DeModify: A Dataset for Analyzing Contextual Constraints on Modifier Deletion

Vivi Nastase[◇], Devon Fritz, Anette Frank[◇]
Leibniz ScienceCampus "Empirical Linguistics and Computational Language Modeling"[◇]
Institute of Computational Linguistics, Heidelberg University
Heidelberg, Germany
{nastase, frank}@cl.uni-heidelberg.de, devon.s.fritz@gmail.com

Abstract

Tasks such as knowledge extraction, text simplification and summarization have in common the fact that from a text fragment a smaller (not necessarily contiguous) portion is obtained by discarding part of the context. This may cause the text fragment to acquire a new meaning, or even to become false. The smallest units that can be considered disposable in a larger context are modifiers. In this paper we describe a dataset collected and annotated to facilitate the study of the influence of modifiers on the meaning of the context they are part of, and to support the development of models that can determine whether a modifier can be removed without undesirable semantic consequences.

Keywords: annotation, modifier deletion, text simplification

1. Introduction

Knowledge extraction, keyword identification, text simplification, or summarization are useful tasks that rely on the assumption that certain information from texts can be discarded without negative consequences. Certain details are peripheral and can be disregarded. Removing subordinate clauses is a common practice in extractive summarization, for example (Vanderwende et al., 2007; Zajic et al., 2007).
While modifiers can be quite complex – ranging from a single word to a full clause – we focus here on single-word modifiers, the smallest unit of context that can be removed. While even their name suggests a peripheral role in the overall meaning of a text, modifiers can impact both their local and global context. The impact of modifiers on their local context (themselves plus their syntactic head), also called modification distortion (Murphy, 2002), has been studied mostly out of context, and there are several datasets that allow for the study of this phenomenon (Kruszewski and Baroni, 2014; Schulte im Walde et al.,). The dataset we built and present here is focused on the effect of modifiers on the larger context. For example, the following title of a TED talk would mean something completely different should the modifier *nearly* be removed:

> AJ Jacobs: How healthy living nearly killed me

While *nearly* could be argued to be a special modifier, the same situation may arise for "normal" modifiers, such as the adjective *old* – it can be removed from the following context without a dramatic impact on the meaning of the sentence:

> Then she saw the old parish priest pull up in his car.

but it is an essential element of the story, when the entire context is included – a short story from the ROC corpus (Mostafazadeh et al., 2016):

> Joan entered the confessional and kneeled. She thought she was confessing to the old parish priest. Joan confessed she had fantasized about the young visiting priest. Joan felt relief as she left the confessional. Then she saw the old parish priest pull up in his car.

Understanding the influence of modifiers is important, as it affects compositional models of language, as well as higher-level tasks such as summarization or textual entailment. As seen in the above examples, modifiers, while syntactically omissible, can make important semantic contributions to the information conveyed by a larger context, such that deleting them may considerably alter the meaning of the sentence and its context.
In this paper we describe a dataset of complete short texts (approximately 5 sentences each) in which one open-class modifier has been annotated with one of three classes based on its impact on the text it appears in: crucial, not-crucial, ungrammatical. The dataset consists of 3632 instances, which we provide with their multiple annotations obtained through CrowdFlower. We describe in the paper two potential gold standards – one obtained by using only instances where all annotators agree consisting of 1767 instances, and one obtained with majority voting, consisting of 3542 instances. We also include a split into 5 folds, to be used for future experiments.[1]

2. Related work

Text simplification can occur at different levels of granularity – extracting a sentence from a document, deleting a phrase from a sentence or a sub-phrase from a larger chunk. Vanderwende et al. (2007), Zajic et al. (2007) propose syntax-based trimming, where branches of a syntactic tree are scored using a combination of features that

[1] http://www.cl.uni-heidelberg.de/english/research/downloads/resource_pages/deModify/deModify_data.shtml

marks them for potential deletion. Wubben et al. (2012) approach the problem of text simplification as a machine translation problem trained on pairs of texts from Wikipedia and SimpleWikipedia. Wang et al. (2016), Zhang and Lapata (2017) reformulate this approach in the form of neural encoder-decoder models.

Focusing on the modifiers, modification can be viewed from many different perspectives. From a linguistic point of view, several typologies of modifiers have been proposed (McNally, 2013). Of these, the semantic impact of modifiers is taken into account in: (a) the *entailment-based* typology, in which modifiers are grouped into three broad categories based on the inferences they license, which stem from potential interpretations of the extension of modifiers, head nouns and the compounds as sets: *intersective* modifiers (*male nurse*), *subsective* modifiers (*molecular scientist*), *intensional* modifiers (*alleged crook*); (b) *pragmatic/discourse-related* typologies which partition modifiers based on their impact on the utterance in which they appear – those that affect the interpretation of the utterance, and those that do not, and modifiers are considered separately by POS or phrase type.

The entailment-based typology is the focus of Amoia and Gardent (2007), who study the inferential properties of adjectives and how these classes influence the omissibility of adjectives under truth-conditional aspects. Amoia and Gardent (2008) published a data set where these and other syntactic and semantic properties of adjectives are tested in an RTE (recognizing textual entailment) setting. In this work, the *context* in which the semantic effects of adjectives are tested is the sentence, and the relevant criterion is preservation of truth when, e.g., deleting the adjective or the head noun as in the following inference pairs: *Daisy is a big mouse* $\rightarrow$ *Daisy is a mouse* or *Daisy is a big mouse* $\rightarrow$ *Daisy is big*. Along similar lines, Stanovsky and Dagan (2016) describe the construction process and resulting dataset of non-restrictive noun phrase modification. Non-restrictive modifiers – e.g. *The speaker thanked president Obama who just came into the room* – can be removed to shorten sentences. The dataset gathered and annotated trough crowd sourcing has no restrictions on the length of the modifiers, which often span phrases. The context provided for annotation is a sentence for each modifier.

From a conceptual point of view, modifiers were studied with respect to the distortion effect they have on the concept denoted by the head noun (Murphy, 2002). Kruszewski and Baroni (2014) focus on the computational study of the effect of *modifier-triggered (head) distortion*, and built and annotated a dataset of (out of context) noun compounds, with respect to their "place" in a hierarchy of concepts. Each modifier-head compound (e.g., *perfume bottle*) is rated with membership and typicality scores against 3 criteria: how well it fits under the concept denoted by the head (*perfume bottle* $\rightarrow$ *bottle*), how well it fits under a superconcept of the head (*perfume bottle* $\rightarrow$ *drinkware*), and how prototypical the concept denoted by the head is of the super concept (*bottle* $\rightarrow$ *drinkware*). The final ratings are averages over individual scores gathered through surveys on CrowdFlower. The data thus collected is tested from the point of view of compositionality, expecting that more typical instances of a class are modelled more successfully using compositional operations on the individual vectors. Schulte im Walde et al. () built a dataset of 868 German noun-noun compounds, where one of the annotations quantifies the compositionality of the compound on a scale of 1 (semantically opaque) to 6 (semantically transparent).

In contrast to Kruszewski and Baroni (2014) and Schulte im Walde et al. (), our dataset focuses on the semantic effects that modifier deletion may have on the wider context, i.e., is not limited to the modified phrase. Modifiers in our dataset may be considered crucial to the larger context in which they appear, although they may not have a distorting effect on their local syntactic head – e.g. the adjective *old* in the noun phrase *old parish priest* ("old parish priest" is still "parish priest") – nor within the full sentence.

We also go beyond Amoia and Gardent (2008)'s and Stanovsky and Dagan (2016) work in that we consider a full (short) story context for judging omissibility – the example with the adjective *old* in the Introduction section shows that the sentence context (where the deletion might be deemed acceptable) can be overriden by the larger context. Furthermore, we judge *informativeness* as opposed to *preservation of truth*, which yields a more natural criterion for omissibility in e.g. text simplification tasks.

3. The *DeModify* Dataset: Data Selection, Annotation Process and Data Statistics

3.1. Data selection and annotation categories

The problem that data selection poses is that the phenomenon we target is not explicitly marked. A random selection of texts may not contain a significant number of positive instances. We follow a deliberate strategy, inspired by Grice's conversational maxims (Grice, 1975), and target collections of short texts with a specific communication intent, which are likely to contain (mostly) relevant and important details. Furthermore, we choose short self-contained texts as a basis for our annotations to ensure that at the onset we have a complete context which is easy and fast to read for the annotators. In these texts we mark one modifier (following the methodology described below), and we present the text and the marked modifier to annotators through CrowdFlower[2]. The annotators are asked to assign the marked modifier one of three categories:

crucial – the modifier is crucial for preserving the intended meaning, removing it leads to a different/erroneous/false interpretation of the remaining text:

> ... She thought she was confessing to the old parish priest. Joan confessed she had fantasized about the young visiting priest. ... Then she saw the old parish priest pull up in his car.

not crucial – the modifier is not crucial, removing it would not result in a distortion of the meaning of the remaining text:

> ... The trading of information is obviously driven by greed of gain. ...

[2] https://www.crowdflower.com

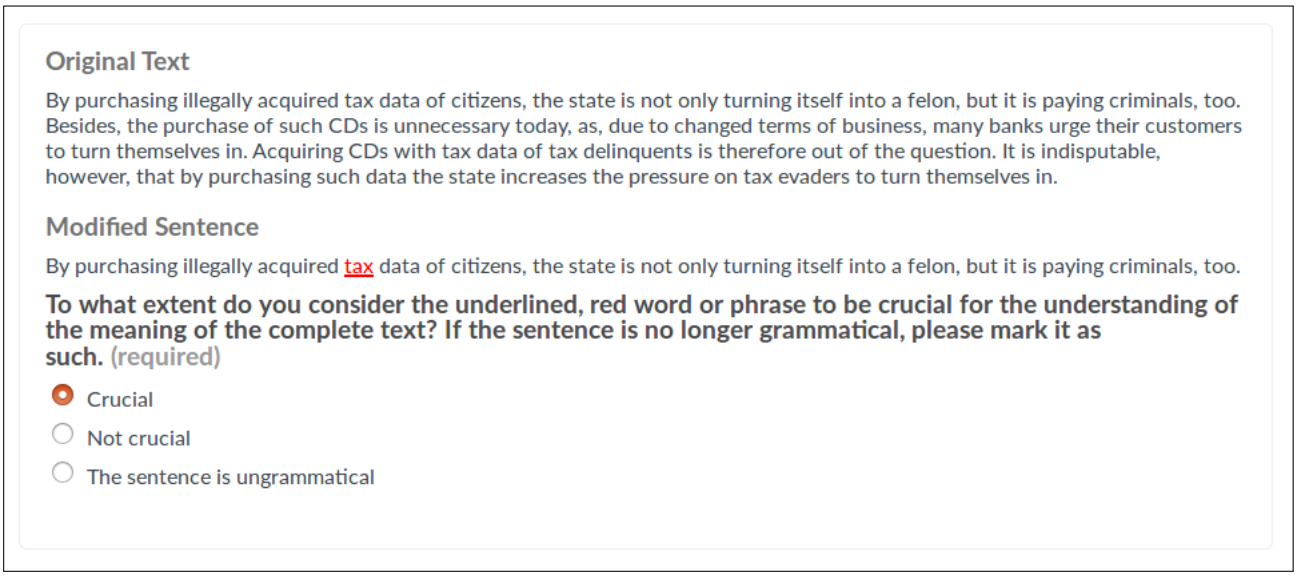


Figure 1: CrowdFlower interaction for annotation.

dependency relation	frequency in ROCStories	frequency in microtexts	examples
advmod	48,477	317	She is happy that she finally baked a cherry pie.
amod	69,066	576	Josh patted himself on the back for making a good decision.
compound	37,714	267	Bella made dessert for her family dinner.
nmod:tmod	9,140	10	Katarina lost her first tooth yesterday.

Table 1: Examples for the four selected dependency relations for modifier selection.

ungrammatical – removing the modifier results in an ungrammatical sentence:

> ... this ideal appears unworthy of support in many ways. ...

We selected instances in our dataset from short stories from the ROCStories corpus[3] (Mostafazadeh et al., 2016) and the short argumentative texts[4] (Peldszus and Stede, 2016). All these texts are approximately 5 sentences long, and self-contained.

Modifier selection is based on a syntactic analysis of the texts: the Stanford Dependency Parser is used to parse the data, and we compute statistics on the frequency of grammatical dependencies other than root/subject/object that connect open-class words. We select the 4 most frequent dependency relations, presented in Table 1.

Based on frequency statistics for the modifier lemmas that occur in the chosen dependency relations, they are split into 5 frequency bands. We then chose at most 15 instances for a random selection of modifiers from each band for inclusion in the final dataset, which contains a total of 3632 instances:

ROCStories – 3026 instances. The ROCStories corpus (Mostafazadeh et al., 2016) is a collection of about 49,000 self-contained stories of at most 5 sentences. From this large dataset we selected approximately 4000 stories. After the filtering criteria mentioned above, 3026 stories with one marked modifier from each story were kept.

Argumentative Microtexts – 606 instances. The argumentative microtexts corpus (Peldszus and Stede, 2016) consists of 112 texts of 3-5 sentences each. We selected more than one modifier from each of these texts according to the same process as for the ROCStories.

3.2. Annotation Process

This raw dataset is presented to CrowdFlower users. For each instance the complete text of the story/microtext is shown, and below it the targeted modifier is shown in red in its sentence context. The users are given three choices for annotation – *crucial (C), not crucial (N), ungrammatical (U)*. Before the actual annotation exercise started, the CrowdFlower users were given instructions and shown both positive and negative commented examples. A partial screenshot of the annotation interaction is shown in Figure 1 and the instructions for the annotation process are included in Figure 3 in the Appendix.

The set-up was first tested in several iterations (on friends and family). We started with 5+1 options: 5 options ranging in stages from *modifier deletable without consequences* to *modifier not deletable* + one additional label: *ungrammatical*. However, we found that the task was clearer with a binary decision regarding the impact of the modifier being *crucial* vs. *not crucial*, plus the ungrammaticality option.

The data was released in batches of 100. 32 control in-

[3] http://cs.rochester.edu/nlp/rocstories

[4] https://github.com/peldzsus/arg-microtexts

label combination	frequency	percentage
C C C	442	12.2%
N N N	1248	34.4%
U U U	77	2.1%
C C N	566	15.6%
C C U	132	3.6 %
N N C	843	23.2%
N N U	86	2.4%
U U C	120	3.3%
U U N	28	0.8%
C N U	89	2.4%

Table 2: Agreement percentages for the different label combinations: strict agreement: rows 1-3.

strict GS: 1089 unique heads, 399 unique modifiers

microtexts	265	N	174	U	7	C	84
ROCStories	1502	N	1074	U	70	C	358
all	1767	N	1248	U	77	C	442
in %	100	N	70.6	U	4.4	C	25.0

relaxed GS: 1773 unique heads, 585 unique modifiers

microtexts	587	N	336	U	32	C	219
ROCStories	2955	N	1841	U	193	C	921
all	3542	N	2177	U	225	C	1140
in %	100	N	61.4	U	6.4	C	32.2

Table 3: Gold standard datasets statistics: number of instances from each text source, and annotation frequencies.

stances – where the answer was known (annnotated by one of the authors who is a native speaker of English) – were included in the batches, and CrowdFlower's internal mechanisms were used to filter out annotators with low levels of (automatically computed) trust. In the end each instance was annotated by at least three CrowdFlower users. One of the authors (native English speaker) annotated the control instances and performed an additional evaluation on a preliminary test run of the system on 100 instances. On a random selection of 20 instances our annotator confirmed agreement with the judges (not all of whom are native English speakers) in 17 of the 20 cases. Table 2 shows the agreement counts for each combination of labels in the data.

3.3. Data Statistics

We derived two gold standard versions from the annotations: one by **strict** agreement (all annotators agree), and one by **relaxed** agreement (majority voting). The distribution of the instances into annotation classes for both versions, divided by text sources, is presented in Table 3.

The **strict GS** is about half the size of the **relaxed GS**. The proportion of categories (N, C, U) is comparable in both versions, with N(oncrucial) covering about 70/60 percent of the instances, followed by C(rucial) with about a quarter/third of the instances and a small U(ngrammatical) class. Table 2 shows the detailed counts for all annotation

gold standard	nb. of classes		modifiers	heads
strict	1		253	939
		C	88	209
		N	156	688
		U	9	42
	2		123	137
		C-N	110	122
		C-U	4	3
		N-U	9	12
	3		23	13
relaxed	1		332	1364
		C	132	408
		N	181	884
		U	19	72
	2		173	365
		C-N	154	306
		C-U	8	22
		N-U	11	37
	3		80	44

Table 4: Number of classes (out of our three – C, N, U) in which modifiers and heads appear, for strict and relaxed (majority voting) gold standards.

combinations.

The purpose for constructing this dataset was to facilitate the study of the impact of context on the deletability of modifiers. For a robust study it would be interesting to have a spread of the instances over the three classes, or rather, mostly *crucial* and *not-crucial* (which as Table 3 shows is the case), but it would also be extremely interesting if the same modifiers appear with different class annotations. This is an issue that we could not control for during annotation, but we have tested the resulting dataset whether such phenomena were captured. If the dataset includes modifiers and heads that appear in more than one class, an automatic system would be forced to take into account the context for prediction. The plots in Figure 2 and the statistics in Table 4 confirm that this is accomplished: 30% (39%) of the modifiers appear in both the *crucial* and *not-crucial* classes for the strict (relaxed) gold-standard. This also confirms the point we made in the Introduction, that the same modifiers will behave differently in different contexts with respect to the phenomenon targeted here.

An example from the annotated data of the same modifier belonging to different classes in different contexts is included in Table 5 for the adjective *long*.

We noted a certain degree of confusion between the *ungrammatical* and *crucial* for the adjective *long*: when it appears in the phrase *as long as* it is sometimes annotated as *crucial*, and sometimes as *ungrammatical*.

3.4. Data files

We provide an archive through our institute's website (`http://www.cl.uni-heidelberg.de/english/research/downloads/resource_pages/deModify/deModify_data.shtml`) with two main files: the dataset with CrowdFlower annotations, and a file with a proposed split into 5 folds.

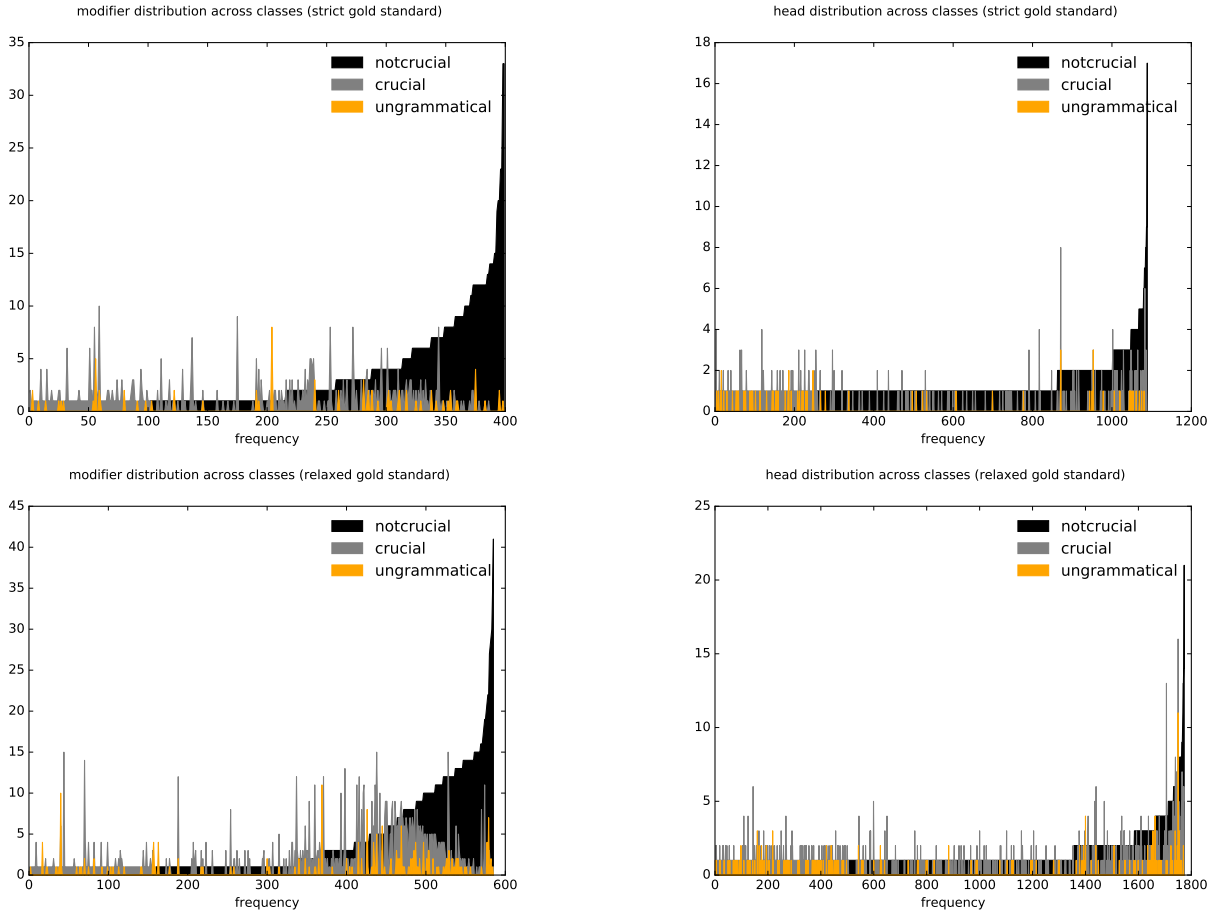


Figure 2: Histogram for modifier (left) and head (right) frequencies by annotation classes, when considering strict (all annotators agree, top) or relaxed (majority voting, bottom) gold standards: every point on the x axis corresponds to a word, and the colored bars show its frequency in each of the three classes (words on the x axis are ordered by increasing frequency in the *not-crucial* (N) class).

long: instances in different classes	
crucial	However, a death would not be of any more use to those affected and their relatives than if the felon receives a **long** sentence. (instance id: 1101714199)
not-crucial	They loved to go on **long** walks together. (instance id 1102590593)
ungrammatical	We put on **long** sleeves and jackets. (instance id 1102590621)
	You should watch less television, as too much TV makes you stupid in the **long** run, like your brother. (instance id 1101713906)

Table 5: Instances in different classes for the adjective **long**

The data (file demodify.tsv) consists of 3632 entries on 10896 lines, each entry consisting of 3 lines which provide the information listed in Table 6 (all lines provide the same information).

We include a file (demodify.data_split.tsv) that gives a proposed split of the data into 5 balanced folds (with respect to the classes). The file contains an assignment to fold for both the strict (full agreement) and relaxed (majority agreement) class assignments.

4. Conclusion

We described a novel dataset for the study of the influence of single-token modifiers on the larger textual context. The DEMODIFY dataset consists of short (up to 5 sentences) self-contained texts, in which selected modifiers are marked and annotated with one of three categories: *crucial* (removing them would distort the meaning of the remaining text), *not crucial* (the modifier can be removed without drastic consequences) or *ungrammatical* (removing the modifier would result in an ungrammatical sentence). The final dataset consists of 3632 instances annotated through CrowdFlower by at least 3 judges. Control instances and CrowdFlower trustiness measures were used to filter the judges and ensure high-quality annotations.

The DEMODIFY dataset is publicly available[5] It can be

[5] http://www.cl.uni-heidelberg.de/english/research/downloads/resource_pages/deModify/

_unit_id	unique instance id for each entry, shared by all lines that contain annotations for this entry
_created_at	time of creation
annotation	one of three classes: crucial, not crucial, ungrammatical
head	a head word
head_word_index	the position of this word in the sentence
modifier	the modifier that is marked for deletion, whose head is in the "head" column
modifier_type	the dependency relation between the head and the modifier
sentence_number	the sentence number relative to the full text
source	the name of the source corpus: microtexts/ROCStories
storyid	the id of the story/microtext
title	title of the story/microtext – if there is one
full_story	the full text of the story/microtext
original_sentence	the sentence in which the modifier appears
_trust	the trust measure of the annotator, computed automatically by CrowdFlower
_country	the country of the annotator
_region_city	the city of the annotator

Table 6: Data file description

used to investigate *linguistic factors* underlying the observed behaviour of modifiers in context, the range of influence of the modifier on a context and the interacting elements of the context. Ultimately we hope this dataset will contribute towards a better understanding of pragmatic influences in text semantics.

5. Acknowledgments

We thank our reviewers for their comments. This research was funded by the Leibniz Science Campus Empirical Linguistics & Computational Language Modeling, supported by Leibniz Association grant no. SAS-2015-IDS-LWC and by the Ministry of Science, Research, and Art of Baden-Württemberg.

6. Bibliographical References

Amoia, M. and Gardent, C. (2007). A first order semantic approach to adjectival inference. In *Workshop on textual entailment and paraphrasing (WTEP)*, Prague.

Amoia, M. and Gardent, C. (2008). A testsuite for inference involving adjectives. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Grice, H. P. (1975). Logic and conversation. In P. Cole et al., editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. Academic Press, New York, N.Y.

Kruszewski, G. and Baroni, M. (2014). Dead parrots make bad pets: Exploring modifier effects in noun phrases. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 171–181, Dublin, Ireland.

McNally, L. (2013). Modification. In Maria Aloni et al., editors, *Cambridge Handbook of Semantics*. Cambridge University Press.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J. (2016). A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California.

Murphy, G. (2002). *The Big Book of Concepts*. MIT Press.

Peldszus, A. and Stede, M. (2016). An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015*, pages 801–815, London. College Publications.

Schulte im Walde, S., Hätty, A., Bott, S., and Khvtisavrishvili, N.). GhoSt-NN: A Representative Gold Standard of German Noun-Noun Compounds. In *Proceedings of the 10th Conference on Language Resources and Evaluation (LREC)*, Portoroz, Slovenia.

Stanovsky, G. and Dagan, I. (2016). Annotating and Predicting Non-Restrictive Noun Phrase Modifications. In *Proc. of the 54th Ann. Meeting of the Assoc. for Computational Linguistics*, pages 1256–1265, Berlin, Germany.

Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. (2007). Beyond SumBasic: Task-focused Summarization with Sentence Simplification and Lexical Expansion. *Information Processing and Management*, 43(6):1606–1618, November.

Wang, T., Chen, P., Rochford, J., and Qiang, J. (2016). Text Simplification Using Neural Machine Translation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 4270–7271. AAAI Press.

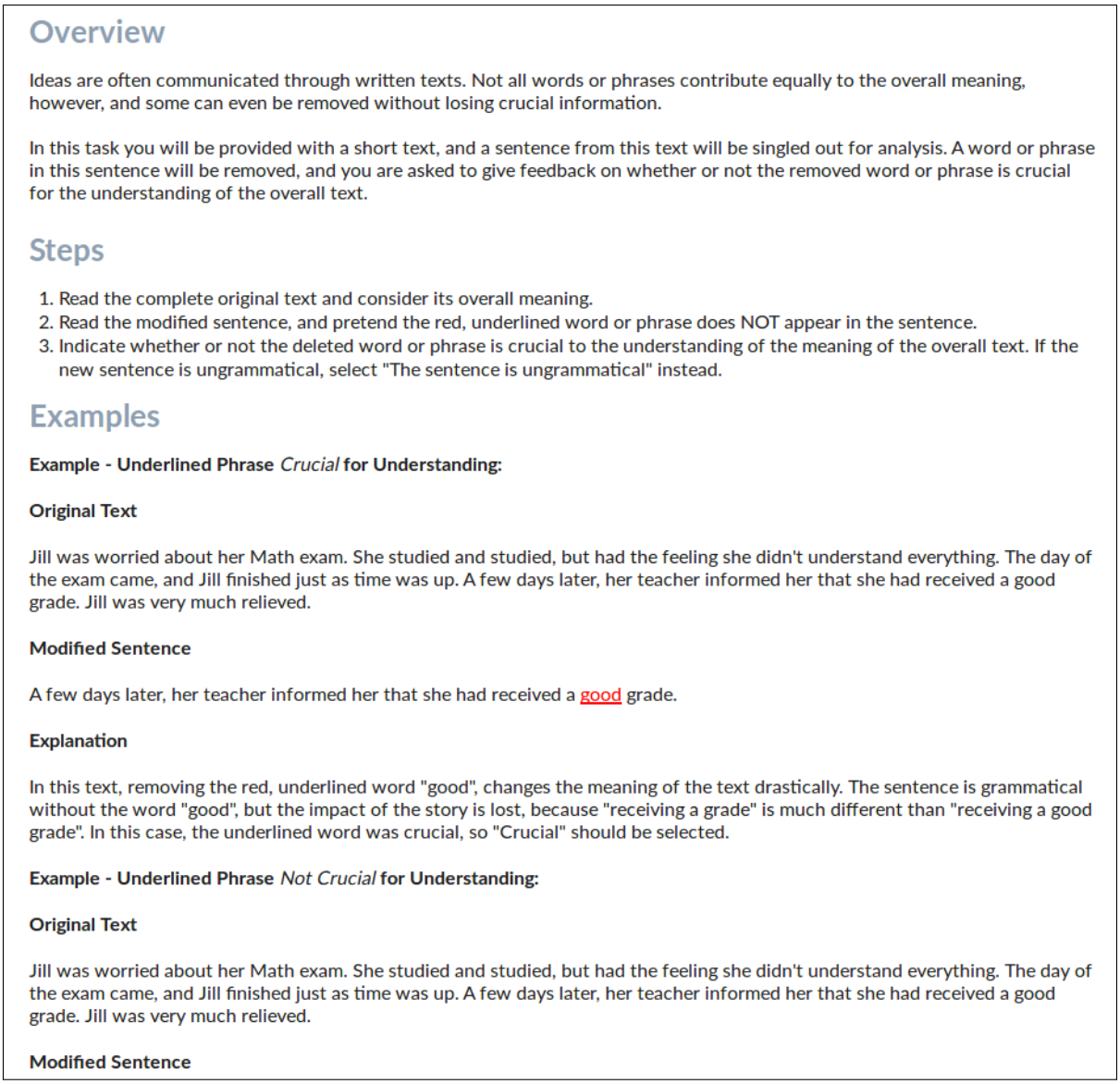
Wubben, S., Van Den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 1015–1024.

Zajic, D., Dorr, B., Lin, J., and Schwartz, R. (2007). Multi-Candidate Reduction: Sentence Compression as a Tool for Document Summarization Tasks. *Information Processing and Management*, 6(43):1549–1570.

Zhang, X. and Lapata, M. (2017). Sentence Simplification with Deep Reinforcement Learning. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Copenhagen, Denmark.

deModify_data.shtml

Appendix



Overview

Ideas are often communicated through written texts. Not all words or phrases contribute equally to the overall meaning, however, and some can even be removed without losing crucial information.

In this task you will be provided with a short text, and a sentence from this text will be singled out for analysis. A word or phrase in this sentence will be removed, and you are asked to give feedback on whether or not the removed word or phrase is crucial for the understanding of the overall text.

Steps

1. Read the complete original text and consider its overall meaning.
2. Read the modified sentence, and pretend the red, underlined word or phrase does NOT appear in the sentence.
3. Indicate whether or not the deleted word or phrase is crucial to the understanding of the meaning of the overall text. If the new sentence is ungrammatical, select "The sentence is ungrammatical" instead.

Examples

Example - Underlined Phrase *Crucial* for Understanding:

Original Text

Jill was worried about her Math exam. She studied and studied, but had the feeling she didn't understand everything. The day of the exam came, and Jill finished just as time was up. A few days later, her teacher informed her that she had received a good grade. Jill was very much relieved.

Modified Sentence

A few days later, her teacher informed her that she had received a good grade.

Explanation

In this text, removing the red, underlined word "good", changes the meaning of the text drastically. The sentence is grammatical without the word "good", but the impact of the story is lost, because "receiving a grade" is much different than "receiving a good grade". In this case, the underlined word was crucial, so "Crucial" should be selected.

Example - Underlined Phrase *Not Crucial* for Understanding:

Original Text

Jill was worried about her Math exam. She studied and studied, but had the feeling she didn't understand everything. The day of the exam came, and Jill finished just as time was up. A few days later, her teacher informed her that she had received a good grade. Jill was very much relieved.

Modified Sentence

Figure 3: CrowdFlower instructions including positive and negative examples for annotation.