Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3319–3330
Marseille, 20-25 June 2022

Clarifying Implicit and Underspecified Phrases in Instructional Text

Talita Anthonio* **Anna Sauer*** **Michael Roth**
University of Stuttgart
Institute for Natural Language Processing
{talita.anthonio,anna.sauer,michael.roth}@ims.uni-stuttgart.de

Abstract

Natural language inherently consists of implicit and underspecified phrases, which represent potential sources of misunderstanding. In this paper, we present a data set of such phrases in English from instructional texts together with multiple possible clarifications. Our data set, henceforth called CLAIRE, is based on a corpus of revision histories from wikiHow, from which we extract human clarifications that resolve an implicit or underspecified phrase. We show how language modeling can be used to generate alternate clarifications, which may or may not be compatible with the human clarification. Based on plausibility judgements for each clarification, we define the task of distinguishing between plausible and implausible clarifications. We provide several baseline models for this task and analyze to what extent different clarifications represent multiple interpretations as a first step to investigate misunderstandings caused by implicit/underspecified language in instructional texts.

Keywords: Corpus, Language Modeling, Semantics, Other

1. Introduction

Natural language inherently consists of elements that are implicit or underspecified because it is assumed that they can be inferred by the reader. For example, anaphoric references are often omitted when the referent is clear from the discourse context (Fillmore, 1986, henceforth *implicit references*). Similarly, relevant aspects of meaning may also be unspecified in explicit references, as is the case of pronominal or metonymic referring expressions. As a consequence, inferences can be difficult for a reader when different interpretations seem plausible. For instance, consider the instructional text in Table 1. Here, different clarifications could be made to specify the implied information, represented by the blank ___, as terms such as *body posture*, *walking posture* and *core posture* could all make sense in the given context. However, these clarifications are not semantically equivalent.

Various NLP tasks deal with settings in which models have to predict the most likely continuation of a text, such as the story cloze task (Mostafazadeh et al., 2016), the LAMBADA word prediction test (Paperno et al., 2016), or the HellaSwag sentence completion task (Zellers et al., 2018). Models in such tasks are typically evaluated by their capability of detecting the correct filler. This ignores the notion that different completions could be plausible, as discussed by Anthonio and Roth (2021) for the case of implicit references. To bridge this gap, we believe that it is important to create resources that accurately reflect that multiple plausible completions or clarifications may exist for implicit/underspecified elements in language.

Another contribution of such a resource would be to facilitate research on misunderstandings that arise from texts, which has received little attention in our community so far. Misunderstandings can be viewed as diverging interpretations of an utterance (Macagno, 2017; Yang et al., 2010a; Yang et al., 2010b), which can be denoted by semantically incompatible clarifications of an implicit or underspecified element.

How to Walk Gracefully				
Improving your posture (. . .) 5. Use your core. (. . .) 6. Practice. In order to perfect your ___ posture, you will need to devote some time to practicing.				
✓body	✓walking	✓core	✗ gym	✗ target

Table 1: An example from our data set, consisting of multiple plausible (✓) and implausible (✗) clarifications for a sentence in its discourse context.

In this paper, we address that challenge by providing **CLAIRE**[1], a large collection of sentences with plausible and implausible clarifications in instructions. The instructions are taken from wikiHow[2], a website where users can collaboratively write and revise how-to guides. Previous studies on wikiHow revisions have showed that such edits can have clarifying functions (Debnath and Roth, 2021; Anthonio and Roth, 2021). Therefore, we use wikiHowToImprove (Anthonio et al., 2020), a data set with sentences and their revised versions, as a basis for CLAIRE. We use the implicit references from our previous studies, and create several additional subsets, consisting of fused heads, metonymic references and generic nouns (which are revised to compound nouns) in the original sentence.

Since we want to investigate cases with multiple plau-

* The first two authors contributed equally.

[1] **Cla**rifying **I**nsertions from **R**evision **E**dits, available at https://github.com/acidAnn/claire

[2] https://www.wikihow.org

sible clarifications, we automatically generate artificial revisions in addition to an editor's manual revision using a cloze test setup. We then ask human annotators to rate the semantic plausibility of each clarification in the given discourse context.

We show that CLAIRE can be used to train computational models for distinguishing between plausible and implausible clarifications of an instruction by providing several baselines for this task. In addition, we demonstrate that our resource contains diverging interpretations of underspecified/implicit elements in instructional texts, which can be used to investigate misunderstandings.

In sum, we make the following contributions:

- We introduce the task of distinguishing between plausible and implausible clarifications in instructional texts, for which we create and release a data set of underspecified and implicit elements (§3).
- We extract and analyze cases for which multiple plausible clarifications exist and shed more light on the differences and similarities of the corresponding diverging interpretations (§4).
- We provide several baseline models for the task of distinguishing plausible clarifications from implausible clarifications (§5).

2. Related Work

In this section, we discuss related studies on revision histories of wikiHow articles (§2.1), on sources of misunderstandings and clarifications thereof (§2.2), and on models for resolving underspecified and implicit language phenomena (§2.3).

2.1. wikiHowToImprove

The wikiHowToImprove corpus was introduced by Anthonio et al. (2020), who defined the task of distinguishing between older and newer versions of a sentence based on the revision histories of wikiHow articles. They provided two baseline models for this task. However, many of the revisions were found to be made for purposes other than clarification, such as spelling or grammar correction. Therefore, the insights on modeling revisions with clarifying functions were limited. To address this gap, two follow-up studies focused on a subset of revisions that follow specific patterns: Anthonio and Roth (2020) analyzed substitutions of nouns with semantically related nouns and showed that such substitutions often increased specificity. In a similar vein, Debnath and Roth (2021) considered verb substitutions and showed that many of them resulted in a more specific or more focused perspective. Both studies included computational experiments that demonstrated the feasibility of modeling the differences before and after a clarifying revision.

2.2. Misunderstandings and Clarifications

Most work on causes of misunderstanding is based on dialogues (McRoy and Hirst, 1993, inter alia). There are only few studies on the computational modeling of misunderstandings that arise from texts (Yang et al., 2010a; Yang et al., 2010b). Specifically, these studies focus on the detection of *nocuous ambiguities*, which Yang et al. define as instances of coordination or anaphora ambiguity that give rise to diverging interpretations. Yang et al. obtain promising results with classical machine learning algorithms using features such as semantic similarity and collocation frequency. Our work is similar in that we view misunderstandings as diverging interpretations. However, our focus lies on implicit and underspecified language phenomena, for which different interpretations can be observed via clarifications and which may not reflect a specific type of ambiguity.

Another body of work examines the effect of misunderstandings, in particular how they are resolved. A clarification can be seen as a revision that is applied to an utterance. It resolves (or prevents) potential misunderstandings by making the intended interpretation of the utterance explicit. This is linked to the notion of repair (Schegloff et al., 1977) in dialogue (Purver et al., 2018; Marge and Rudnicky, 2019). Some previous research in NLP has focused on generating clarification questions that ask for missing information (Rao and Daumé, 2018), e.g., in human-machine dialogue (Khalid et al., 2020; Aliannejadi et al., 2021), in information retrieval settings like question answering (Xu et al., 2019) and conversational search (Bi et al., 2021; Sekulic et al., 2021) or as feedback for human-edited texts (Majumder et al., 2021; Zhang and Zhu, 2021). In contrast, there are no explicit clarification questions in CLAIRE. Instead, the resource focuses on different clarifications by themselves, which might be seen as answers to (hypothetical) clarification questions. More similar to the notion of a clarification as a revision, AmbiQA (Min et al., 2020) introduced the QA-related task of identifying questions with multiple plausible answers and rewriting the questions such that there is a unique answer for each interpretation.

2.3. Implicit and Underspecified Language

The implicit and underspecified phenomena addressed in this work have been the focus of a growing body of related work in NLP, specifically in tasks on how to recover missing elements. For example, Elazar and Goldberg (2019) addressed the task of finding the missing head of numeric fused-heads with a model inspired by coreference resolution. Similarly, metonymy interpretation or resolution is the task of determining the hidden intended entity or event that a metonymic expression refers to (Utiyama et al., 2000; Lapata and Lascarides, 2003; Shutova, 2009; Zarcone et al., 2012; Chersoni et al., 2017). Rambelli et al. (2020) and Pedinotti and Lenci (2020) both use masked language modeling to recover the specific hidden entity or event, which resembles our approach. For noun compounds, instead of a recovery of a missing element, there is re-

lated work by Günther and Marelli (2016) and Dhar and van der Plas (2019) on predicting how plausible a noun–noun composition is. However, none of the studies have considered the possibility of several plausible interpretations or insertions for these phenomena, nor their link to misunderstandings.

Finally, the most closely related work to ours is Anthonio and Roth (2021). In this work, we extracted a subset from wikiHowToImprove with sentences in which a word or phrase was inserted that referred to an entity in the preceding discourse. We used a pattern based approach to collect a set of instances where there was an implicit reference in the original sentence (e.g., *Call for an appointment*) that was made explicit in the revised version through insertion (e.g., *Call the salon for an appointment*). In our experiments, we masked the entity mention of the insertion and used Generative Pre-trained Transformer (Radford et al., 2018) to generate the top-100 completions. In a second step, we used the perplexity to re-rank the top-100 completions, which we found to increase the likelihood of finding the human-inserted reference among the top 10. Besides, we found in a set of annotation experiments that alternate fillers generated by the language model can be just as good as the human-produced insertion.

3. Task and Resource

The aim of this work is to accomplish several steps towards assessing potentially misunderstood instances of implicit and underspecified language and examining possible clarification alternatives for their plausibility. For this purpose, we first define the task of determining plausibility of possible clarifications based on insertions in revisions (§3.1). For this task, we create subsets of data representing selected phenomena, for which several different clarifications seem potentially possible (§3.2). The data creation is done semi-automatically in three steps: First, we extract relevant instances of each phenomenon in a rule-based manner based on automatic preprocessing (§3.3. Second, we use general-purpose language models to generate and select alternative clarifications that seem appropriate for the extracted instances (§3.4). Finally, in a third step, we use crowdsourcing to collect human plausibility ratings for all clarifications (§3.5).

3.1. Task Definition

Our task is to predict whether a clarification for an instruction is *plausible* or *implausible*. We additionally introduce a class label *neutral* for clarifications that are neither clearly plausible or implausible. Formally, we define each instruction in terms of its textual content $c = (s, d)$ and each clarification as a filler f. The (original) sentence s of an instruction is represented as a sequence of tokens $t_0 t_1 \dots b \dots t_n$, with a special blank token b = ___ to represent the position of the filler f. $d = (d_{\text{before}}, d_{\text{after}})$ is additional discourse context around s, consisting of token sequences that represent the previous sentences d_{before} and follow-up sentence d_{after}. $f = [x_0 x_1 \dots x_m]$ is a clarification, a sequence of one or more tokens that can replace b in s to obtain a new revised sentence version $s_f = [t_0 t_1 \dots f \dots t_n]$. The model must then select how plausible f is in the given context c, classifying it as *implausible*, *neutral* or *plausible*.

3.2. Phenomena

Revisions in collaboratively edited texts can have purposes other than clarification, such as spelling or grammar correction. Therefore, we specifically extract a subset of revisions that are made to clarify implicit or underspecified elements. We consider the following linguistic phenomena (see Table 2 for examples):

Implicit references ($N = 6,014$): instances with a non-verbalized reference in the original sentence which was clarified in the revised sentence through insertion. The considered reference refers to an entity in the previous discourse context. We re-use the set of implicit references from Roth and Anthonio (2021).

Fused heads ($N = 1,929$): instances of noun phrases for which the head noun was implicit in the original sentence, which was clarified in the revised sentence through insertion. The considered phrases may or may not refer to an entity mentioned in the previous discourse context and do not overlap with the set of implicit references.

Noun compounds ($N = 5,759$): instances of underspecified noun phrases, which were clarified in the revised sentence through the insertion of a dependent noun to form a more specific compound. The resulting compounds are mostly cases of endocentric compounds (Nakov, 2013): the head noun defines a set of entities H and the dependent noun attributes a particular property to the head, with the combination describing a more specific subset $C \subseteq H$ (e.g., "garden gloves" are a subset of "gloves").

Metonymy ($N = 1,855$): instances in which a revision adds a noun y to a noun x to make explicit to which component or aspect of x the text refers. The insertion follows the possessive pattern of "y of x" or the genitive "x'(s) y".

3.3. Extraction

We collect instances of the aforementioned phenomena by extracting revisions where a single *contiguous* insertion was made and where the insertion was the only difference between the original and revised version. We compute differences and extract relevant instances automatically using the Python library `difflib`[3] and the following preprocessing tools: `spaCy`[4] for sentence splitting and tokenisation, the Berkeley Neural

[3] `https://docs.python.org/3/library/difflib.html`

[4] `https://github.com/explosion/spaCy`

Phenomenon	Pattern	Clarification question	Example	Potential fillers
Implicit reference	Ø→ DET NOUN or NOUN	Who(m)? What?	Rinse ___ before re-assembling.	✓each piece ✓your hands
Fused head	DET/JJ Ø → DET/JJ NOUN	Who(m)? What?	Some ___ like tricks, some like races, and some like speed control.	✓cyclists ✗ races
Noun compound	NOUN → **NOUN** NOUN	What type of . . . ?	Line a large baking sheet with ___ foil.	✓aluminium ✗ bronze
Metonymy	NOUN → **NOUN of** NOUN	Which aspect or part of . . . ?	Turn the cup upside-down and tape it to ___ an aluminum pie pan.	✓the insides of ✓the bottom of ✗ the wood of
	NOUN → NOUN**'s/' NOUN**		The blanket should be snug around your baby ___, but not tight.	✓'s body ✓'s belly

Table 2: Phenomena of implicit and underspecified language in the data set.

Parser (Kitaev and Klein, 2018) for constituency parsing and `Stanza` (Qi et al., 2020) for POS tagging, dependency parsing and coreference resolution.

Implicit references. We select the cases from Anthonio and Roth (2021) with insertions containing a single noun or a determiner followed by a noun. In this set of the data, each insertion (co-)refers to an entity mentioned in the previous context.

Fused heads. We search for a fused head noun phrase with a determiner or adjective head in the original sentence and select those instances where a single noun was inserted in the revision. Note that there is a risk of false positives here because the original sentence might have been ungrammatical without the head noun (cf. Section 4). A fair portion of the fused heads in this subset also refer to an entity or concept in the previous context, as exemplified by the many cases of the fused head "this".

Noun compounds. We select instances of single noun insertions in which the inserted noun is a `compound` dependent of another noun that has already been present in the original sentence. Some inserted noun compounds refer to concepts that have been mentioned in previous context, but the majority is only related to the context by commonsense knowledge.

Metonymy. For the genitive "x'(s) y", we select insertions including an apostrophe and a noun y that is in a dependency relation `nmod:poss` with a noun x. For the "y of x" pattern, we select insertions that consist of a noun y and the token *of* and that was made right in front of a noun x, allowing for intervening determiners and adjectives. Most inserted nouns in this subset do not appear in the previous context and need to be inferred by commonsense knowledge.

Discourse context. Finally, we extract the discourse context d for each sentence s. To avoid having unreasonably long texts for annotation, we limit the context to the article title, the paragraph title and at most two preceding sentences as well as one follow-up sentence. Implementation details are described in Appendix A.

3.4. Generating Clarifications

We produce a set of possible clarifications for each instance as follows: First, we generate the top-100 fillers in place of an observed insertion using language modeling. Second, we select a subset of potentially suitable clarifications by filtering and clustering the top-100.

Filler generation. For the implicit references, we take the top-100 generated clarifications from Anthonio and Roth (2021). For the other phenomena, we generate alternative clarifications automatically using the same approach as Anthonio and Roth (2021). That is, we feed the original sentence s with the surrounding sentences from the same paragraph to a language model. We then compute the top-100 completions for the token position(s) where an insertion was added in the revised sentence.[5] We use BERT (Devlin et al., 2019) instead of GPT (Radford et al., 2018) to generate the clarifications, as the required insertions consist of only one token and BERT makes it possible to also consider follow-up context directly. The BERT checkpoint `bert-base-uncased` in `Transformers` (Wolf et al., 2020) was used without additional pre-training.

Filler selection. From the top-100 clarifications provided by the language model, we select four fillers with the goal of producing a semantically diverse set of clarifications. First, we remove unsuitable fillers from the

[5]For metonymy, only y is treated as the filler and the other elements (*of*, determiners, etc.) are assumed as given in order to concentrate on the metonymic aspect. For the examples in Table 2, this means that *the ___ of an aluminium pie pan* and *your baby's* ___ would be given and the fillers would be *insides*, *bottom* and *wood* or *body* and *belly*.

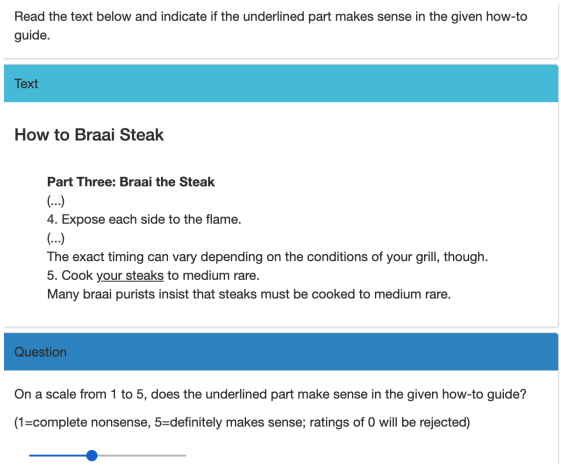

Figure 1: Interface for collecting annotations.

top-100, including cases that only consist of digits or non-alphanumerical characters and fillers that do not have the right part of speech based on `Stanza` (retaining only "NOUN" for fused heads and metonymy and "NN" for noun compounds to exclude plural nouns). For all instances with ≥ 4 candidate fillers, we select the observed insertion from the revised sentence as one filler. To select semantically different fillers as alternate candidates, we apply k-means clustering with $k = 4$ to the remaining candidates, using Elkan (2003)'s algorithm as implemented in `sklearn` (Pedregosa et al., 2011). We obtain vector representations for clustering from BERT (`bert-base-uncased`) by averaging over the last hidden state for all tokens in a filler. After clustering, we select the fillers closest to the four cluster centroids based on cosine similarity.

3.5. Plausibility Annotation

Task. After selecting fillers for each sentence, we collect plausibility judgements on Amazon Mechanical Turk for our train set (19,975 instances, i.e. 3995 sentences with 1 human and 4 generated fillers each[6]), development and test sets (2,500 instances each, i.e., 125 sentences per phenomenon with 5 fillers per sentence). Each clarification is annotated by two crowdworkers in the train set and four crowdworkers in the development and test set. In particular, we ask participants to indicate on a scale from 1 to 5 whether a highlighted clarification makes sense in the context of a given how-to-guide. An example and interface as shown in our Human Intelligence Task (HIT) is depicted in Figure 1.

Qualifications. We use several qualifications to increase the annotation quality. First, we require participants to be located in the United States or in the United Kingdom, to increase the chance that the participants are native speakers of English. Secondly, participants need to have a HIT approval rate $\geq 95\%$ and their number of approved HITS has to be ≥ 1000. Finally, annotators are required to pass a qualification test in which they are asked to judge a list of clearly plausible and implausible cases that were pre-selected unanimously by the authors.

[6] 1000 each for noun compounds and metonymy, 996 for implicit references and 999 for fused heads.

	Train	Dev	Test
Implausible	5,474 (27%)	982 (39%)	858 (34%)
Neutral	7,162 (36%)	602 (24%)	672 (27%)
Plausible	7,339 (37%)	916 (37%)	970 (39%)
Total	19,975	2,500	2,500

Table 3: Distribution of class labels in our training, development and test sets.

Class labels. For the task as described in Section 3.1, we average over the real-valued judgements collected for a clarification and map this plausibility score to one of the three classes labels. Specifically, we label clarifications with an average score ≤ 2.5 as *implausible*, clarifications with a score ≥ 4.0 as *plausible*, and all clarifications between these thresholds as *neutral*.

Statistics. We show the frequency distribution of the labels in the train, development and test set in Table 3. Since we are particularly interested in cases with multiple plausible clarifications, we also compute the average number of *plausible* clarifications per sentence s, which we found to be 1.84, 1.87 and 1.84 in the training, development and test set, respectively. This means that, on average, each annotated sentence in the dataset has between 1 and 2 clarifications that the annotators deem plausible.

4. Data Analysis

In this section, we investigate the extent to which the clarifications in CLAIRE reflect diverging interpretations of an implicit/underspecified element. Therefore, we analyze the sentences for which there are several plausible clarifications that are potentially conflicting to one another. We define *conflicting clarifications* as clarifications referring to different persons/objects/aspects considering the context.

4.1. Method

We select a subset of sentences with potentially conflicting clarifications from the development set in three steps. First, we take the sentences for which there are at least two plausible clarifications with an average score ≥ 4.5 ($N = 416$).[7] Next, we exclude the sentences and their clarifications for which (1) the sentence would be ungrammatical without the clarification (e.g., *guidance* in *seek professional ___ if you feel out of control*) and (2) the clarifications did not represent the phenomenon that we were interested in. These two issues are due to error propagation in components like part-of-speech

[7] We use 4.5 as a threshold because plausible clarifications with a lower score often contained minor issues (e.g., typos).

Phenomenon	Sentences	Plausible clarifications
Implicit references	55	137 (average: 2.5)
Fused heads	36	86 (average: 2.4)
Metonymic reference	35	83 (average: 2.4)
Noun compound	33	75 (average: 2.3)
Total	**159**	**381** (average: 2.4)

Table 4: Distribution of development set instances with multiple plausible clarifications across phenomena.

(POS) tagging or parsing during the selection procedure described in Section 3. The result is a collection of 159 sentences. We show the distribution of those sentences in Table 4. We find most cases with multiple plausible clarifications for sentences with implicit references ($N = 139$). An example of such a sentence is: *Stir slightly to combine* ___, for which the annotators marked *the ingredients* and *the mixture* as plausible.
Finally, two of the authors identify for the 159 sentences and their multiple plausible clarifications whether the clarifications are conflicting or equivalent to one another. The agreement is computed as the ratio between instances in which the annotators identified the same set of clarifications as being conflicting or equivalent to one another.

4.2. Results

Table 5 shows the agreement between both annotators. Clarifications for fused heads were the most difficult to annotate, because of the relationship between the preceding word (e.g., *this, most, some*) to the clarification in the given context. Among the 135 sentences with agreement, we found 116 sentences with (one or more) conflicting clarifications. We discuss different categories of conflicts and examples for each phenomenon.

Implicit references ($N = 43$): In most instances ($N = 21$), the conflicting clarifications denote objects that stand in a part–whole relationship. For example, *Refrigerate the oil for 1–2 weeks to infuse the oil/the ingredients*. The second largest set of conflicts ($N = 14$) involves separate entities, many of which are related in their domain (e.g., *Anger can only trouble your life/your heart/your soul*). Some further conflicts ($N = 4$) are related to clarifications that are not necessarily exclusive but that clarify different aspects of the instruction. For instance, inserting *protest/day* into the paragraph heading *How to Find and Hire a Charter Bus Company (for a ___)* would address two different clarification questions ("For what purpose?"/"For which time period?").

Fused heads ($N = 27$): A number of clarifications are conflicting because they denote sets that stand in a subgroup/group relationship ($N = 12$), such as: *Most people/teenagers hate scary movies*. Other conflicts ($N = 12$) involve clarifications that are not directly related in meaning, but are within the same domain, such as: *For this project/purpose/process, you will need* .... Another two conflicts involve clarifications that answer different types of clarification questions, such as: "What?"/"Who?" for *Most medications/people take four to eight weeks to show any effects* ....

Phenomena	Agreement by sentence
Implicit references	46/55 (83.66%)
Fused heads	30/48 (62.50%)
Metonymic reference	29/35 (82.86%)
Noun compound	30/33 (90.91%)
Total	**135**

Table 5: Absolute and relative number of sentences for which both annotators agreed on the set of clarifications that are (non-)conflicting.

Metonymy ($N = 21$): In most instances ($N = 16$), the conflicting clarifications refer to different aspects of an entity, such as *the absorption/amount of sunlight*. In a small number of cases ($N = 5$), the conflicting clarifications refer to different parts of an entity, such as: *the hairs/skins of pelts* or *your baby's belly/skin/body*.

Noun compounds ($N = 26$): Most conflicting clarifications ($N = 12$) address different clarification questions, such as: *road/racing bike* ("Where?/For what?"), *summer/dance class* ("When?/Which type?"). Most of the remaining conflicts ($N = 7$) address the same question, such as *guitar/audio amps* ("What is amplified?").

Our analysis shows that all of the four considered phenomena involve instances that can have conflicting clarifications in context. In total, we identified 117 such instances in our development set (23% of the whole development set). Some similarities exist across phenomena: the conflicting clarifications can be related in their domain or denote differences in granularity/specificity. Some categories of conflict are however specific to each phenomenon, such as references to different parts of an entity in case of metonomy.

5. Computational Experiments

In this section, we investigate how we can computationally distinguish between plausible, neutral and implausible clarifications. We approach the task described in 3.1 as a supervised three-class classification problem. For training, hyperparameter tuning and evaluation of our models, we use the training, development and test sets, respectively. As evaluation measure, we calculate accuracy as the ratio of correct predictions among all predictions of a model.

5.1. Models and Hyperparameters

We compare different ways of modeling the relation between a filler f and its context in the sentence s and the surrounding discourse d. A concatenation of the clarified sentence s_f and its discourse context d serves as the input to our models: $d_{\text{before}} :: s_f :: d_{\text{after}}$ (cf. 3.1).

Model	Dev	Test
NAIVE BAYES	36.20%	38.20%
BERT VANILLA	44.51%	45.66%
$-d$	43.07%	–
$-d\ -s_f$	42.36%	–
BERT + RANKING	48.53%	39.80%
BERT + FILLER MARKERS	51.39%	47.37%

Table 6: Accuracy of different model architectures on the development and test set

Baselines. Our first goal is to see if a model can learn to classify the plausibility of f by just seeing the filler in context. Therefore, we work with two baseline models that process s_f without treating the filler span f any different than the other tokens. The first baseline is a multinominal NAIVE BAYES classifier with tf-idf weighted unigram features that takes $d_{\text{before}} :: s_f :: d_{\text{after}}$ as input and predicts a plausibility label as output. We use `sklearn` (Pedregosa et al., 2011) for the implementation. For the second baseline, BERT VANILLA, we fine-tune a BERT model that takes $d_{\text{before}} :: s_f :: d_{\text{after}}$ as input and whose last hidden state of the first sequence token (`[CLS]`) is then passed into a linear classification layer. This model is based on the BERT checkpoint `bert-base-uncased` from `Transformers` (Wolf et al., 2020).

Extensions. Two aspects neglected in the BERT VANILLA model are that only specific tokens are part of the clarification to be classified and that other competing clarifications exist for each classification instance. Therefore, we hypothesize that the following two extensions can provide a better training signal to the model: In our first extension, BERT + FILLER MARKERS, we explicitly mark the span of a filler f with special tokens "[F]" and "[/F]". This technique is adapted from work on relation extraction, where the spans of related (subject and object) entities are highlighted by special tokens (Soares et al., 2019). Our second extension, BERT + RANKING, takes into account that different fillers are provided for each clarified sentence, which can be ranked in terms of language modeling perplexities. For a sentence s and its fillers $F = [f_1, f_2, f_3, f_4, f_5]$, we compute the perplexity p_i of each $f_i \in F$ within s using GPT (Radford et al., 2018) and use the numerical rank $r_i \in [1, 2, 3, 4, 5]$ as an additional feature in the linear classification layer.[8]

Hyperparameters. We utilize the development set for choosing model architectures and hyperparameters. We conduct preliminary experiments with different learning rates, dropout values and optimizers. For BERT VANILLA and BERT + FILLER MARKERS, we train the models for 10 epochs with the Adam optimizer (Kingma and Ba, 2017), a learning rate of e^{-4} and a dropout rate of 50%. We freeze the parameters of the first 11 BERT layers and only fine-tune the parameters of the last layer. For BERT + RANKING, we use a dropout rate of 25% and we freeze all the BERT layers without fine-tuning the parameters.

[8] We also tried to combine RANKING and FILLER MARKERS, but we did not observe any improvements from this combination in preliminary experiments.

Actual / Predicted	Implausible	Neutral	Plausible
Implausible	**406**	321	255
Neutral	128	**187**	287
Plausible	84	140	**692**

Table 7: Confusion matrix for BERT + FILLER MARKERS on the development set.

5.2. Results and Discussion

The results, shown in Table 6, indicate that the NAIVE BAYES baseline obtains the lowest accuracy scores, namely 36.20% and 38.20% on the development and test set, respectively. In comparison, BERT VANILLA achieves an improvement by 8.31 and 7.46 percentage points. BERT + RANKING further outperforms NAIVE BAYES by 12.33 and percentage points on the development and by 1.60 percentage points on the test set. The best model, BERT + FILLER MARKERS, improves on NAIVE BAYES by 15.19 and 9.17 percentage points.

Discussion. We analyze the predictions of our best model BERT + FILLER MARKERS on the development set. Table 7 shows its confusion matrix. The model most often predicts instances to be plausible, with 1,234 predictions vs. 648 predictions as neutral and 618 as implausible. In contrast, BERT + RANKING classifies only 19 out of 2,500 instances as neutral. While BERT + RANKING seems to mainly learn the two extremes of the plausibility scale, BERT + FILLER MARKERS manages to cover the full spectrum. For BERT + FILLER MARKERS, the class-wise accuracy is much higher for plausible (75.55%) than for implausible (41.34%) and for neutral (31.06%). Thus, a clear semantic match between filler and context seems easier to detect than a mismatch or a borderline case.

The accuracy that BERT + FILLER MARKERS achieves per phenomenon is lowest on noun compounds (46.88%) and highest on metonymy (57.92%), with implicit references (48.48%) and fused heads (52.32%) in between. Predicting the plausibility of an inserted compound might be difficult because the semantic match does not only depend on the global context, but mainly on the complex relation between head noun and compound.

To find out if the models can differentiate between the 5 different fillers for a given context c, we count for how many of the 500 development sentences the models predict the same class for all 5 fillers (e.g., 5 times plausible). This is the case in 499 out of 500 sentences for NAIVE BAYES and in 208 out of 500 sentences for BERT VANILLA. In comparison, BERT +

	Example	Prediction	Label
1	**How to Clean a Vacuum** - *Performing a Basic Cleaning* (...) It is particularly important that you do not reinsert components of the canister, like the filter, until they have dried. Putting moist objects in a confined space can foster disease. Leave no traces out in the sun, if possible. (Human insertion: the vacuum)	plausible	implausible
2	**How to Draw a Human Figure from the Side** - *Steps* Use a pencil to draw the corners of a half circle. (Human insertion: top) This circle should be higher on the page as it will later become the top of the shoulder.	plausible	implausible
3	**How to Tap a Tree for Maple Syrup** - *Steps* 1. Find a cedar tree. (Human insertion: maple) In the summer, search around the neighbourhood if you're property doesn't have a maple tree.	plausible	implausible
4	**How to Calculate Dog Years** - *Using Your Dog's Physical Traits* 1. Look at the thickness of the teeth. (Human insertion: condition) If you're unsure of your dog's age, or want to determine if they are already entering into the senior territory, try the teeth.	implausible	plausible

Table 8: Examples of incorrect predictions of BERT + FILLER MARKERS on the development set.

FILLER MARKERS predicts five identical class labels for only 23 sentences. The baselines seem to be insensitive to the small change that the insertion of different fillers makes, whereas the emphasis on the filler span in BERT + FILLER MARKERS seems to increase this sensitivity substantially.

Ablation. We consider two ablations on BERT VANILLA to analyse which information is essential for predicting the semantic plausibility of a clarification (see Table 6). First, we remove the discourse context d, only providing the sentence s_f. The accuracy on the development set decreases by 1.44 percentage points, indicating that cross-sentence context is useful. Second, we remove both d and the sentence s_f, only providing the filler f. This is inspired by Poliak et al. (2018; Gururangan et al. (2018), who showed that models can obtain a high accuracy on some NLI datasets solely based on the hypothesis without considering the premise. We observe that performance drops only by 2.15 percentage points. This suggests that the model can reconstruct the semantic plausibility of a filler in a given context to a degree by learning how common or universally applicable f is in and of itself.

Error analysis. Table 8 shows examples of errors that BERT + FILLER MARKERS makes on the development set. We focus on the more serious errors that confuse implausible and plausible. The model often predicts incongruous fillers as plausible (255 out of 2,500 instances). One potential reason is that the filler fits in well with neighbouring words although it does not match the topic of the larger discourse context. In Example 1, *Leave no traces* forms a commonly used phrase, but does not match the context of drying a cleaned vacuum out in the sun. On the other hand, a filler like *corners* in Example 2 might be classified as plausible if it is related to the text domain (drawing shapes) despite not making sense in the specific sentence (*corners of a half circle*). Another frequent problem is that commonsense knowledge would be needed to correctly assess the plausibility of a filler, e.g. for identifying that maple syrup cannot be obtained from a *cedar tree* or that the age of a dog can in fact be estimated based on the *thickness of the teeth*.

6. Conclusion

In this paper, we proposed the task of classifying a clarification for an instruction as plausible, neutral or implausible. To address this task, we presented CLAIRE, a data set that contains several alternative clarifications for wikiHow instructions in English In an analysis, we found that the clarifications in our data can represent diverging interpretations of an implicit/underspecified element in instructional texts. We found conflicting clarifications in all phenomena and that the extent to which they are conflicting can differ. We further showed that the clarifications share similarities across phenomena, such as that they denote differences in specificity/granularity. However, some clarification types were only specific to one phenomenon.

Among the different model architectures that we tried on the plausibility classification task, the use of special tokens to highlight the clarification within its sentence context proved most successful. The resulting model was able to learn the full plausibility spectrum and to differentiate between different versions of a sentence. Nevertheless, there is still a lot of room for improvement, especially in correctly modeling more complex relations between a clarification and its larger context that depend on commonsense knowledge. Based on our results, we believe that fully reconstructing these relations is still beyond the capability of current general-purpose language models.

7. Acknowledgements

The research presented in this paper was funded by the DFG Emmy Noether programme (RO 4848/2-1)

8. Bibliographical References

Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., and Burtsev, M. (2021). Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Anthonio, T. and Roth, M. (2020). What can we learn from noun substitutions in revision histories? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1359–1370, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Anthonio, T. and Roth, M. (2021). Resolving implicit references in instructional texts. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 58–71, Punta Cana, Dominican Republic and Online, November. Association for Computational Linguistics.

Anthonio, T., Bhat, I., and Roth, M. (2020). wikihowtoimprove: A resource and analyses on edits in instructional texts. In Nicoletta Calzolari, et al., editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5721–5729. European Language Resources Association.

Bi, K., Ai, Q., and Croft, W. B. (2021). Asking clarifying questions based on negative feedback in conversational search. In Faegheh Hasibi, et al., editors, *ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*, pages 157–166. ACM.

Chersoni, E., Lenci, A., and Blache, P. (2017). Logical metonymy in a distributional model of sentence comprehension. In Nancy Ide, et al., editors, *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics, *SEM @ACM 2017, Vancouver, Canada, August 3-4, 2017*, pages 168–177. Association for Computational Linguistics.

Debnath, A. and Roth, M. (2021). A computational analysis of vagueness in revisions of instructional texts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 30–35, Online, April. Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, et al., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Dhar, P. and van der Plas, L. (2019). Learning to predict novel noun-noun compounds. In Agata Savary, et al., editors, *Proceedings of the Joint Workshop on Multiword Expressions and WordNet, MWE-WN@ACL 2019, Florence, Italy, August 2, 2019*, pages 30–39. Association for Computational Linguistics.

Elazar, Y. and Goldberg, Y. (2019). Where's my head? definition, dataset and models for numeric fused-heads identification and resolution. *Trans. Assoc. Comput. Linguistics*, 7:519–535.

Elkan, C. (2003). Using the triangle inequality to accelerate k-means. In *Proceedings of the 20th international conference on Machine Learning (ICML-03)*, pages 147–153.

Fillmore, C. J. (1986). Pragmatically controlled zero anaphora. In *Annual Meeting of the Berkeley Linguistics Society*, volume 12, pages 95–107.

Günther, F. and Marelli, M. (2016). Understanding karma police: The perceived plausibility of noun compounds as predicted by distributional models of semantic representation. *PloS one*, 11(10):e0163200.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June. Association for Computational Linguistics.

Khalid, B., Alikhani, M., and Stone, M. (2020). Combining cognitive modeling and reinforcement learning for clarification in dialogue. In Donia Scott, et al., editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4417–4428. International Committee on Computational Linguistics.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Kitaev, N. and Klein, D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia, July. Association for Computational Linguistics.

Lapata, M. and Lascarides, A. (2003). A probabilistic account of logical metonymy. *Comput. Linguistics*, 29(2):261–315.

Macagno, F. (2017). Evidence and presumptions for

analyzing and detecting misunderstandings. *Pragmatics & Cognition*, 24(2):263–296.

Majumder, B. P., Rao, S., Galley, M., and McAuley, J. J. (2021). Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In Kristina Toutanova, et al., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4300–4312. Association for Computational Linguistics.

Marge, M. and Rudnicky, A. I. (2019). Miscommunication detection and recovery in situated human-robot dialogue. *ACM Trans. Interact. Intell. Syst.*, 9(1):3:1–3:40.

McRoy, S. W. and Hirst, G. (1993). Abductive explanation of dialogue misunderstandings. In *EACL*.

Min, S., Michael, J., Hajishirzi, H., and Zettlemoyer, L. (2020). Ambigqa: Answering ambiguous open-domain questions. In Bonnie Webber, et al., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5783–5797. Association for Computational Linguistics.

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, June. Association for Computational Linguistics.

Nakov, P. (2013). On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19(3):291–330.

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, N. Q., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany, August. Association for Computational Linguistics.

Pedinotti, P. and Lenci, A. (2020). Don't invite BERT to drink a bottle: Modeling the interpretation of metonymies using BERT and distributional representations. In Donia Scott, et al., editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6831–6837. International Committee on Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018). Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, June. Association for Computational Linguistics.

Purver, M., Hough, J., and Howes, C. (2018). Computational models of miscommunication phenomena. *Top. Cogn. Sci.*, 10(2):425–451.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In Asli Celikyilmaz et al., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 101–108. Association for Computational Linguistics.

Radford, A., Narasimhan, K., salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI.

Rambelli, G., Chersoni, E., Lenci, A., Blache, P., and Huang, C.-R. (2020). Comparing probabilistic, distributional and transformer-based models on logical metonymy interpretation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 224–234, Suzhou, China, December. Association for Computational Linguistics.

Rao, S. and Daumé, III, H. (2018). Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In Iryna Gurevych et al., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2737–2746. Association for Computational Linguistics.

Roth, M. and Anthonio, T. (2021). UnImplicit shared task report: Detecting clarification requirements in instructional text. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 28–32, Online, August. Association for Computational Linguistics.

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Sekulic, I., Aliannejadi, M., and Crestani, F. (2021). Towards facet-driven generation of clarifying questions for conversational search. In Faegheh Hasibi, et al., editors, *ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*,

pages 167–175. ACM.

Shutova, E. (2009). Sense-based interpretation of logical metonymy using a statistical method. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Student Research Workshop*, pages 1–9. The Association for Computer Linguistics.

Soares, L. B., FitzGerald, N., Ling, J., and Kwiatkowski, T. (2019). Matching the blanks: Distributional similarity for relation learning. In Anna Korhonen, et al., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2895–2905. Association for Computational Linguistics.

Utiyama, M., Murata, M., and Isahara, H. (2000). A statistical approach to the processing of metonymy. In *COLING 2000, 18th International Conference on Computational Linguistics, Proceedings of the Conference, 2 Volumes, July 31 - August 4, 2000, Universität des Saarlandes, Saarbrücken, Germany*, pages 885–891. Morgan Kaufmann.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Xu, J., Wang, Y., Tang, D., Duan, N., Yang, P., Zeng, Q., Zhou, M., and Sun, X. (2019). Asking clarification questions in knowledge-based question answering. In Kentaro Inui, et al., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1618–1629. Association for Computational Linguistics.

Yang, H., Roeck, A. N. D., Gervasi, V., Willis, A., and Nuseibeh, B. (2010a). Extending nocuous ambiguity analysis for anaphora in natural language requirements. In *RE 2010, 18th IEEE International Requirements Engineering Conference, Sydney, New South Wales, Australia, September 27 - October 1, 2010*, pages 25–34. IEEE Computer Society.

Yang, H., Roeck, A. N. D., Willis, A., and Nuseibeh, B. (2010b). A methodology for automatic identification of nocuous ambiguity. In Chu-Ren Huang et al., editors, *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 1218–1226. Tsinghua University Press.

Zarcone, A., Utt, J., and Padó, S. (2012). Modeling covert event retrieval in logical metonymy: probabilistic and distributional accounts. In David Reitter et al., editors, *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics, CMCL@NAACL-HLT 2012, Montrèal, Canada, June 7, 2012*, pages 70–79. Association for Computational Linguistics.

Zhang, Z. and Zhu, K. Q. (2021). Diverse and specific clarification question generation with keywords. In Jure Leskovec, et al., editors, *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3501–3511. ACM / IW3C2.

9. Language Resource References

Talita Anthonio and Irshad Bhat and Michael Roth. (2020). *wikiHowToImprove: A Resource and Analyses on Edits in Instructional Texts*. European Language Resources Association.

Rowan Zellers and Yonatan Bisk and Roy Schwartz and Yejin Choi. (2018). *SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference*. Association for Computational Linguistics.

A. Appendix

Context extraction The basis for the context extraction are the wikiHowToImprove data files[9]. In these files, several sentences that belong together in an instruction step or another coherent text component are on the same line. Therefore, we also use these line breaks here for choosing the context sentences.

For the previous context, if there are two or more sentences before the target sentence on the same line, we take the first one and the one immediately before the target sentence. If there is only one previous sentence on the same line, we take this one and the first sentence from the previous line. Otherwise, we skip back to the previous line and take the first and the last sentence from there. If the line is part of an enumeration (lines starting with ordinal numbers like "1.", "2." etc.), we skip back to the last previous line in the ordered list (e. g. from "10." to "9."), leaving out minor notes and additions in lines with bullet points (starting with "*", "-", etc.).

For the follow-up context, if there are sentences after the target sentence on the same line, we take the first one. Otherwise, we select the first sentence of the following line.

In this whole process, we use `Stanza` (Qi et al., 2020) for sentence splitting. Left out context is replaced by "(. . .)" to mark the omission.

[9] `https://github.com/irshadbhat/wikiHowToImprove`.