used for multi-document summarization in Task 4, while the articles were used for single-document summarization in Task 3. Three different sets of English translations were provided: human translations from the Linguistic Data Consortium (LDC) and two sets of machine translations provided by IBM and the Information Sciences Institute (ISI)[4]. Human annotators wrote ten-word reference summaries for each article, using the human translations from the LDC.

Document: washington 10-23 (afp) was signed by benjamin netanyahu and yasser arafat on friday at the white house agreed on the israeli military withdrawal from the west bank in return for palestinian additional security guarantees.

NYT Baseline: washington 10-23, signed by benjamin netanyahu and yasser arafat, agrees on israeli military withdrawal from west bank in return for palestinian additional security guarantees.

Cross-lingual Mixed: benjamin netanyahu and yasser arafat agree on israeli military withdrawal from west bank in return for palestinian security guarantees.

Figure 7.3: An Arabic article, automatically translated into English, and its baseline and mixed model summaries.

Figure 7.3 shows a machine-translated Arabic document (the ISI translation) and its NYT Baseline and Cross-lingual Mixed summaries[5]; phrases that are copied from the input document are color-coded. While the difference between the baseline and mixed models is

[4]Both IBM and ISI provided multiple machine translation variants; we use only the highest-scoring translation for each article.

[5]All four cross-lingual summarizers produce identical summaries on this document.

less dramatic in this example than in the weblog example in the previous section, the mixed model clearly produces a fluent summary, while the baseline does not. NYT Baseline again copies heavily from the beginning of the document: the unnecessary "washington 10-23" (shown in red) and the phrase "signed by" (shown in green), whose subject is missing.

Model	ROUGE-1	ROUGE-2	ROUGE-L
DUC 2004 Best MT	21.41	5.04	17.14
DUC 2004 Summarize-Translate	23.56	5.16	20.72
DUC 2004 Best Human	**25.87**	**7.63**	**21.98**
NYT Baseline	26.56	5.86	15.76
Cross-lingual Somali*	28.64	6.66	19.62
Cross-lingual Swahili* †	28.08	6.39	18.36
Cross-lingual Tagalog* †	**29.43**	**7.02**	**19.89**
Cross-lingual Mixed*	28.79	6.74	19.79

Table 7.6: DUC 2004 with ISI translations. * indicates significant improvement over NYT Baseline ($p < 2.09 \times 10^{-6}$); † indicates significant difference between cross-lingual models ($p < 0.05$).

Table 7.6 shows the performance of our cross-lingual abstractive summarizers on the DUC 2004 Task 3 data, demonstrating their ability to generalize and improve the fluency of input documents automatically translated from a previously unseen language, yielding a significant improvement in ROUGE. Task 3 had two subtasks: summarizing the machine-translated documents and summarizing the human-translated documents. There were ten submitted systems, each of which performed and was evaluated on both subtasks. An additional submitted system took the summarize-then-translate approach instead, summarizing the documents in Arabic and translating the summaries into English.

Table 7.6 shows the ROUGE scores of best-performing DUC 2004 system in each category – the row "DUC 2004 Best MT," for example, does not show the results of a single system, but rather the best scores that any DUC 2004 submitted system achieved using

machine-translated input documents. Compared to the ten DUC 2004 systems, our baseline performs reasonably well – some of the performance gain of our systems over the DUC 2004 systems must be due to the fifteen years of general advances in summarization research that separate us. However, we still find that the cross-lingual abstractive summarizers give improved performance over the baseline, despite never having been trained on Arabic data, suggesting that our approach does generalize to new and unseen languages. Our cross-lingual abstractive summarizers would have ranked first on summarizing the machine-translated documents. Further, despite our use of these lower-quality translations, we would have performed extremely well even in comparison with the DUC 2004 systems on high-quality, human-translations – we would have ranked first on ROUGE-1, fourth on ROUGE-2, and fifth on ROUGE-L.

7.6 Conclusion

The main contributions described in this chapter (based on work published in Ouyang, Song, and McKeown (2019)) are as follows:

- We develop a new approach for training cross-lingual abstractive summarization systems for low resource languages. For such languages, there would be no training data available for summarization systems, and as a result, documents must first be translated into English, likely by relatively poor-performing machine translation systems, which would also have suffered from a lack of training data. Our new approach provides a potential summarization solution for thousands of such low-resource languages.

- We create summarization corpora simulating automatically translated documents from three low-resource languages: Somali, Swahili, and Tagalog. Our corpora consist of noisy English "translations" paired with clean English reference summaries. Our method for producing these corpora can be easily applied to new source-target language pairs.

- By training on simulated "translated" input and clean reference summaries, many of which were highly abstracted, indicative summaries, we outperform a standard copy-attention abstractive summarizer on real-world Somali, Swahili, and Tagalog documents. Our cross-lingual abstractive summarizers are rated more highly in both content and fluency by human judges.

- Our evaluation on Arabic documents demonstrates that our cross-lingual summarization systems are robust and can generalize to new, unseen languages. Our evaluations show that while training a system for a specific source language gave the strongest performance, our cross-lingual systems can improve summarization performance for any source language.

The key advantage of our approach is in making use of an abstractive summarization system's ability to delete phrases that were translated awkwardly or incorrectly, and to generate new text to use instead. The approach is also general enough to apply to any source-target language pair, and it can be adapted to new languages very easily. The main limitation is that our approach assumes the existence of a machine translation system for the language pair, which may not be the case for extremely low-resource languages. Although our systems are able to handle errorful, disfluent translations, they do not perform the

full cross-lingual summarization task directly; they cannot take a non-English document as input directly. In the case where there are no translations of any kind available, our systems would not work; in such a case, a language-independent approach or a language-pair-specific approach, such as cross-lingual word embeddings, would be necessary.

A relatively new and very interesting idea that contrasts with this work is the problem of *summary faithfulness*. Cao et al. (2018) observed that, since abstractive summarization systems rewrite their input sentences, there is nothing preventing them from producing summaries that are factually incorrect. In their experiments on the Gigaword corpus, previously used by Rush et al., Chopra et al., and Nallapati et al., Cao et al. found that almost 30% of the summaries generated by a basic pointer-generator-style abstractive summarizer were factually incorrect, often due to mistaking a word that happened to appeared near a predicate in the input sentence for subject of that predicate.

Summary faithfulness is a serious concern, particularly as in this work, we want our abstractive summarizer to heavily rewrite its input sentences. In the example shown in Figure 7.2, our summarizer goes so far as to invent four words out of whole cloth – words that are not present in the input document at all. In our experiments, our approach seems to have avoided the problem of factually incorrect summaries; our systems are rated slightly higher in correctness and faithfulness than is the baseline pointer-generator model in our human weblog evaluation. Still, the problem of faithfulness is one to keep in mind, even if we do not encounter it in these experiments – Cao et al.'s proposed solution of encoding dependency relations found in the input document, in addition to the document itself, could not be applied to the disfluent and errorful translations for which our approach is designed.

Conclusion

With the tremendous amounts of text published on the Web every day, automatic text summarization is more relevant than ever. As the Web has changed how people seek out and interact with information, so must summarization techniques change and adapt to these new sources of information. In this thesis, we have developed novel approaches for automatic summarization of two new sources of information: personal narratives and non-English documents.

Contributions

Our main contributions are as follows:

A summarization corpus for personal narrative.

Our Reddit summarization corpus is unique in three respects. First, it is the only summarization corpus for the genre of personal narrative. While other collections of personal narratives exist, they are of significantly lower quality; our method for collecting narratives from Reddit achieved 94% precision on extracting only narrative text, compared to Gordon, Cao, and Swanson (2007)'s 50% and Gordon and Swanson (2009)'s 66%. Our corpus also includes meta information not found in other collections of narrative: the topical prompt that elicited a narrative, readers' comments, and TL;DR short summaries – we were the first to use the TL;DR for summarization, predating the Webis-TLDR-17 corpus (Völske

et al. 2017) by two years.

Second, our Reddit summarization corpus is unusual among summarization corpora in that it not only provides both human-written abstractive summaries and human-selected extractive summaries, but that the extractive summaries are explicitly constructed to match the abstractive summaries as closely as possible and are aligned back to the abstractive summaries. Summarization corpora that provide both abstractive and extractive summaries are already uncommon; only one other corpus, the ISCI Meeting Corpus (Murray et al. 2005), provides paired and aligned abstractive and extractive summaries. There has been recent interest in two-stage, extract-then-abstract summarization (Nallapati, Zhai, and Zhou 2017; Liu et al. 2018; Gehrmann, Deng, and Rush 2018), and although our corpus is not large enough to train a neural summarizer, it allowed us to perform two-stage summarization using a mix of non-neural techniques and neural models with other distantly-labeled datasets.

Finally, our extractive-abstractive summary alignments are annotated with the six summary rewriting operations identified by Jing and McKeown (1999). Our Reddit summarization corpus captures interactions among rewriting operations, rather than focusing on only one operation in isolation. The presence of multiple different rewriting operation annotations in each alignment allows the corpus to serve as a very challenging dataset for research on compression, fusion, reordering, generalization, specification, and paraphrasing.

A narratology-inspired approach to extractive summarization.

The success of our change-based approach to extractively summarizing personal narrative validated the contextualist theory that authors draw attention to important content in their narratives by varying their writing style. Our work shows that this theory can be empirically verified: using only stylistic and affectual features to compare sentences within a narrative to their neighbors, we can learn to make sentence selection decisions that match those of human extractive summarizers. The change-based approach outperforms traditional extractive summarization techniques designed for the news genre, demonstrating the importance of tailoring summarization approaches to one's genre of input and goals for output.

An investigation of paraphrasing for summarization.

We conduct the first experiments on paraphrasing specifically in the context of abstractive summarization. Prior work on paraphrase generation has resorted to a "shotgun" approach of generating as many paraphrases as possible, in the hopes that one will be useful, but our evaluations quantify the profound limitations of this approach. No prior work has attempted to learn which inputs ought to be paraphrased, and which ought to be left alone; it was assumed that if a sentence could be paraphrased, then it should be. Our two-stage, classify-then-paraphrase approach shows that, not only is it possible to identify which document sentences should or should not be paraphrased for summarization, but doing so improves the performance of the downstream paraphrasing system by removing spurious inputs before it encounters them.

Our analysis of the affectual and stylistic features we used to augment the word embeddings in our neural classifier and paraphraser also represent the first attempt to learn why human summarizers perform paraphrasing: to modify the activeness and subjectivity – and to a lesser extent, the pleasantness, imagery, formality, and complexity – of a document sentence into something more appropriate for a summary.

An alignment system for sentential quasi-paraphrases.

Our pointer-network-based alignment system is one of only six phrase-based monolingual aligners in existence. It is specifically designed to be able to align the longer and more lexically diverse paraphrases that have been neglected in prior work, and it is the first to perform alignment using a neural network.

The pointer-aligner can align phrases that mean approximately, rather than exactly, the same thing, in constrast with previous approaches, which aligned shared words or close synonym pairs. We represent each phrase with a learned phrase embedding, which captures the general meaning of the phrase as a whole, rather than representing each phrase by the individual words it contains; a difference of a single function word, for example, would have little to no effect on the phrase embedding level. Our use of phrase embeddings also allows the pointer-aligner to align phrases of any length. Where previous approaches were limited due to the computational expense of aligning long phrases, our system does not distinguish between full sentences and single words; both are represented by a single embedding.

While it is able to align very long phrases, the pointer-aligner is not required to do so.

Unlike previous approaches, where the maximum allowable phrase length must be set by the user, or where the system must commit to a single arbitrary chunking of the input, we are able to discover the optimal phrase size during alignment. By aligning multiple different chunkings of the input and allowing the different alignments to vote on the final output, the pointer-aligner allows high-confidence alignments between phrases of the optimal size to outvote low-confidence alignments between phrases of sub-optimal sizes.

The success of our approach, even when pared down for faster run times, on the paraphrases in the Reddit summarization corpus and on both the *sure* and *possible* alignments in the MSR RTE corpus (Brockett 2007), proves that the long, lexically diverse paraphrases that were ignored in prior work can be aligned.

A cross-lingual summarization pipeline for low-resource languages.

We develop cross-lingual, abstractive summarization systems for three low-resource languages, Somali, Swahili, and Tagalog, for which no summarization systems of any kind, cross-lingual or otherwise, previously existed. Our approach of creating synthetic, low-quality "translations" for training can be applied to any low-resource source language/high-resource target language pair. However, as long as English remains the target language, our evaluation on Arabic shows that, while retraining the summarizer for a new source language may give the best performance, it is not necessary to do so; the cross-lingual system trained on three languages is already robust enough to perform well on unseen languages.

Limitations

The limitations of the work described in this thesis are as follows:

The Reddit summarization corpus alignments and rewriting annotations are imprecise.

The main difficulty we encounter in lexical paraphrasing is that the alignments in the Reddit summarization corpus can be of any size. Many are full clauses, or even sentences, and as a result, they contain multiple other rewriting operations, in addition to lexical paraphrasing. While capturing the presence of multiple rewriting operations is a feature of the corpus, the operations are difficult to disentangle in downstream tasks. An additional layer of annotations indicating precisely which word(s) in an alignment demonstrate a given rewriting operation may have been useful, although, having looked at a great many examples over the course of the work described in this thesis, we are not confident that it is always possible to identify where exactly an operation applies.

The pointer-aligner is computationally expensive.

Towards the end of Chapter 5, we analyzed the run time of our paraphrase alignment system. While the pointer-aligner's run time is independent of the lengths of the phrases to be aligned – which was a key design goal – it remains bound by the lengths of the input sentences. On long sentences of around 70 words, the full pointer-aligner takes several hours to run. The pared-down version of the aligner that we used in Chapter 6 trades performance for speed; its aligned word pair precision and recall are both lower than those of the full

system, but it takes only seconds to align input sentences, regardless of their length. However, one source of computational overhead that cannot be disentangled from our approach is the need to perform constituent parsing on all input sentences. While it is possible to arbitrarily chunk the input sentences without using a constituent parse (e.g., by imposing a chunk boundary every so many words), we have seen in our experiments in Chapter 6 that sentence structure does matter to neural encoders, so we suspect that arbitrary chunkings would perform poorly.

The cross-lingual summarization system assumes the availability of machine translation.

Although our cross-lingual abstractive summarization system is able to handle errorful, disfluent translations, for extremely low-resource languages there may be no machine translations of any kind available. In such a situation, our system will not work; it requires that the input document be in some form of English, even if it is poor English. Similarly, for source/target language pairs for which machine translations are not available, our approach and training pipeline will not work. Another approach, such as summarization using cross-lingual word embeddings, would be necessary.

Two-stage approaches cannot recover from errors in the first stage.

The final limitation is a more general one. We use two different two-stage approaches in this thesis: first, we used an extract-then-abstract summarization approach over the course of most of this thesis; second, we used a classify-then-paraphrase approach to lexical para-

phrasing for summarization in Chapter 6. The limitations of the two-stage approach can be seen in the examples shown in Chapter 6, which not only contained a two-stage approach of its own, but was the second stage of the overarching two-stage approach covering the bulk of this thesis.

The difficulty with two-stage approaches is that errors in the first stage are propagated to, and often cannot be recovered by, the second stage. For example, a character in a narrative may be referenced several times but named only once; if the sentence containing the character's name is not extracted during the first stage of summarization, the abstractive second-stage system will never see it. In this particular example, some post-processing after the first stage would help: performing coreference resolution on the input document and replacing pronouns in selected extractive summary sentences with their antecedents would solve this particular problem.

However, not all first-stage errors can be so easily addressed. In the case of the classify-then-paraphrase approach, the sentence-level paraphrasing evaluations shown in Chapter 6 used only the gold standard paraphrasing input sentences. If we evaluate instead on the input sentences classified as paraphrasing by the first-stage classifier (reverting to the end-to-end evaluation procedure of comparing paraphrasing sentence outputs against their target paraphrases and non-paraphrasing sentence outputs against the input), we find that the word error rate increases by about 7.48, and ROUGE decreases by about 6.52 across the board. There is no easy solution to the problem of classification error; there is no way for the paraphrasing system to tell that the classification system has made an error in the first place. Similarly, if it is not a pronoun's antecedent that is omitted from the extractive summary, but an entire event, the second-stage abstractive summarization system cannot

recover the event unless it miraculously manages to invent it from whole cloth.

Future Directions

Some future directions for adapting automatic summarization to even more new sources of information are as follows:

Cross-Lingual Summarization.

Combining ideas from our work in both cross-lingual summarization and paraphrasing, we would like to make use of multiple translations of non-English input documents. We can obtain multiple translations of a document, not only by using multiple machine translation systems, but also by taking advantage of the n-best word candidate lists produced by neural machine translation systems. Different translation systems may have different strengths and weaknesses. For example, a statistical machine translation system that copies input words not found in its phrase table is better at preserving non-English named entities than are neural systems that simply delete unknown words without any indication to the user, but the neural systems generally produce more fluent English. Some ideas we hope to explore are

- applying multi-document summarization techniques to combine multiple translations into a single high-quality summary, and
- designing a summarization system that takes a lattice of candidate translations as input, rather than a single sequence, allowing it to learn which path will produce the best summary.

Query-Focused Summarization

People make use of query-focused extractive summaries every day, in the form of search engine results, where relevent web pages are accompanied by excerpts showing where the user's search term occurs in each web page. These query-focused extractive summaries are sometimes helpful, and sometimes not – we find the summaries for results from Stack Overflow, for example, to be distressingly unhelpful. To improve the coherence and usefulness of these summaries, we need better techniques for query-focused summarization.

While there has been recent interest in query-focused abstractive summarization Hasselqvist, Helmertz, and Kågebäck (2017) and Nema et al. (2017), the summarization part of query-focused summarization has been lost – recent work considers the query-focused abstractive summarization task to be simply question-answering with complete sentences. We believe that a query-focused summary should address the query, but it should also function as a reasonable summary of the whole document; to continue the search engine example, most queries are not specific questions, but rather topics that the user wants to know more about. Some ideas we hope to explore are

- using reinforcement learning to fine-tune a pre-trained abstractive summarization system to focus on user queries, without losing its ability to produce general-purpose summaries; and
- providing the query to the system in different ways, such as learning a query embedding, or extracting a topic based on both the query and the document and focusing on words from that topic.

Multimodal Summarization.

Besides new text genres, information on the Web is increasingly found in other media, such as videos or podcasts. Previous work on summarizing recorded speech has been split between two approaches: transcribe into text using automatic speech recognition and summarize the transcript, or extract speech segments directly from the audio signal. We hope to combine these two modes of information in order to make use of both what is said (the transcript) and how it is said (the audio) in creating a summary, and videos introduce a third medium that can likewise assist in selecting content for a summary.

A challenge of multimodal summarization is that, in a video or audio recording, there may be multiple speakers who interrupt or speak over each other, or who hold different or even opposing views. In this case, the traditional informative summary may not be the most useful format. When one speaker responds to another, for instance, he or she will likely refer to what the other person said, without repeating it; if such an utterance is excerpted for the summary, it will be impossible for a user to understand in isolation. We hope to experiment with different summary structures, such as highly-abstracted, indicative summaries, or separate summaries for each speaker, to better organize and present the content of multimodal documents.

User-Oriented Summarization.

Finally, in the course of our work on summarization, we have performed many human evaluations, and we have found that users are very sensitive to how information is presented – a user is likely to reject a poorly-written summary as useless, even if it contains all of

the information they need. How something is said can be just as important as what is said. From our work on lexical paraphrasing in Chapter 6, we discovered that 1) encoder-decoder models are good at learning sentence structure, and 2) they can also learn human summarizers' affectual and stylistic choices. Can we apply these abilities to customize summaries in other ways? Some ideas we hope to explore are

- customizing summaries for a particular dialect of English, and
- tailoring summaries for a particular user by customizing the text to mimic the writing style of that user's demographic group.

Bibliography

Agarwal, Apoorv, Fadi Biadsy, and Kathleen McKeown (2009). "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.

Avinesh, PVS, Maxime Peyrard, and Christian M Meyer (2018). "Live Blog Corpus for Summarization." In: *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Bannard, Collin and Chris Callison-Burch (2005). "Paraphrasing with bilingual parallel corpora." In: *Proceedings of 43th Annual Meeting of the Association for Computational Linguistics*.

Bar-Haim, Roy et al. (2006). "The Second PASCAL Recognizing Textual Entailment Challenge." In: *Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment*.

Barone, Antonio Valerio Miceli et al. (2017). "Deep architectures for Neural Machine Translation." In: *Proceedings of the Second Conference on Machine Translation*.

Baroni, Marco and Roberto Zamparelli (2010). "Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space." In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Barthes, Roland (1975). "An Introduction to the Structural Analysis of Narrative." In: *New Literary History*. Translated by Lionel Duisit.

Barzilay, Regina (2003). "Information Fusion for Multidocument Summerization: Paraphrasing and Generation." PhD thesis. Columbia University.

Barzilay, Regina and Michael Elhadad (1999). "Using lexical chains for text summarization." In: *Advances in Automatic Text Summarization*.

Barzilay, Regina and Noemie Elhadad (2003). "Sentence alignment for monolingual comparable corpora." In: *Proceedings of the 2003 Conference on Empirical Mmethods in Natural Language Processing*.

Barzilay, Regina and Lillian Lee (2003). "Learning to paraphrase: an unsupervised approach using multiple-sequence alignment." In: *Proceedings of the 2003 Conference*

of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.

Barzilay, Regina and Kathleen R McKeown (2001). "Extracting paraphrases from a parallel corpus." In: *Proceedings of the 39th annual meeting on Association for Computational Linguistics*.

— (2005). "Sentence fusion for multidocument news summarization." In: *Computational Linguistics*.

Barzilay, Regina, Kathleen R McKeown, and Michael Elhadad (1999). "Information fusion in the context of multi-document summarization." In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Bauer, John (2014). *Shift-Reduce Constituency Parser*. https://nlp.stanford.edu/software/srparser.html. Accessed: 2018-11-30.

Berg-Kirkpatrick, Taylor, Dan Gillick, and Dan Klein (2011). "Jointly learning to extract and compress." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*.

Bhagat, Rahul and Eduard Hovy (2013). "What is a paraphrase?" In: *Computational Linguistics*.

Botel, Morton and Alvin Granowsky (1972). "A formula for measuring syntactic complexity: A directional effort." In: *Elementary English*.

Brockett, Chris (2007). *Aligning the RTE 2006 corpus*. Tech. rep. Microsoft Research.

Cao, Ziqiang et al. (2018). "Faithful to the original: Fact aware neural abstractive summarization." In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Carenini, Giuseppe, Raymond T Ng, and Xiaodong Zhou (2007). "Summarizing email conversations with clue words." In: *Proceedings of the 16th International Conference on World Wide Web*.

Carletta, Jean et al. (2005). "The AMI meeting corpus: A pre-announcement." In: *International Workshop on Machine Learning for Multimodal Interaction*.

Celikyilmaz, Asli et al. (2018). "Deep Communicating Agents for Abstractive Summarization." In: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Chambers, Nathaniel and Dan Jurafsky (2008). "Unsupervised learning of narrative event chains." In: *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*.

Chen, Hsin-Hsi and Chuan-Jie Lin (2000). "A muitilingual news summarizer." In: *Proceedings of the 18th International Conference on Computational Linguistics*.

Cheng, Jianpeng and Mirella Lapata (2016). "Neural summarization by extracting sentences and words." In: *Proceedings of ACL 2016: Annual Meeting of the Association for Computational Linguistics*.

Chopra, Sumit, Michael Auli, and Alexander M Rush (2016). "Abstractive sentence summarization with attentive recurrent neural networks." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Cohn, Trevor, Chris Callison-Burch, and Mirella Lapata (2008). "Constructing corpora for the development and evaluation of paraphrase systems." In: *Computational Linguistics* 34.4.

Cohn, Trevor and Mirella Lapata (2008). "Sentence compression beyond word deletion." In: *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*.

Daumé III, Hal and Daniel Marcu (2004). "Generic Sentence Fusion is an Ill-Defined Summarization Task." In: *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

— (2006). "Bayesian query-focused summarization." In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*.

Dolan, Bill, Chris Quirk, and Chris Brockett (2004). "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources." In: *Proceedings of the International Conference on Computational Linguistics*.

Dolan, William B and Chris Brockett (2005). "Automatically constructing a corpus of sentential paraphrases." In: *Proceedings of the Third International Workshop on Paraphrasing*.

Duboue, Pablo and Jennifer Chu-Carroll (2006). "Answering the question you wish they had asked: The impact of paraphrasing for question answering." In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*.

Edmundson, Harold P (1969). "New methods in automatic extracting." In: *Journal of the ACM (JACM)* 16.

Elson, David K and Kathleen McKeown (2010). "Building a bank of semantically encoded narratives." In: *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Elson, David (2012). "DramaBank: Annotating Agency in Narrative Discourse." In: *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

Erkan, Günes and Dragomir R Radev (2004). "LexRank: Graph-based lexical centrality as salience in text summarization." In: *Journal of Artificial Intelligence Research*.

Evans, David Kirk, Judith L Klavans, and Kathleen R McKeown (2004). "Columbia Newsblaster: Multilingual news summarization on the web." In: *Demonstration Papers at HLT-NAACL 2004*.

Filippova, Katja 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*.

Filippova, Katja and Michael Strube (2008). "Sentence fusion via dependency graph compression." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Fivush, Robyn, Jennifer G. Bohanek, and Marshall Duke (2005). "The Intergenerational Self: Subjective Perspective and Family History." In: *Individual and Collective Self-Continuity*. Ed. by F Sani. New York, NY: Psychology Press.

Fuentes Fort, Maria, Enrique Alfonseca, and Horacio Rodríguez Hontoria (2007). "Support vector machines for query-focused summarization trained and evaluated on pyramid data." In: *Proceedings of the Annual Meeting of the Association for Compuational Linguistics, Companion Volume: Proceedings of the Demo and Poster Sessions*.

Ganitkevic, Jurij (2018). "Large-Scale Paraphrasing for Text-to-Text Generation." PhD thesis. Johns Hopkins University.

Ganitkevitch, Juri, 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*.

Gehrmann, Sebastian, Yuntian Deng, and Alexander Rush (2018). "Bottom-Up Abstractive Summarization." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Giannakopoulos, George (2013). "Multi-document multilingual summarization and evaluation tracks in acl 2013 MultiLing workshop." In: *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*.

Giannakopoulos, George et al. (2011). "TAC 2011 MultiLing pilot overview." In:

Giannakopoulos, George et al. (2015). "MultiLing 2015: multilingual summarization of single and multi-documents, on-line fora, and call-center conversations." In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Giannakopoulos, George et al. (2017). "MultiLing 2017 overview." In: *Proceedings of the MultiLing 2017 workshop on summarization and summary evaluation across source types and genres*.

Goldman, S.R., A.C. Graesser, and Paul van den Broek (1999). *Narrative Comprehension, Causality and Coherence: Essays in Honor of Tom Trabasso*. Mahwah, NJ: Lawrence Erlbaum.

Gong, Yihong and Xin Liu (2001). "Generic text summarization using relevance measure and latent semantic analysis." In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Gordon, Andrew S., Qun Cao, and Reid Swanson (2007). "Automated Story Capture From Internet Weblogs." In: *Proceedings of the 4th International Conference on Knowledge Capture*.

Gordon, Andrew, Cosmin Adrian Bejan, and Kenji Sagae (2011). "Commonsense Causal Reasoning Using Millions of Personal Stories." In: *Proceedings of the 25th AAAI Conference on Artificial Intelligence*.

Gordon, Andrew 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*.

Goyal, Amit, Ellen Riloff, and Hal Daumé III (2010). "Automatically producing plot unit representations for narrative text." In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Grefenstette, Edward et al. (2013). "Multi-step regression learning for compositional distributional semantics." In: *Proceedings of the 10th International Conference on Computational Semantics*.

Grusky, Max, Mor Naaman, and Yoav Artzi (2018). "NEWSROOM: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies." In: *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Gu, Jiatao et al. (2016). "Incorporating copying mechanism in sequence-to-sequence learning." In: *Proceddings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Guo, Weiwei and Mona Diab (2012). "Modeling sentences in the latent space." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Hasselqvist, Johan, Niklas Helmertz, and Mikael Kågebäck (2017). "Query-based abstractive summarization using neural networks." In: *arXiv preprint arXiv:1712.06100*.

Hermann, Karl Moritz et al. (2015). "Teaching machines to read and comprehend." In: *Advances in Neural Information Processing Systems 28*.

Hill, Felix et al. (2016). "Learning to understand phrases by embedding the dictionary." In: *Transactions of the Association for Computational Linguistics*.

Hokamp, Chris and Qun Liu (2017). "Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Hu, J Edward et al. (2019). "ParaBank: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation." In: *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*.

Ibrahim, Ali, Boris Katz, and Jimmy Lin (2003). "Extracting structural paraphrases from aligned monolingual corpora." In: *Proceedings of the Second International Workshop on Paraphrasing*.

Jacquemin, Christian, Judith L Klavans, and Evelyne Tzoukermann (1997). "Expansion of multi-word terms for indexing and retrieval using morphology and syntax." In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*.

Janin, Adam et al. (2003). "The ICSI meeting corpus." In: *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processings*.

Jing, Hongyan (2000). "Sentence reduction for automatic text summarization." In: *Sixth Applied Natural Language Processing Conference*.

Jing, Hongyan 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*.

Junczys-Dowmunt, Marcin et al. (2018). "Marian: Fast Neural Machine Translation in C++." In: *Proceedings of ACL 2018, System Demonstrations*.

Kajiwara, Tomoyuki and Mamoru Komachi (2016). "Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings." In: *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*.

Kauchak, David and Regina Barzilay (2006). "Paraphrasing for automatic evaluation." In: *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*.

Kedzie, Chris, Kathleen McKeown, and Hal Daume III (2018). "Content selection in deep learning models of summarization." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Knight, Kevin and Daniel Marcu (2000). "Statistics-based summarization-step one: Sentence compression." In: *Proceedings of the 12th Conference on Innovative Applications of Artificial Intelligence*.

Labov, William (1966). *The Social Stratification of New York City*. Washington, D.C.: Center for Applied Linguistics.

— (1997). "Some further steps in narrative analysis." In: *Journal of narrative and life history* 7.

— (2010). "Oral narratives of personal experience." In: *Cambridge Encyclopedia of the Language Sciences*. Ed. by Patrick Hogan. Cambridge, UK: Cambridge University Press.

— (2013). *The Language of Life and Death*. Cambridge, UK: Cambridge University Press.

Labov, William and Joshua Waletzky (1967). "Narrative Analysis: Oral Versions of Personal Experience." In: *Essays on the Verbal and Visual Arts*.

Lan, Wuwei et al. (2017). "A Continuously Growing Dataset of Sentential Paraphrases." In: *Proceedings of The 2017 Conference on Empirical Methods on Natural Language Processing*.

Lapata, Maria (2001). "A corpus-based account of regular polysemy: The case of context-sensitive adjectives." In: *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*.

Legrand, Joël, Michael Auli, and Ronan Collobert (2016). "Neural Network-based Word Alignment through Score Aggregation." In: *Proceedings of the First Conference on Machine Translation*.

Lehnert, Wendy G (1981). "Plot units and narrative summarization." In: *Cognitive Science* 5.

Li, Boyang et al. (2013). "Story generation with crowdsourced plot graphs." In: *Proceedings of the 27th AAAI Conference on Artificial Intelligence*.

Lin, Chin-Yew and Eduard Hovy (2000). "The automated acquisition of topic signatures for text summarization." In: *Proceedings of the 18th conference on Computational Linguistics*.

Linde, Charlotte (1993). *Life STories: The Creation of Coherence*. Oxford, UK: Oxford University Press.

Litvak, Marina, Mark Last, and Menahem Friedman (2010). "A new approach to improving multilingual summarization using a genetic algorithm." In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Liu, Peter J et al. (2018). "Generating wikipedia by summarizing long sequences." In: *Proceedings of the 6th International Conference on Learning Representations*.

Luhn, Hans Peter (1958). "The automatic creation of literature abstracts." In: *IBM Journal of Research and Development* 2.

MacCartney, Bill, Michel Galley, and Christopher D. Manning (2008). "A phrase-based alignment model for natural language inference." In: *Proceedings of the conference on empirical methods in natural language processing*.

Madnani, Nitin and Bonnie J. Dorr (2010). "Generating phrasal and sentential paraphrases: A survey of data-driven methods." In: *Computational Linguistics*.

Maharjan, Nabin et al. (2016). "SemAligner: A Method and Tool for Aligning Chunks with Semantic Relation Types and Semantic Similarity Scores." In: *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

Mallinson, Jonathan, Rico Sennrich, and Mirella Lapata (2017). "Paraphrasing revisited with neural machine translation." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Manshadi, Mehdi, Reid Swanson, and Andrew S Gordon (2008). "Learning a Probabilistic Model of Event Sequences from Internet Weblog Stories." In: *Proceedings of the 21st FLAIRS Conference*.

McCowan, Iain et al. (2005). "The AMI meeting corpus." In: *Proceedings of Measuring Behavior 2005, the 5th International Conference on Methods and Techniques in Behavioral Research*.

McIntyre, Neil and Mirella Lapata (2009). "Learning to Tell Tales: A Data-driven Approach to Story Generation." In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.

— (2010). "Plot induction and evolutionary search for story generation." In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

McKeown, Kathleen et al. (1999). "Towards multidocument summarization by reformulation: Progress and prospects." In: *Proceedings of AAAI*.

McKeown, Kathleen et al. (2010). "Time-Efficient Creation of an Accurate Sentence Fusion Corpus." In: *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mihalcea, Rada (2005). "Language independent extractive summarization." In: *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*.

Mihalcea, Rada and Paul Tarau (2004). "TextRank: Bringing Order into Texts." In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Mikolov, Tomas et al. (2013). "Distributed representations of words and phrases and their compositionality." In: *Advances in Neural Information Processing Systems*.

Miller, George (1995). "WordNet: A Lexical Database for English." In: *Communications of the ACM* 38.

Mitchell, Jeff and Mirella Lapata (2008). "Vector-based Models of Semantic Composition." In: *Proceedings of the 2008 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Montfort, Nick (2011). "Curveship's automatic narrative style." In: *Proceedings of the 6th International Conference on Foundations of Digital Games*.

Murray, Gabriel et al. (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*.

Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou (2017). "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents." In: *31st AAAI Conference on Artificial Intelligence*.

Nallapati, Ramesh et al. (2016). "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond." In: *Proceedings of the 2016 SIGNLL Conference on Computational Natural Language Learning*.

Napoles, Courtney, Matthew Gormley, and Benjamin Van Durme (2012). "Annotated Gigaword." In: *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*.

Nema, Preksha et al. (2017). "Diversity driven attention model for query-based abstractive summarization." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Nenkova, Ani (2005). "Automatic text summarization of newswire: Lessons learned from the Document Understanding Conference." In: *Proceedings of the 20th National Conference on Artificial Intelligence*.

Nenkova, Ani and Kathleen McKeown (2011). "Automatic Summarization." In: *Foundations and Trends in Information Retrieval* 5.

Nenkova, Ani and Rebecca Passonneau (2004). "Evaluating content selection in summarization: The pyramid method." In: *Proceedings of the Human Language Technology Conference of the north American Chapter of the Association for Computational Linguistics*.

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

Nenkova, Ani, Lucy Vanderwende, and Kathleen McKeown (2006). "A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization." In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Och, Franz Josef and Hermann Ney (2003). "A Systematic Comparison of Various Statistical Alignment Models." In: *Computational Linguistics* 29.1.

Ochs, Eleanor and L. Capps (2001). *Living Narrative*. Cambridge, MA: Harvard University Press.

Orăsan, Constantin and Oana Andreea Chiorean (2008). "Evaluation of a Cross-lingual Romanian-English Multi-document Summariser." In: *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Osborne, Miles (2002). "Using maximum entropy for sentence extraction." In: *Proceedings of the ACL-02 Workshop on Automatic Summarization*.

Ouyang, Jessica, Serina Chang, and Kathleen McKeown (2017). "Crowd-sourced iterative annotation for narrative summarization corpora." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Ouyang, Jessica 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*.

— (2019). "Neural network alignment for sentential paraphrases." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Ouyang, Jessica and Kathy McKeown (2014). "Towards Automatic Detection of Narrative Structure." In: *Proceedings of the 9th International Conference on Language Resources and Evaluation*.

Ouyang, Jessica, Boya Song, and Kathleen McKeown (2019). "A robust abstractive system for cross-lingual summarization." In: *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Pang, Bo, Kevin Knight, and Daniel Marcu (2003). "Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences." In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

Paszke, Adam et al. (2017). "Automatic differentiation in PyTorch." In: *NIPS Autodiff Workshop*.

Paulus, Romain, Caiming Xiong, and Richard Socher (2018). "A deep reinforced model for abstractive summarization." In: *Proceedings of the 6th International Conference on Learning Representations*.

Pavlick, Ellie and Ani Nenkova (2015). "Inducing Lexical Style Properties for Paraphrase and Genre Differentiation." In: *Proceedings of NAACL-HLT 2015*.

Pavlick, Ellie et al. (2015). "PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Pew Research Center (2016). *The 2016 Presidential Campaign – a News Event That's Hard to Miss*. Washington, D.C. URL: https://www.journalism.org/2016/02/04/the-2016-presidential-campaign-a-news-event-thats-hard-to-miss/.

Pitler, Emily, Annie Louis, and Ani Nenkova (2009). "Automatic sense prediction for implicit discourse relations in text." In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.

Polanyi, Livia (1981). "What stories can tell us about their teller's world." In: *Poetics Today* 2.

— (1985). *Telling the American Story : A Structural and Cultural Analysis of Conversational Storytelling*. Norwood, NJ: Ablex Publishing.

Prince, Gerald (1973). *A Grammar of Stories: An Introduction*. The Hague: Mouton.

Propp, Vladimir (1928). *Morphology of the Folktale*. Translated by Laurence Scott. Austin, TX: University of Texas Press.

Radev, Dragomir et al. (2002). *Evaluation of text summarization in a cross-lingual information retrieval framework*. Tech. rep. Center for Language and Speech Processing, Johns Hopkins University.

Rahimtoroghi, Elahe et al. (2013). "Evaluation, Orientation, and Action in Interactive StoryTelling." In: *Proceedings of Intelligent Narrative Technologies 6*.

Ricoeur, Paul (1984). *Time and Narrative (Temps et Récit*. Translated by Kathleen McLaughlin and David Pellauer. Chicago: University of Chicago Press.

Riedl, Mark and R. Michael Young (2010). "Narrative planning: Balancing plot and character." In: *Journal of Artificial Intelligence Research* 39.1.

Rishes, Elena et al. (2013). "Generating different story tellings from semantic representations of narrative." In: *International Conference on Interactive Digital Storytelling*.

Rush, Alexander M, Sumit Chopra, and Jason Weston (2015). "A Neural Attention Model for Abstractive Sentence Summarization." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Sagae, Kenji et al. (2013). "A Data-Driven Approach for Classification of Subjectivity in Personal Narratives." In: *Proceedings of the 4th Workshop on Computational Models of Narrative, OASICS OpenAccess Series in Informatics*.

Sandhaus, Evan (2008). "The New York Times Annotated Corpus." In: *Linguistic Data Consortium*.

See, Abigail, Peter J Liu, and Christopher D Manning (2017). "Get to the point: Summarization with pointer-generator networks." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Sharifi, Beaux P, David I Inouye, and Jugal K Kalita (2013). "Summarization of Twitter Microblogs." In: *The Computer Journal* 57.3.

Shriberg, Elizabeth et al. (2004). "The ICSI meeting recorder dialog act (MRDA) corpus." In: *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*.

Smith, Barbara Herrnstein (1980). "Narrative versions, narrative theories." In: *Critical inquiry* 7.

Socher, Richard et al. (2012). "Semantic compositionality through recursive matrix-vector spaces." In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Sultan, Md Arafat, Steven Bethard, and Tamara Sumner (2014a). "Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence." In: *Transactions of the Association for Computational Linguistics*.

— (2014b). "DLS @ CU: Sentence Similarity from Word Alignment." In: *Proceedings of the 8th International Workshop on Semantic Evaluation*.

— (2015). "Feature-rich two-stage logistic regression for monolingual alignment." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Swanson, Reid and Andrew S Gordon (2012). "Say anything: Using textual case-based reasoning to enable open-domain interactive storytelling." In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2.

Swanson, Reid et al. (2014). “Identifying Narrative Clause Types in Personal Stories.” In: *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Tamura, Akihiro, Taro Watanabe, and Eiichiro Sumita (2014). “Recurrent neural networks for word alignment model.” In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Thadani, Kapil, Scott Martin, and Michael White (2012). “A joint phrasal and dependency model for paraphrase alignment.” In: *Proceedings of the 24th International Conference on Computational Linguistics*.

Thadani, Kapil and Kathleen McKeown (2011). “Optimal and syntactically-informed decoding for monolingual phrase-based alignment.” In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Todorov, Tzvetan (1969). “Structural Analysis of Narrative.” In: *NOVEL: A Forum on Fiction*. Translated by Arnold Weinstein. Providence, RI: Brown University.

Ulrich, Jan, Gabriel Murray, and Giuseppe Carenini (2008). “A publicly available annotated corpus for supervised email summarization.” In: *AAAI08 EMAIL Workshop*.

Vanderwende, Lucy et al. (2007). “Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion.” In: *Information Processing & Management* 43.

Vaswani, Ashish et al. (2017). “Attention is all you need.” In: *Advances in Neural Information Processing Systems*.

Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly (2015). “Pointer networks.” In: *Advances in Neural Information Processing Systems*.

Völske, Michael et al. (2017). “TL;DR: Mining Reddit to learn automatic summarization.” In: *Proceedings of the Workshop on New Frontiers in Summarization*.

Wan, Xiaojun (2011). “Using bilingual information for cross-language document summarization.” In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Wan, Xiaojun, Huiying Li, and Jianguo Xiao (2010). “Cross-language document summarization based on machine translation quality prediction.” In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Wan, Xiaojun et al. (2018). "Cross-language document summarization via extraction and ranking of multiple summaries." In: *Knowledge and Information Systems*.

Whissell, Cynthia (1989). "The dictionary of affect in language." In: *Emotion: Theory, research, and experience*.

Wieting, John and Kevin Gimpel (2018). "ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Wieting, John, Jonathan Mallinson, and Kevin Gimpel (2017). "Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Wong, Kam-Fai, Mingli Wu, and Wenjie Li (2008). "Extractive summarization using supervised and semi-supervised learning." In: *Proceedings of the 22nd International Conference on Computational Linguistics*.

Yang, Nan et al. (2013). "Word alignment modeling with context dependent deep neural network." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Yao, Jin-ge, Xiaojun Wan, and Jianguo Xiao (2015). "Phrase-based compressive cross-language summarization." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Yao, Xuchen (2014). "Feature-driven Question Answering with Natural Language Alignment." PhD thesis.

Yao, Xuchen et al. (2013a). "A Lightweight and High Performance Monolingual Word Aligner." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

— (2013b). "Semi-Markov Phrase-based Monolingual Alignment." In: *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing*.

Yessenalina, Ainur and Claire Cardie (2011). "Compositional matrix-space models for sentiment analysis." In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Yu, Mo and Mark Dredze (2015). "Learning composition models for phrase embeddings." In: *Transactions of the Association for Computational Linguistics*.

Zanzotto, Fabio Massimo et al. (2010). "Estimating linear models for compositional distributional semantics." In: *Proceedings of the 23rd International Conference on Computational Linguistics*.

Zhang, Jiajun, Yu Zhou, and Chengqing Zong (2016). "Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.10.

Zhao, Shiqi et al. (2008). "Combining multiple resources to improve SMT-based paraphrasing model." In: *Proceedings of ACL-08: HLT*.

Zhou, Liang et al. (2006). "Paraeval: Using paraphrases to evaluate summaries automatically." In: *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*.