Decontextualization: Making Sentences Stand-Alone

Eunsol Choi[2*], **Jennimaria Palomaki**[1], **Matthew Lamm**[1],
Tom Kwiatkowski[1], **Dipanjan Das**[1], **Michael Collins**[1]
[1]Google Research
[2]Department of Computer Science, The University of Texas at Austin
eunsol@cs.utexas.edu,
{jpalomaki,mrlamm,tomkwiat,dipanjand,mjcollins}@google.com

Abstract

Models for question answering, dialogue agents, and summarization often interpret the meaning of a sentence in a rich context and use that meaning in a new context. Taking excerpts of text can be problematic, as key pieces may not be explicit in a local window. We isolate and define the problem of sentence decontextualization: taking a sentence together with its context and rewriting it to be interpretable out of context, while preserving its meaning. We describe an annotation procedure, collect data on the Wikipedia corpus, and use the data to train models to automatically decontextualize sentences. We present preliminary studies that show the value of sentence decontextualization in a user-facing task, and as preprocessing for systems that perform document understanding. We argue that decontextualization is an important subtask in many downstream applications, and that the definitions and resources provided can benefit tasks that operate on sentences that occur in a richer context.

1 Introduction

Many applications of natural language processing need to be able to interpret, or present, text independently from the rich context in which it occurs. For example, summarization systems extract salient information from documents and present it in a reduced context. Many systems also segment documents prior to interpretation of retrieval for computational efficiency. In all of these cases, we would like the context-reduction step to be *meaning preserving* but, to date, there has been no independent method of ensuring this.

In this paper we isolate and define the problem of *sentence decontextualization*: taking a sentence together with its context and rewriting it to be interpretable out of context if feasible, while preserving its meaning.[1] Having defined the problem, we operationalize this definition into a high quality annotation procedure; use the resulting data to train models to automatically decontextualize sentences; and present preliminary results that show the value of automatic decontextualization in a user-facing task, and as preprocessing for systems that perform document understanding. We argue that decontextualization is an important sub-task in many downstream applications, and we believe this work can benefit tasks that operate on sentences that occur in a wider context.

One contribution of this work is to release a dataset of decontextualized sentences that can be used as training and evaluation data, together with the evaluation script: On publication of this paper the data will be available at `https://github.com/google-research/language/tree/master/language/decontext`.

Figure 1 shows an example decontextualization. In this example we have a coreference resolution step (their → The Croatia national football team's) and a bridging step (insertion of the prepositional phrase ''in the FIFA World Cup'' to modify ''Croatia's best result thus far''). Decontextualization involves various linguistic phenomena, including coreference resolution, global scoping, and bridging anaphora (Clark, 1975). We present a linguistically motivated definition of decontextualiation in Section 2 and show that this definition can be reliably applied by crowdworkers in Section 3.

[*]Work done at Google.

[1] More precisely the truth-conditional meaning or *explicature* (Sperber and Wilson, 1986); see section 2 for discussion.

Transactions of the Association for Computational Linguistics, vol. 9, pp. 447–461, 2021. https://doi.org/10.1162/tacl_a_00377
Action Editor: Chris Brew. Submission batch: 10/2020; Revision batch: 12/2020; Published 4/2021.

> **Page Title:** Croatia at the FIFA World Cup
> **Paragraph:** Croatia national football team have appeared in the FIFA World Cup on five occasions (in 1998, 2002, 2006, 2014 and 2018) since gaining independence in 1991. Before that, from 1930 to 1990 Croatia was part of Yugoslavia. Their best result thus far was reaching the 2018 final, where they lost 4-2 to France.
> **Decontextualized Sentence:** The Croatia national football team's best result thus far in the FIFA World Cup was reaching the 2018 final, where they lost 4-2 to France.

Figure 1: An example decontextualization. The sentence to decontextualize is highlighted in gray.

We generate a corpus of decontextualized sentences corresponding to original sentences drawn from the English Wikipedia. We show that a high proportion of these original sentences can be decontextualized using a relatively simple set of re-write operations, and we use the data to define a new *automatic decontextualization* task in which a computer system needs to create a decontextualized sentence from an original sentence presented in paragraph context. We discuss the implications of choosing Wikipedia as a domain in Section 3.4.

We present two methods for automatic decontextualization based on state-of-the-art coreference (Joshi et al., 2020) and generation (Raffel et al., 2019a) models. We evaluate the output of these models with automatic measures (derived from Xu et al. [2016]), as well as through human evaluation. Both automatic and human evaluations show that the largest sequence-to-sequence model produces high quality decontextualizations in the majority of cases, although it still lags human performance in the thoroughness and accuracy of these decontextualization edits.

Finally, we present two demonstrations of the utility of decontextualization. The first is a user study giving evidence that decontextualized sentences can be valuable when presented to users as answers in a question-answering task—raters judge that they balance conciseness with informativeness. In the second one, we use decontextualization as a preprocessing component for generating a retrieval corpus for open domain question answering. Decontextualizing the sentences to be indexed by retrieval system enables more efficient answer string retrieval for information seeking queries. These demonstrations are presented as preliminary results, and we argue that decontextualization is an important sub-task for a wide range of NLP applications.

2 Linguistic Background

We start with the following definition:

Definition 1 (Decontextualization) *Given a sentence-context pair (s, c), a sentence s' is a valid* decontextualization *of s if: (1) the sentence s' is interpretable in the empty context; and (2) the truth-conditional meaning of s' in the empty context is the same as the truth-conditional meaning of s in context c.*

A context c is a sequence of sentences preceding s, and the empty context is the empty sequence.

We have been careful here to use the more specific term ''truth conditional meaning'' rather than ''meaning''. Here we follow the distinction in semantics/pragmatics between truth conditional meaning and implicature, and deliberately exclude implicatures (which can also be considered part of the meaning of an utterance) from our definition. There is a rich history of work in semantics and pragmatics on truth-conditional meaning and implicatures, going back to Grice (1975). Our concept of ''truth conditional meaning'' is very close to ''explicature'' as used in Relevance Theory (Sperber and Wilson, 1986). Consider this description of explicature from Birner (2012) (pages 96–97, our own emphasis added):

> The explicature in an utterance is the result of enriching the semantic content with the sorts of pragmatic information necessary to provide us with a truth-evaluable proposition. This includes calculating the referents for pronouns, working out the intended interpretation for deictic phrases like here and later ..., *disambiguating lexically and structurally ambiguous words and phrases, making any ''bridging'' inferences necessary for reference resolution* ... and so on.

We will see in the next section that our annotation task follows this definition quite closely.

As an example consider the following exchange:

> Susan: Has the Croatia national football team ever won the FIFA World Cup?
> Jon: Their best result thus far was reaching the 2018 final, where they lost 4-2 to France.

Here the truth conditional meaning of Jon’s reply is equivalent to ‘‘Croatia’s best result thus far in the FIFA World Cup was reaching the 2018 final, where they lost 4-2 to France’’, whereas the implicature would be ‘‘the Croatia national football team has never won the FIFA World Cup’’ (which answers Susan’s question). In our definition the decontextualized sentence s' should preserve the truth-conditional meaning, but is not required to preserve the implicature(s) of the sentence.[2]

Remark (extra-linguistic context): In addition to its document context, a given sentence s and its counterpart s' also come with a temporal, cultural, and geographic context—that is, where and when they are being written or read and by whom.[3] We assume that these aspects of context are preserved during decontextualization. The effect of this is that elements of s that derive their meaning from outside of the document context will receive equivalent interpretation in s', and hence do not require decontextualization. For example, the expression ‘‘thus far’’ in Figure 1 is interpreted relative to the time of utterance, not relative to what has been previously said in the Wikipedia article, and hence it appears in both the original and decontextualized sentences.

3 Task Definition

An annotator is provided with an entire document d with a target sentence within the document, represented as a start and end index s_{st}, s_{end}. First, the annotator decides whether the target sentence can be decontextualized or not, labeling it as FEASIBLE or INFEASIBLE. If the example is marked as FEASIBLE, the annotator decontextualizes the sentence, producing y, a new sentence that satisfies the conditions in Definition 1.

3.1 Feasibility

Sentences in FEASIBLE include sentences that do not require any modification to be decontextualized (e.g., *‘‘Émilie du Châtelet proposed the hypothesis of the conservation of total energy, as distinct from momentum’’*), and sentences that require edits to stand alone.

Page Title: Postage stamps and postal history of India
Paragraph: ... In the opinion of Geoffrey Clarke , the reformed system was to be maintained “ for the benefit of the people of India and not for the purpose of swelling the revenue.” The Commissioners voted to abolish the earlier practice of conveying official letters free of postage (“franking”). The new system was recommended by the Governor - General , Lord Dalhousie , and adopted by the East India Company ’s Court of Directors.

Page Title: Thermodynamic temperature

Paragraph: ... To completely melt ice at 0 C into water at 0 C, one must add roughly 80 times the thermal energy as is required to increase the temperature of the same mass of liquid water by one degree Celsius. The metals’ ratios are even greater, typically in the range of 400 to 1200 times.

Figure 2: Decontextualization examples falling into the INFEASIBLE category. The sentence to be decontextualized is highlighted in gray.

In the decontextualization step, we instructed annotators to make only minor modifications, which includes copying and pasting a few phrases from the document to the target sentence and deleting phrases from the target sentence. When it is too challenging to decontextualize, it is classified into the INFEASIBLE category. Often, sentences in this category are a part of a narrative story, or rely heavily on the preceding few sentences. See Figure 2 for examples.

3.2 Edit Types and Linguistic Phenomena

When an example is classified as FEASIBLE, the annotator makes edits to decontextualize the sentence. Table 1 shows the different edit types. They fall into four broad categories:

NAME COMPLETION, PRONOUN / NP SWAP correspond to replacement of a referring expression that is unclear out of context with a referring expression that is unambiguous out of context. For example, replacing the pronoun ‘‘She’’ with ‘‘Cynthia Nixon’’, the definite NP ‘‘the copper statue’’ with ‘‘The Statue of Liberty’’, or the abbreviated name ‘‘Meg’’ with ‘‘Megan ‘‘Meg’’ Griffin’’.

DM REMOVAL involves removal of discourse markers (DMs) such as ‘‘therefore’’.

BRIDGING, GLOBAL SCOPING involve addition of a phrase (typically a prepositional phrase) that

[2] We have not necessarily given up on recovering implicatures: The decontextualized sentence will likely be a valuable intermediate step in deriving the implicatures of an utterance.

[3] Research on text simplification (Xu et al., 2015; Bingel et al., 2018) also shows how target output depends on the expected audience.

Edit Type	Description	Example	%
PRONOUN/NP SWAP	Replacement of a definite pronoun / noun phrase with another referring expression	⟅ -The copper statue , +The Statue of Liberty ⟆, a gift from the people of France to the people of the United States, was designed by French sculptor Frédéric Auguste Bartholdi and built by Gustave Eiffel.	40.5
NAME COMPLETION	Expansion of acronyms or partial names	⟅ -Meg , +Megan ‘‘Meg’’ Griffin ⟆ made her first appearance on television when Family Guy debuted on Fox on January 31, 1999, with the episode ‘‘Death Has a Shadow’’.	11.5
DM REMOVAL	Removal of discourse markers that can be only understood in context	⟅ - For instance, ⟆ Alaska could be regarded as the highest state because Denali, at 20,310 feet, is the highest point in the US.	3.5
BRIDGING	Addition of a modifier (typically a PP) to a noun phrase	In all fights ⟅ +in the Ultimate Fighting Championship ⟆, each round can be no longer than five minutes.	13
GLOBAL SCOPING	Addition of a phrase (typically a PP) that modifies the entire sentence	The Japanese film Shoplifters, directed by Hirokazu Kore-eda, won the Palme d’Or ⟅ +at the 2018 Cannes Film Festival. ⟆	7
ADDITION	Addition of background information that is not necessary but helps readability significantly	Charles Darwin⟅ +, an English naturalist and biologist, ⟆ was among the first to suggest that physiological changes caused by an emotion had a direct impact on , rather than being just the consequence of that emotion.	10

Table 1: The list of possible edits in decontextualization. The last column represents how frequently the phenomena occurs in the data, from manual analysis on 200 examples, including examples that belongs to INFEASIBLE categories and examples that does not require any edits. The bag notation removes x and adds y (⟅ -x , +y ⟆) at its position.

modifies either a particular noun phrase (‘‘bridging’’) or the entire sentence (‘‘global scoping’’). For example, adding ‘‘in the Ultimate Fighting Championship’’ as a modifier to ‘‘all fights’’, or adding ‘‘at the 2018 Cannes Film Festival’’ at the end of the sentence. The additional phrase essentially spells out a modifier that is implied by the context.

ADDITION inserts background information that significantly improves readability: In many cases, this involves adding an appositive or premodifier to a named entity to add useful background information about that entity. Unlike other edits described above, edits in this category are optional. For example, replacing ‘‘The Eagles’’ with ‘‘The American rock band The Eagles.’’

3.3 Variability

We note that for a given sentence frequently there will be more than one possible decontextualization. While this inherent subjectivity makes the task challenging to crowdsource and evaluate, we argue this is important feature, as shown in recent literature (Aroyo and Welty, 2015; Pavlick and Kwiatkowski 2019; Kwiatkowski et al., 2019), and propose to collect multiple references per example. Table 2 shows examples where there can be multiple different correct decontextualizations. In the first example, while the semantics of the edits are roughly equivalent (i.e., the annotators agreed on what noun phrases have to be disambiguated and information has to be added), they differ in *how* to rewrite the sentence. In the second example, we see disagreement on *what* information should be added to the sentence. We do not make any explicit assumptions about what is known and salient to the reader, and instructed annotators to use their best judgment to rewrite such that the new sentence is fluent, unambiguous and clear when posed alone. In the last example, annotators disagree on the feasibility. While the sentence is a part of a bigger narrative, two annotators judged it could be edited to alone, by adding a global scoping modifier, ‘‘In Greek mythology.’’

3.4 Scope of Current Task Formulation

Our data comes from the English portion of the Wikipedia corpus. We sampled sentences as follows. We first pick a (question, Wikipedia, short answer) triple from the Natural Questions

Page title / Section title: We Don't Talk Anymore (Charlie Puth song) / Music video
Paragraph: The music video premiered on August 2 , 2016 , on BuzzFeed and was directed by Phil Pinto . It shows Puth and Mirella Cardoso as his love interest . . . .
Decontextualization 1: ⟅ -It , +We Don't Talk Anymore music video ⟆ shows ⟅ -Puth , +Charlie Puth ⟆ and Mirella Cardoso as his love interest.
Decontextualization 2: ⟅ -It , +The ''We Don't Talk Anymore''(Charlie Puth song) music video ⟆ shows Puth and Mirella Cardoso as his love interest.

Page title: The American Baking Competition
Paragraph: CBS placed casting calls for participants on November 14, 2012 . Auditions were held between December 1 and December 15, 2012. The competition took place at the Gibbs Gardens in Ball Ground , Georgia in March 2013.
Decontextualization 1: The ⟅ -competition , +American Baking Competition ⟆ took place at the Gibbs Gardens in Ball Ground , Georgia in March 2013.
Decontextualization 2: The ⟅ -competition , +American Baking Competition, a reality competition television series, ⟆ took place at the Gibbs Gardens in Ball Ground , Georgia in March 2013 .

Page title: Gemini (Constellation)
Paragraph: In Greek mythology, Gemini was associated with the myth of Castor and Pollux, the children of Leda and Argonauts both. Pollux was the son of Zeus, who seduced Leda, while Castor was the son of Tyndareus, king of Sparta and Leda's husband. Castor and Pollux were also mythologically associated with St. Elmo's fire in their role as the protectors of sailors. When Castor died, because he was mortal, Pollux begged his father Zeus to give Castor immortality, and he did, by uniting them together in the heavens.
Decontextualization 1: Infeasible
Decontextualization 2: ⟅ +In Greek mythology, ⟆ when Castor died, because he was mortal, Pollux begged his father Zeus to give Castor immortality, and he did, by uniting them together in the heavens.

Table 2: Examples showing the diversity of valid decontextualization edits.

(Kwiatkowski et al., 2019) uniformly at random from the questions that have a short answer. We include the sentence containing the short answer as one example; as a second example we choose a sentence at random from the Wikipedia page. After sampling we exclude (1) sentences under a ''Plot'' category as they are often infeasible to decontextualize; (2) any sentence that is the first sentence of the page; and (3) any sentence from a paragraph containing only a single sentence.

We designed this data selection process to ensure that a large proportion of examples (90%) could be decontextualized using simple edits described in Section 3.2.

Before settling on Wikipedia, we conducted an initial pilot study which revealed that encyclopedic text is substantially easier to decontextualize compared to newswire or literary text. In the latter genres, the context required for the comprehension of any given sentence appears to be much more complex in structure. Similarly, it is difficult to posit decontextualization for sentences that appear on social media platforms, because they are situated within complex and highly specific social contexts. In contrast, being written for a general audience, Wikipedia makes limited assumptions about its reader.

Within Wikipedia, we similarly found that articles on popular historical or cultural entities and events were easier to decontextualize by crowdworkers compared to articles from technical domains, such as ones on medical or mathematical concepts. Comprehension of such articles requires a considerable body of background knowledge or information from preceding paragraphs. Articles in our dataset cover topics that require little background knowledge to comprehend.

We focus on decontextualization of *sentences*, where the space of edits is restricted, to make the task easier to quality control and annotate. However alternate formulations, such as decontextualization on *paragraphs* could also be studied. One could even also consider allowing wider range of edits, such as multi-sentence outputs and edits beyond copy-and-pasting, such as paraphrasing and re-ordering. We anticipate exploring such alternative formulations would help to extend the scope of decontextualization to the more challenging domains previously mentioned.

We stress however that in spite of our restriction to single sentences in Wikipedia, the decontextualization task is nevertheless valuable: Wikipedia (and other encyclopedic sources) contain a wealth of factual information, and a high proportion (over

	#	par. len	sent. len	FEASIBLE (%)		INFEASIBLE (%)
				w/ edit	as is	
Train	11290	695	156	60	31	9
Dev	1945	695	162	67	21	12
Test	1945	711	160	68	20	12
Expert	100	658	163	63	26	12

Table 3: Data statistics. par. len refers to paragraph length in bytes, and sent. len refers to sentence length in bytes. All lengths are in bytes. The development and test set is five-way annotated, and the expert data is four-way annotated.

60%; see Table 3) of sentences both require decontextualization and can be decontextualized under our definitions (only 30% of sentences are interpretable out of context without any edits).

4 Data Collection

Annotation Interface The annotator is presented a sentence in the context of an entire Wikipedia page. In the first step the annotator judges whether the example is FEASIBLE or INFEASIBLE. If the example is marked as FEASIBLE, the annotator can use delete, add, or swap operations within a user interface to produce a decontextualized string.

Data Statistics We collected one reference for each example in the training data, and five references for each example in the evaluation data. Annotators are native speakers of English located in the United States, and on average, they took 4 minutes to annotate a single example.

In total, 28 annotators annotated the examples, with 11 annotators annotating more than 1K examples each.

Table 3 represents some overall data statistics. Decontextualization is possible for the majority of examples, with the INFEASIBLE category covering roughly 10% of the data. We note a slight discrepancy between train and evaluation dataset distribution, potentially due to a change in the annotation interface. A small subset of data is annotated by the authors to be compared with the crowd-sourced data (last row in the table).

Annotation Quality We quantify the annotation agreement on the category classification. The Fleiss' kappa on category classification is 0.51 among expert annotators, and is 0.30 among the crowd annotators (binary agreement is at 85%). We observed more variability in crowdworkers as annotators' background is more diverse, and some annotators have a loose concept of ''stand alone'' and consistently attempted decontextualization.

We also measured agreement among the individual edits. For each of the edit operations (as defined in Section 3.2), we compare the output sentence after the single edit and to a set of output sentences, each after a single edit by other annotators. About 32.5% of edits were covered.

Because of the inherent annotation variability, four of the authors manually evaluated 100 crowd-sourced annotations from the evaluation data based on two measures: (1) whether the sentence is sufficiently and correctly decontextualized, and (2) whether the sentence is grammatically correct and fluent. Overall, 88% of annotations were valid in both, 89% on the content and 88% on form.

5 Automatic Decontextualization

5.1 Models

We present two models for decontextualization: a coreference resolution model and a sequence-to-sequence generation model. For both models, the input is a concatenation of the title of the Wikipedia document, the section titles, and the paragraph containing the target sentence. During the annotation pilots, we found that the document title is crucial for decontextualization, while section headers were frequently necessary or missing. We denote the title of the Wikipedia page as the sequence of tokens t, section titles of the paragraph as the sequence of tokens t_s and the n sentences in the paragraph where the target sentence is coming from as $x_1 \ldots x_n$, where each x_i is a sequence of tokens, and x_t is the target sentence ($1 \leq t \leq n$). The model considers the concatenation of a subset of the document,

$$\texttt{[CLS]}\, t\, \texttt{[S]}\, t_s\, \texttt{[S]}\, x_1 \cdots x_{t-1}\, \texttt{[S]}\, x_t\, \texttt{[S]}\, x_{t+1} \cdots x_n\, \texttt{[S]}$$

where [S] is a separator token. This representation differs from the setting of annotators, where they were given the full document context. As an approximation, we include article and section titles in the inputs, as these often contain salient contextual elements. We did experiment with giving more context, namely, adding the first paragraph of the article as an additional input, but did not observe a performance improvement. On the initial pilot, annotators marked that 10–20% of examples required access to the full document.

The Coreference Model As many decontextualization edits can be recovered by a coreference resolution module, we adapt the output from the state-of-the-art coreference resolution system of Joshi et al. (2020), trained on the CoNLL dataset (Pradhan et al., 2012), as a decontextualization system. We used the publicly available pre-trained checkpoint of SpanBERT-Large with the original hyper parameters.[4]

We run this model on the input sequence, and map the coreference cluster predictions to modify the sentence as follows. We only consider clusters with a mention in the target sentence. For each such cluster, we find its first mention inside the target sentence, and find another mention in the same cluster that was presented earlier in the input and is longer than the current mention. If such a mention is found, we replace the current entity mention string with the earliest such mention string (e.g., ''She'' is replaced with ''Taylor Swift''). On average, 36.5% of examples were modified through this process.

The Seq2Seq Generation Model is based on the recent T5 model (Raffel et al., 2019b). We show two variations of the model, BASE and 11B, which mainly differ in the model capacity. We fine-tune the model on our crowdsourced training set, by setting the target sequence to be `[CAT] [SEP]` y, where `[CAT]` $\in \{\text{UNNECESSARY}, \text{FEASIBLE}, \text{INFEASIBLE}\}$ and y is a decontextualized sentence when `[CAT]` $=$ FEASIBLE and the original sentence when `[CAT]` $\in \{\text{UNNECESSARY}, \text{INFEASIBLE}\}$. UNNECESSARY are examples where the original sentence without any edit can stand alone.

We limit the input/output to 512/128 tokens for both variants, and fine-tuned from pre-trained checkpoints[5] with a batch size of 100 examples until the validation loss stopped decreasing, after about 32K for the larger and 500K steps for the smaller model.

5.2 Evaluation

5.2.1 Feasibility Detection

We first evaluate the accuracy of models in making the feasible vs. infeasible decision. To do this we compute the binary agreements with all human references and average them to get an accuracy.

Results For the feasible vs. infeasible classification task, a baseline that always predicts FEASIBLE will have 88% accuracy. The larger variant of T5, T5-11B, achieves 89% accuracy, outperforming human agreement (85% accuracy), affirming the strong performance of pre-trained language models on classification tasks (Devlin et al., 2018). This model predicts the INFEASIBLE category infrequently for the larger variant (5% of examples), while humans classify an example as INFEASIBLE for 12% of examples. We observe the smaller variant, T5-Base, is less accurate, over-predicting the INFEASIBLE category (for 20% of examples), getting 77% accuracy. The coreference model cannot decide the decontextualization feasibility, as an untrained baseline.

5.2.2 Decontextualized Sentence Generation

Setup For development / test examples, we have five human annotations per example. We only consider examples marked by three or more annotators (out of five) as FEASIBLE for decontextualized sentence generation. For each of these examples, we discard annotations which mark the example as INFEASIBLE. For automatic evaluation and comparison, we need a human output, which will be compared to model outputs, and a set of reference annotations that will be considered as correct, gold annotations. The single human output provides a reference point for evaluation measures to which the automatic output can be compared.

We observed comparing a longer decontextualized sentence to shorter decontextualized sentences often erroneously results in low scores automatic metrics (e.g., in the last example of Table 2, adding extra information will be erroneously punished). Thus, instead of randomly selecting one annotation to be used as the representative human output, we sort the annotations by the length of the output sentence (raw bytes), and take the annotation with median length[6] as a human output and take the remaining annotations as a set of reference annotations. From manual inspection of the data the median-length output appeared often to be optimal in terms of balancing length versus accuracy of the decontextualization.

Metric For each model prediction and human output, we report:

- Length increase, the average value of (len(decontext)-len(original)) / len(original).

[4] `https://github.com/facebookresearch/SpanBERT/`.

[5] `https://github.com/google-research/text-to-text-transfer-transformer`.

[6] When there are four references, we take the second shortest sentence.

- % edited, the proportion of examples that were modified for decontextualization (as opposed to being left unchanged).
- Sentence match, a binary score computed between the output and a set of references, indicating whether the output matches any of the references after normalization (stripping away articles and punctuation and lowercasing). We report two numbers, a score on all examples, and a score on examples where all references edited the sentence.
- SARI (**s**ystem output **a**gainst **r**eferences and against the **i**nput sentence) metric (Xu et al., 2016). To compute this, for each reference, we calculate a set of **add** edits, corresponding to which unigrams are seen in the reference but not in the original sentence. Conversely, we can calculate the set of **delete** edits, corresponding to unigrams that are in the original sentence but not in the reference. We calculate precision/recall/F1-measure on add and delete edits. We look at unigrams only, and use fractional counts for the words in the references (i.e., a word appearing in one of r references will be counted as $1/r$). We compute micro average across examples, that is, globally by counting the total true positives, false negatives, and false positives, as many examples do not require any edits.[7]

While the sentence match score is the easiest to interpret, it punishes longer outputs, making comparisons across systems producing outputs of different lengths challenging, and it overly rewards conservative strategies that simply copy across the original sentence. Thus, we use the SARI metric as our main evaluation metric. SARI can be thought of as a precision/recall measure on topics (unigrams) that should be added or deleted.

	len inc.	% edited	match all / edited	SARI add F1 (P/R)	SARI del F1 (P/R)
Repeat	0	0	38 / 0	0 (0/0)	0 (0/0)
Coref	7	42	39 / 13	22 (51/14)	31 (34/28)
T5-Base	8	40	48 / 21	29 (67/19)	40 (54/32)
T5-11B	12	59	53 / 32	42 (72/30)	46 (49/43)
Human	24	76	45 / 29	56 (64/49)	58 (61/55)

Table 4: Development set performance. Len inc. is the average percentage increase in length from decontextualization. % edited is the proportion of examples that have at least one edit. match-all shows percentage of outputs that have at least one match in the human references; match-edited shows the match value calculated on cases where all references include at least one edit.

	len inc.	% edited	match all / edited	SARI add F1 (P/R)	SARI del F1 (P/R)
Repeat	0	0	36 / 0	0 (0/0)	0 (0/0)
Coref	8	42	38 / 13	23 (50/15)	36 (40/32)
T5-11B	13	61	52 / 32	43 (69/31)	47 (49/46)
Human	23	77	44 / 28	56 (64/49)	58 (61/56)

Table 5: Test set results. See Table 4 caption for a key.

Automatic Evaluation Tables 4 and 5 show development and test performance. A successful decontextualization system would result in high sentence match, adequate changed ratio (experts edited about 79% of examples), and length change ratio (the experts' ratio is 1.19), as well as high SARI addition and deletion scores. As a sanity check, we report Repeat, which outputs the original sentence. This alone results in high sentence match score, around 40%, meaning that on this number of examples, at least one of the annotators deemed the sentence can stand alone without any edits.

The coreference system has an exact match of about 13% of examples that require edits, without any task-specific fine-tuning. Its SARI add scores shows high precision and low recall, and its deletion scores are low as it cannot delete discourse markers. The Seq2seq generation model achieves high scores across all measures. The bigger variant is substantially better, editing more than its smaller variant without losing precision. We observe the larger variants outperform the average human on sentence match measure, but not in SARI measures. The T5 model modifies fewer examples than the annotator, and edits involve fewer tokens, benefiting it on the sentence match measure. However, the model is more likely to miss required edits, as shown in low recall for the SARI add and deletion measures. We discuss this further in the following human evaluation section.

Human Evaluation We sampled 100 examples in the evaluation set, where at least two annotators

[7] Similar to BLEU in machine translation, SARI is a useful measure for comparing different systems; however, due to the relatively large space of possible decontextualizations it will not be possible to achieve anything close to 100% F1 on SARI measures, and thus the absolute score is harder to interpret. A SARI score of for example 50% should *not* be interpreted as indicating a system with 50% accuracy.

	T5	either	Annotator	Sum
T5	13	12	2	27
either	7	22	4	33
Annotator	1	15	24	40
Sum	21	49	30	100

Table 6: Preference between T5 output and human annotation. Columns represents the judgement of the expert A, rows that of the expert B. We see high agreement between two expert annotators, despite one expert annotator (column annotator) is ambivalent more frequently.

and our best model made decontextualization edits. We randomized the order or presentation of the T5 and human outputs so as to not bias the annotation. On this set, we (two of the authors) conducted a manual evaluation. Given two decontextualized sentences, one from the best model and another randomly selected from a set of annotations with decontextualization edits, we evaluated each on two dimensions: (a) is it fluent and grammatically correct? (b) is it sufficiently and correctly decontextualized? Lastly, we chose the preference between two outputs (A, B, or either).

Expert annotators marked as ''sufficient'' those items for which all possible referential ambiguities had been resolved. Given the subjective nature of the task, some ''insufficient'' decontextualizations by the expert annotator could be valid for the another annotator with a different world knowledge. We report averaged binary scores from two experts. The model output scored 88.0% on fluency, and 67.5% on correct decontextualization, while the human reference output scored 84.5% on fluency and 78.5% on correct decontextualization. Both annotators found T5 to be slightly more fluent, while humans are more thorough and accurate in decontextualizating. Table 6 shows the preferences of two annotators. Both preferred human output, and their preferences exhibit high agreement (matching on 37 out of 40 examples when both had preferences).

We briefly characterize common error patterns for annotators and the T5 model. Similar error patterns emerge between the annotations and model outputs. Both occasionally fail to identify generics that need to be replaced with referring NPs, phrases that require bridging, and temporal contexts that should be provided. Additionally, we noticed that the T5 model heavily relies on the title cues, and sometimes fail to clarify ambiguous entities that are not the main entity of the page. We noticed very few examples where T5 hallucinates factually incorrect contents.

	Prefer			log odds
Opt.A vs. Opt.B	A	B	either	intercept [CI]
Dec. vs. Ori.	730	426	364	0.85 [0.4,1.3]
Dec. vs. Par.	850	456	234	0.55 [0.1,1.0]
Ori. vs. Par.	741	505	274	0.31 [−0.2,0.8]

Table 7: User study results. Dec. refers to decontextualized sentence answer, Ori. means original sentence answer, Par. means paragraph answer. We present raw counts of preferences and the log odds of preferring option A and its 95% confidence interval.

6 Two Applications

We present two demonstrations of the utility of decontextualization. First, we argue that the decontextualized sentences can be valuable in themselves in question answering, and show that they can be useful as a preprocessing step.

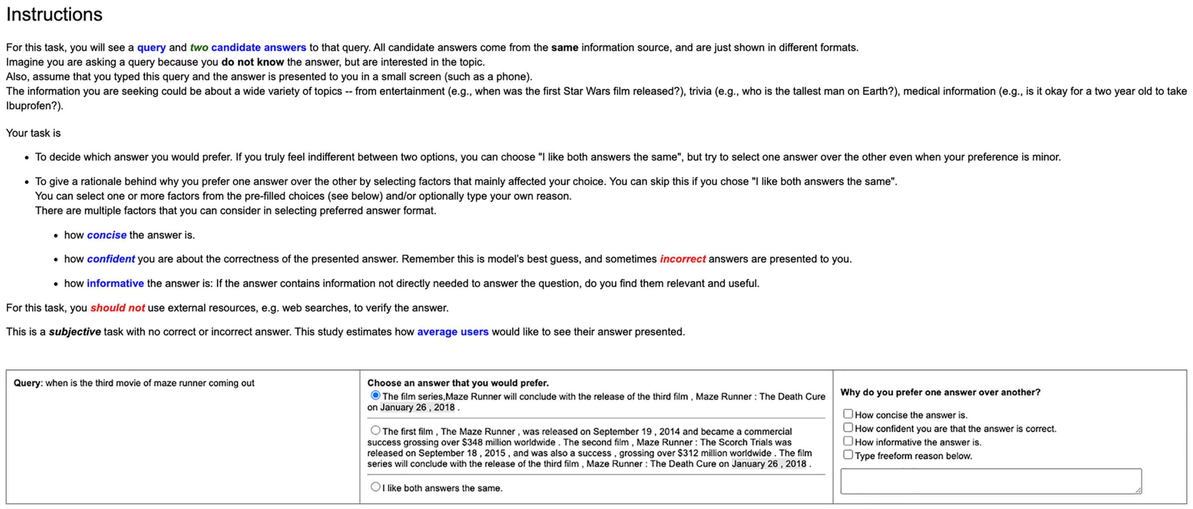
6.1 Decontextualized Answer As Is

We showcase a use case of decontextualized sentences as providing a succinct yet informative answer to open domain factoid questions (Kwiatkowski et al., 2019). We design a user study where people compare a decontextualized-sentence answer with an original-sentence answer and a paragraph answer to the same query.[8]

Setup Given a question and two presentations of the same answer, raters were tasked with marking their preference between the two answer presentations (option A, option B, or either). The actual short span answer in the sentence is always highlighted (similar to seen in Table 8) (See Figure 3 for a screenshot).

We conduct three comparison studies on the same set of 150 questions: (a) decontextualized sentence vs. original sentence, (b) original sentence vs. original paragraph, (c) decontextualized sentence vs. original paragraph. For each example in each study, we collected 10 user ratings. The questions are randomly chosen from a set of questions that have a short answer, and such that the sentence containing the short answer is categorized as FEASIBLE by the annotators and

[8] Understanding how to present answers to users is a complex problem with many desiderata, e.g., preserving the original content, crediting the source, interaction with the user interface, which we are not covering comprehensively.



Instructions

For this task, you will see a **query** and ***two* candidate answers** to that query. All candidate answers come from the **same** information source, and are just shown in different formats.
Imagine you are asking a query because you **do not know** the answer, but are interested in the topic.
Also, assume that you typed this query and the answer is presented to you in a small screen (such as a phone).
The information you are seeking could be about a wide variety of topics -- from entertainment (e.g., when was the first Star Wars film released?), trivia (e.g., who is the tallest man on Earth?), medical information (e.g., is it okay for a two year old to take Ibuprofen?).

Your task is

- To decide which answer you would prefer. If you truly feel indifferent between two options, you can choose "I like both answers the same", but try to select one answer over the other even when your preference is minor.
- To give a rationale behind why you prefer one answer over the other by selecting factors that mainly affected your choice. You can skip this if you chose "I like both answers the same".
  You can select one or more factors from the pre-filled choices (see below) and/or optionally type your own reason.
  There are multiple factors that you can consider in selecting preferred answer format.
  - how ***concise*** the answer is.
  - how ***confident*** you are about the correctness of the presented answer. Remember this is model's best guess, and sometimes ***incorrect*** answers are presented to you.
  - how **informative** the answer is: If the answer contains information not directly needed to answer the question, do you find them relevant and useful.

For this task, you ***should not*** use external resources, e.g. web searches, to verify the answer.

This is a ***subjective*** task with no correct or incorrect answer. This study estimates how **average users** would like to see their answer presented.

Query: when is the third movie of maze runner coming out

Choose an answer that you would prefer.

● The film series,Maze Runner will conclude with the release of the third film , Maze Runner : The Death Cure on January 26 , 2018 .

○ The first film , The Maze Runner , was released on September 19 , 2014 and became a commercial success grossing over $348 million worldwide . The second film , Maze Runner : The Scorch Trials was released on September 18 , 2015 , and was also a success , grossing over $312 million worldwide . The film series will conclude with the release of the third film , Maze Runner : The Death Cure on January 26 , 2018 .

○ I like both answers the same.

Why do you prefer one answer over another?

□ How concise the answer is.
□ How confident you are that the answer is correct.
□ How informative the answer is.
□ Type freeform reason below.

Figure 3: A screenshot of the instruction and an example instance comparing the original sentence answer and the paragraph answer shown to annotators for the user study.

Query	Decontextualized answer	Paragraph answer (sentence answer highlighted)	Decont.	Ori.
when was the rising of the moon written	The Rising of the Moon, Irish ballad recounting a battle between the United Irishmen and the British Army has been in circulation since circa 1865 .	The ballad has been in circulation since circa 1865 . The earliest verifiable date found in publication is 1867 .	−2.09	−1.53
what is the most viewed video on youtube in 24 hours	The most viewed music video within 24 hours of its release is Taylor Swift 's Look What You Made Me Do .	This list of most viewed online videos in the first 24 hours contains the top 30 online videos that received the most views within 24 hours of release across the world. This list excludes movie trailers , which are featured on the list of most viewed online trailers in the first 24 hours. The most viewed music video in this time period is Taylor Swift 's Look What You Made Me Do .	1.06	−0.48
when was last time england got to quarter finals in world cup	The England national football team have reached the quarter - finals on nine occasions, the latest of which were at the 2002 and the 2006 .	England did not enter the competition until 1950. . . Their best ever performance is winning the Cup in the 1966, whilst they also finished in fourth place in 1990, and in 2018. Other than that, the team have reached the quarter - finals on nine occasions, the latest of which were at the 2002 and the 2006 .	1.40	0.70

Table 8: Examples from the user study. The last column represent coefficients for preferring original sentence over the original paragraph, and the fourth column presents coefficients for decontextualized sentence over the paragraph. Positive values means preference towards the sentence-length answer over the paragraph-length answer.

edits were necessary to decontextualize. We use crowd-sourced annotations of decontextualized sentences. Figure 3 shows the screenshot of the user study interface.

Result Table 7 shows the results of the user study. We observe that decontextualized sentence answers are preferred to both the original sentence answers and the original paragraph answers. We also note that the users preferred sentence answer compared to paragraph answer in general.

We further investigated the statistical significance of the preferences reported in Table 7. We noticed a quite large amount of question and rater variability—some raters consistently preferred a sentence answer, valuing conciseness, while some raters behaved in the other direction. Similarly, for some questions, all raters preferred a sentence answer. Figure 4 visualizes such variability based on the questions and raters.

To control for the correlations induced by the rater and question groups, we fit a generalized linear mixed model (GLMM) using the `brm` R package (Bürkner, 2017). For this analysis, we excluded data points where users did not show a preference (selected either). We used the formula: `p∼1 + (1|r) + (1|q)`, where `p` is whether a rater chose one option over the other; `r` is the rater id; and `q` is the question id. This formula specifies a regression of the log-odds of the rater preference while allowing for random effects in the raters (`r`) and questions (`q`). The last column

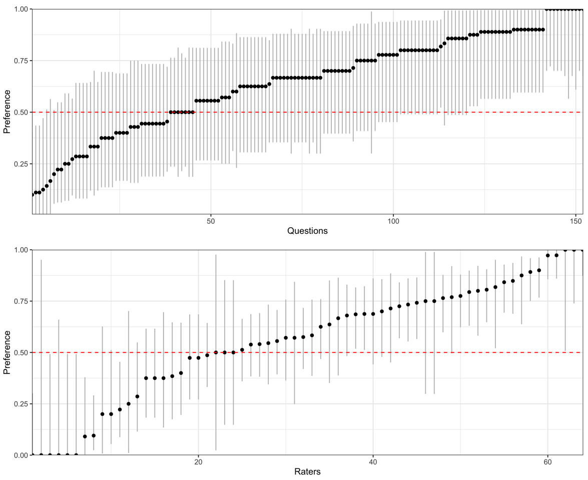


Figure 4: Each dot represents how frequently the decontextualized answer is preferred for a single question / rater to the original sentence for a single question (top plot) and single rater (bottom plot). Questions (top plot) and raters (bottom plot) are sorted by its preference towards the decontextualized answer. The red line is where both are equally preferred, and above the line represents question / rater where decontextualized answers were preferred. While the decontextualized answer is preferred overall, we see a large variability.

of Table 7 shows the fixed effect coefficients and their confidence intervals. The intercept represents the strength of preference towards option A. We found a statistically significant preference for decontextualized sentences over both original sentences and the paragraphs (p-value was smaller than 0.05 for both studies).

Examples We qualitatively investigated which examples benefit from decontextualization, and in which examples raters prefer paragraph answers. Table 8 shows questions together with two answer presentations, along with the predicted fixed effect question coefficient towards decontextualized answer in study (b) and towards the sentence answer in study (c). In the first row, the added information from the decontextualization is not relevant to the question, thus we observe preference *against* decontextualization. In the second and third row, the decontextualized sentence answer is preferred as it provides enough evidence to answer the query, while the original sentence answer does not.

6.2 Decontextualizing System Inputs

Having shown the benefits of decontextualization in a user-facing task, we now investigate the use of decontextualizaton as a preprocessing step. Specifically, we construct a passage retrieval corpus for open domain question answering (Chen et al., 2017) with decontextualized sentences. Experiment shows that decontextualized sentences ensure completeness of the passages while minimizing their length (thus computational cost).

Background Open domain question answering typically consists of pair a passage retrieval (Liu and Croft, 2002) and transformer-based answer extractor (reading comprehension model) based on the retrieved passages (Guu et al., 2020; Karpukhin et al., 2020; Izacard and Grave, 2020). And the computational cost is dominated by the cost of co-encoding the query with the retrieved passages (typically paragraphs or overlapping 100 word windows).

Setup We create a corpus using the 7k documents (233k paragraphs, 868k sentences) from the documents associated with the questions in the NQ-open development set (Lee et al., 2019). We consider a retrieved passage to be correct if it contains one of the answer strings[9] and investigate the number of questions for which we can retrieve a correct passage for a fixed computational cost. Under this measure, we compare paragraphs, windows of 100 words, sentences, and decontextualized sentences as a set of retrieval passages. These segmentation approaches generate different number of passages for the same article (paragraph and a window of 100 words segmentation make fewer passages compared to sentences-level segmentation). To generate decontextualized sentences process all paragraphs with T5-11B model, which are trained on all annotated data (including development and test set). For about 40% of sentences, the model classified the sentence as infeasible to decontextualize or unnecessary to make any edits, we use the original sentence. On the other 60% the model tended to add more information. For example, for a sentence ‘‘Bush was widely seen as a ‘pragmatic caretaker’ president who lacked a unified and compelling long-term theme in his efforts.’’, the decontextualized sentence will be ‘‘George H.W. Bush was widely seen as a ‘pragmatic caretaker’ president of the United States who lacked a unified and compelling long-term theme in his efforts.’’ A paragraph would be the entire paragraph containing this sentence, and a

[9] We adopt the answer match heuristics from Lee et al. (2019).

100-word window will be a chunk without using a sentence boundary as a segmentation. For all, we prepend the document title to the passage, following the literature and use the TFIDF as a retriever model.

Metric Let $\mathbf{q_i}$ be a question; let $\mathbf{A}_i = [a_i^0 \dots a_i^n]$ be the set of valid answers; let $\mathbf{C}_i = [\mathbf{c}_i^1 \dots \mathbf{c}_i^k]$ be a ranked list of evidence passages; and let $H(\mathbf{A}_i, \mathbf{C}_i)$ be the index of the top ranked context that contains one of the valid answers (or $k+1$ if there is no such context). We first define the cost of encoding a single question and passage, $c(\mathbf{q_i}, \mathbf{c_i^m}) = (|\mathbf{q_i}| + 1 + |\mathbf{c_i^m}|)^2$. This captures the fact that the Transformer's computation cost scales quadratically with the length of the encoded text (question + separator + evidence passage).

$$O(\mathbf{q_i}, \mathbf{A_i}, \mathbf{C_i}) = \sum_{m=1}^{H(\mathbf{A}_i,\mathbf{C}_i)} c(\mathbf{q_i}, \mathbf{c_i^m}).$$

Given the per example cost defined above, we define the ***recall*** of the retrieval system at computational cost budget t to be:

$$\frac{1}{N} \sum_{i=0}^{N} \mathbb{1}\left[O(\mathbf{q_j}, \mathbf{a_j}, \mathbf{C_j}) < t\right] \qquad (1)$$

where N is the total number of examples and $\mathbb{1}$ is an indicator function. We use this as an evaluation measure instead of mean reciprocal rank or recall at N, to compare across different retrieval passage length.

Results Figure 5 plots the recall of each retrieval corpus at different computational cost budget t on the whole NQ-open evaluation set. The graph shows that sentence level segmentation is more cost-effective than paragraph or 100-word level segmentation, and using decontextualized sentences is more cost effective than using the original sentences. Decontextualized sentences near the performance of commonly used 100-word windows with 1/10th the cost.

This result exemplifies the way in which decontextualization can be used to ensure that the input to natural language understanding system is concise yet complete. We think this way of using decontextualization as a preprocessing could also aid tasks such as summarization.

7 Related Work

Prior literature in summarization studied how article context affects the understanding of sentences within an article. It has been observed that disambiguating entity mentions and correctly resolving anaphora is crucial for automatic summarization (Otterbacher et al., 2002; Steinberger et al., 2007) and for evaluation of summarization systems (Pitler et al., 2010). Li et al. (2016) identified that information missing from a sentence could be identified in the article context in newswire text 60% of the time. This is considerably less frequent than for the encyclopedic text studied here, but nevertheless hints that decontextualization for newswire text could be feasible. It remains unclear whether information accessible in newswire contexts can be readily incorprated into sentences using controlled edits of the type we employ.

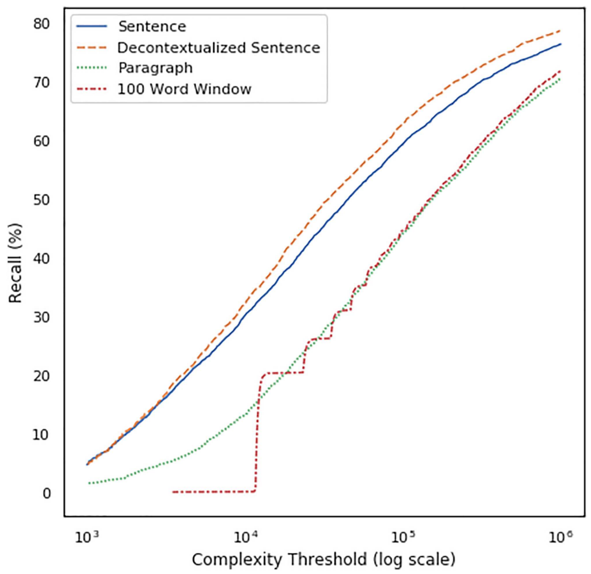


Figure 5: Retrieval recall plotted against computational cost budget (Eqn 1) for different methods of document segmentation.

Successful decontextualization models must resolve entity and event coreferences (Humphreys et al., 1997) as well as other forms of anaphora (Rösiger et al., 2018). These are necessary but insufficient for decontextualization however, which also involves discourse marker removal, acronym expansion, and fluent and grammatical sentence generation.

The term decontextualization was introduced in a recent table-to-text generation dataset (Parikh et al., 2020) where a sentence from a Wikipedia document was *decontextualized* such that it can be interpretable when presented with a table alone. They cover only the sentences that are relevant to the table, and adapt it to the table context. In a recent image captioning dataset (Sharma et al., 2018), sentences are re-written such that information that cannot be inferred from the image is removed. For example, entity names are replaced with generics (e.g., ⟅ -Tom Cruz , +A man ⟆ is waiting.'').

8 Conclusion

We define *decontextualization*, the task of rewriting a sentence from a document to be interpretable in an empty context, while preserving its meaning. We build a crowdsourced dataset and a model for decontextualization, and demonstrate how decontextualization can be used in a user-facing task and as a sub-component of an application system.

We believe that decontextualization will also be helpful in a wide range of other applications. For example, in multi-document summarization (Fabbri et al., 2019), co-referring entities and events must be resolved across different documents and removing ambiguous references may help; extractive summarization (Cheng and Lapata, 2016) could benefit from the type of pre-processing that we presented for open-domain QA; anaphora resolution is crucial for both summarization and machine translation (Susanne et al., 1992); and decontextualizing sentences may help in recovering explicit mentions of entities and relations which can help information extraction (Narasimhan et al., 2016). The current formulation focuses on the English encyclopedic corpus and rewriting for an empty context, and future work can explore different domains of text as well as mapping to a different context.

Acknowledgments

We would like to thank members of Google AI, especially Jacob Eisenstein, Kenton Lee, Santiago Ontanon, Ankur Parikh, Daniel Andor, Chris Alberti, and Slav Petrov for helpful discussions and comments. Lastly, we would like to thank our annotators.

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36:15–24. **DOI:** `https://doi.org/10.1609/aimag.v36i1.2564`

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Betty J. Birner. 2012. *Introduction to Pragmatics*, 1st edition. Wiley Publishing.

Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28. **DOI:** `https://doi.org/10.18637/jss.v080.i01`

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. *Proceedings of the Annual Meeting of the Association for Computation Linguistics (ACL)*. **DOI:** `https://doi.org/10.18653/v1/P17-1171`, **PMCID:** PMC5579958

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 484–494. Berlin, Germany, Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/P16-1046`, **PMCID:** PMC4738087

Herbert H. Clark. 1975. Bridging. In *Theoretical issues in natural language processing*. **DOI:** `https://doi.org/10.3115/980190.980237`, **PMID:** 1166311

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the Annual Meeting of the Association for Computation Linguistics (ACL)*. **DOI:** `https://doi.org/10.18653/v1/P19-1102`

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts*, volume 3 of *Syntax and Semantics*, pages 41–58. Academic Press, New York.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. **DOI:** https://doi.org/10.3115/1598819.1598830

K. Humphreys, R. Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. **DOI:** https://doi.org/10.3115/1598819.1598830

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. **DOI:** https://doi.org/10.1162/tacl_a_00300

V. Karpukhin, B. Oğuz, S. Min, L. Wu, S. Edunov, D. Chen, and W.-T. Yih. 2020. Dense passage retrieval for Open-Domain question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. **DOI:** https://doi.org/10.18653/v1/2020.emnlp-main.550

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association of Computational Linguistics*. **DOI:** https://doi.org/10.1162/tacl_a_00276

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint 1906.00300*.

Junyi Jessy Li, Bridget O'Daniel, Y. Wu, W. Zhao, and A. Nenkova. 2016. Improving the annotation of sentence specificity. In *LREC*.

X. Liu and W. Croft. 2002. Passage retrieval based on language models. In *CIKM '02*. **DOI:** https://doi.org/10.1145/584792.584854

Karthik Narasimhan, Adam Yala, and Regina Barzilay. 2016. Improving information extraction by acquiring external evidence with reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. **DOI:** https://doi.org/10.18653/v1/D16-1261

Jahna Otterbacher, Dragomir R. Radev, and Airong Luo. 2002. Revisions that improve cohesion in multi-document summaries: a preliminary study. In *Proceedings of the Annual Meeting of the Association for Computation Linguistics (ACL)*. **DOI:** https://doi.org/10.3115/1118162.1118166

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, abs/2004.14373. **DOI:** https://doi.org/10.18653/v1/2020.emnlp-main.89

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. **DOI:** https://doi.org/10.1162/tacl_a_00293

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the Annual Meeting of the Association for Computation Linguistics (ACL)*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. ConLL-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *EMNLP-CoNLL Shared Task*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019a. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint 1910.10683*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019b. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Ina Rösiger, Arndt Riester, and Jonas Kuhn. 2018. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computation Linguistics (ACL)*. **DOI:** `https://doi.org/10.18653/v1/P18-1238`, **PMCID:** PMC6266124

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Harvard University Press, USA.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Jeek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43(6):1663–1680. **DOI:** `https://doi.org/10.1016/j.ipm.2007.01.010`

Preusz Susanne, Birte Schmitz, Christa Hauenschild, and Carla Umbach. 1992. *Anaphora resolution in machine translation*.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297. **DOI:** `https://doi.org/10.1162/tacl_a_00139`

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415. **DOI:** `https://doi.org/10.1162/tacl_a_00107`