UnImplicit Shared Task Report:
Detecting Clarification Requirements in Instructional Text

Michael Roth **Talita Rani Anthonio**
University of Stuttgart
Institute for Natural Language Processing
{rothml,anthonta}@ims.uni-stuttgart.de

Abstract

This paper describes the data, task setup, and results of the shared task at the First Workshop on Understanding Implicit and Underspecified Language (UnImplicit). The task requires computational models to predict whether a sentence contains aspects of meaning that are contextually unspecified and thus require clarification. Two teams participated and the best scoring system achieved an accuracy of 68%.

1 Introduction

The goal of this shared task is to evaluate the ability of NLP systems to detect whether a sentence from an instructional text requires clarification. Such clarifications can be critical to ensure that instructions are clear enough to be followed and the desired goal can be reached. We set up this task as a binary classification task, in which systems have to predict whether a given sentence in context requires clarification. Our data is based on texts for which revision histories exist, making it possible to identify (a) sentences that received edits which made the sentence more precise, and (b) sentences that remained unchanged over multiple text revisions.

The task of predicting revision requirements in instructional texts was originally proposed by Bhat et al. (2020), who attempted to predict whether a given sentence will be edited according to an article's revision history. The shared task follows this setup, with two critical differences: First, we apply a set of rules to identify a subset of edits that provide clarifying information. This makes it possible to focus mainly on those edits that are related to implicit and underspecified language, excluding grammar corrections and other edit types. Since the need for such edits may depend on discourse context, a second difference is that we provide context for each sentence to be classified (see Table 1).

	Store Asparagus
✗	Keep the asparagus refrigerated for five to seven. [Cooked asparagus is best within a few days.]
✓	[Transfer the asparagus to a container.] Label the container with the date.

Table 1: Examples of a sentence that requires clarification according to the revision history (✗) and a sentence that remained unedited over many article-level revisions (✓). Annotators and systems were provided with additional context, here shortened in brackets.

2 Task and Data

In our task, sentences from instructional texts are provided in their original context and systems need to predict whether the sentence requires clarification. We define a clarification as a type of revision in which information is added or further specified.

Systems participating in the shared task are required to distinguish between sentences that require clarification and sentences that do not. For simplicity, we assume all sentences that remained unchanged over multiple article-level revisions (until the final available version) to not require clarification. Based on this assumption, we create a class-balanced data set for our task by selecting for each sentence that requires clarification exactly one sentence that does not require clarification.

In the following, we provide details on the collection procedure and an annotation-based verification thereof as well as statistics of the final data set.

2.1 Data Collection

We extract instances of clarifications from a resource of revision edits called wikiHowToImprove (Anthonio et al., 2020). Specifically, we used a state-of-the-art a constituency parser (Mrini et al., 2020) to preprocess all revisions from wikiHow-

Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language (UnImplicit 2021), pages 28–32
Bangkok, Thailand (online), August 5, 2021. ©2021 Association for Computational Linguistics

Edit type	Description	Example
Modifiers	Insertion of an adverbial/adjectival modifier	✗ Try watching one game to see if you like it. (→ Try watching one game alone to see if you like it.)
		✓ Learn about some teams. Article: **Enjoy Football**
Pronouns	Replacement of a pronoun with a noun phrase	✗ Do not be ashamed of it with your parents. (→ Do not be ashamed of your choice with your parents.)
		✓ Stay true to what you want. Article: **Explain Cross Dressing to Parents**
Complements	Insertion of an optional verb complement	✗ Press and hold to take a photo. (→ Press and hold the button to take a photo.)
		✓ Keep on pressing to extend the Snap to up to 30s. Article: **Set Up Snapchat Spectacles**
Quantifier/ Modals	Insertion of a quantifier or modal verb	✗ Dry the shoe off with the hand towel. (→ Dry each shoe off with the hand towel.)
		✓ Avoid using too much water. Article: **Make Your Sneakers Look New Again**
Verbs	Replacement of 'do' with another main verb	✗ The change in temperature does the rest. (→ The change in temperature takes care of the rest.)
		✓ You should do this as soon as you are finished. Article: **Cut a Glass Bottle**

Table 2: Revision types and example sentences that require clarification from our training set (✗). Additionally shown are clarified versions (→ . . .) and sentences that remain unrevised until the final version of an article (✓).

ToImprove and applied a set of rule-based filters to identify specific types of edits (see Table 2).

Sentences that require clarification identified this way are likely to share specific syntactic properties. Accordingly, it might be easy for a computational model to distinguish them from sentences that do not require clarification. We counteract this potential issue by relying on syntactic similarity to pair each sentence that requires clarification with a sentence that does not. Following Bhat et al. (2020), we specifically select sentences that are part of the final version of an article (according to wikiHowToImprove) and that remained unchanged over the past 75% of revisions on the article level. For the syntactic similarity measure, we calculate the inverse of the relative edit distance in terms of part-of-speech tags between two sentences.

Data and data format. We divide the collected data into training, development and test sets, following the splits by article of wikiHowToImprove. For all parts of the data, we provide the article name and the full paragraph in addition to the sentence to be classified. For the sentences that require clarification in the training set, we additionally provide the type of revision and the revised sentence.

Out-of-domain data. We collect a small set of data from other sources, following the procedure outlined above, to create a possibility of testing how well models would generalize beyond the type of instructions provided in wikiHow articles. For this purpose, we create a corpus of board game manuals that consists of modern games for which multiple print-runs and editions of manuals exist.[1] We apply the same preprocessing and filtering criteria to this corpus as described above. In order to increase the size of this data, we allow edits that go beyond the exact match of a syntactic pattern (e.g. we include ✗ *The price. . .* → *This unit price. . .*, which contains a small change in addition to the added modifier).

[1] Board games in this set include *Android: Netrunner*, *Brass: Lancashire*, *Champions of Midgard*, *Descent: Journeys into the Dark (2nd Ed.)*, *Feast for Odin*, *Food Chain Magnate*, *Gloomhaven*, *Istanbul*, *Le Havre*, *Root*, *Teotihuacan: City of Gods*, *T.I.M.E. Stories*, *Unfair* and *War of the Ring (2nd Ed.)*.

	#Sentences	#Tokens	#Types
wikiHowToImprove			
- Training	39 186	552 567	25 297
- Development	3 264	45 622	6 719
- Test	3 414	48 261	6 934
Board game manuals			
- Test	44	885	381
Total	45 908	647 335	27 331

Table 3: Statistics on sentence and word counts.

2.2 Annotation and Statistics

Previous work has found that revisions do not always improve a sentence (Anthonio and Roth, 2020). Based on this insight, we decided to collect human judgements on all edited sentences that would be included as requiring revision in our development, test, and out-of-domain data. We used Amazon Mechanical Turk to collect 5 judgements per edit and only kept sentences that require clarification if a majority of annotators judged the revised version as being better than the original version.

Statistics. Our rule-based extraction approach yielded a total of 24,553 sentences that received clarification edits. We discarded 1,599 of these sentences as part of the annotation process. In these cases, annotators found the edits to be unhelpful or they had disagreements about the need for clarification. Finally, we paired the remaining 22,954 sentences with sentences that received no clarification. Statistics for the training, development, test and out-of-domain sentences as well as for the full data set are provided in Table 3.

3 Participants and Results

Two teams registered for the shared task and submitted predictions of their systems: Wiriyathammabhum (2021) and Ruby et al. (2021). **Wiriyathammabhum** approached the task as a text classification problem and experimented with different training regimes of transformer-based models (Vaswani et al., 2017). **Ruby et al.** combined a transformer-based model with additional features based on entity mentions, specifically addressing clarifications of pronoun references.

Results. We evaluated submitted predictions on the test and out-of-domain data in terms of accuracy, measured as the ratio of correct predictions over all data instances. We compare submitted predictions against the expected performance of a random baseline and against a simple logistic regression classifier that makes use of uni-grams, bi-grams and sentence length as features. The results, summarized in Table 4, show that the participating systems perform substantially better than both baselines on the test set.[2] Compared to this high performance (66.4–68.8%), results on the out-of-domain data are considerably low (59.1%) and they do not exceed the accuracy of the logistic regression classifier (61.4%). We next discuss potential reasons for this and highlight other observations.

	wikiHowToImprove	Games	Overall
Wiriyathammabhum	**68.8**	59.1	**68.4**
Ruby et al. (updated)	66.4	59.1	66.3
Logistic Regression	62.4	**61.4**	62.3
Ruby et al. (official)	50.1	56.8	50.2
Random	50.0	50.0	50.0

Table 4: Accuracy (%) of baselines and participants.

	Wiriyathammabhum	**Ruby et al.**	LR
Modifiers	53.6	46.7	**53.7**
Pronouns	**92.7**	92.2	73.4
Complements	**81.7**	68.7	59.2
Quantifier/modals	54.2	**55.4**	53.0
Verbs	**95.1**	70.7	78.0

Table 5: Test accuracy (%) by edit type.

4 Discussion

The results of the participating teams and the logistic regression baseline provide some insights regarding the task posed and the data sets provided.

Task. The results suggest that it is generally possible to predict whether a sentence requires clarification and models can pick up reliable patterns for most types of revision. In fact, the per-type results shown in Table 5 indicate that the best participating system is able to identify over 90% of cases that require one of the following two types of clarifications: replacements of pronouns and replacements of occurrences of ‘do’ as a main verb. These two types may seem like easy targets because pronouns and relevant word forms can be

[2] Note that due to a software bug during the evaluation phase, we allowed team **Ruby et al.** to submit an *updated* set of predictions after their *official* submission.

found simply by matching strings. However, the results of the logistic regression model show that a simple word-based classification is insufficient. Not all occurrences of pronouns and 'do' require clarification (cf. Table 2).

On the other end, we find that required insertions of modifiers, quantifiers and modal verbs are hard to predict. In fact, the systems only identify up to 56% of such cases, which is only slightly better than the performance of a random baseline (50%). One reason could be that commonsense knowledge plays an important role in such clarifications.

Data. It is worth noting that the distribution of different revision types is not balanced and the overall results are skewed accordingly. In almost half of the test sentences that require clarification, the edit involved the insertion of an adverbial or adjectival modifier (49%, 840 out of 1,707). Predicting the need for such edits is particularly difficult because they often add only subtle and context-specific information. Replacements of pronouns form the second most-frequent clarification type in our data (23%, 398/1707). Both participating systems were able to identify over 92% of sentences that require such a replacement. The remaining cases are distributed as follows: insertions of optional verb complements (15%, 262/1707), insertions of quantifiers and modal verbs (10%, 166/1707) and replacements of 'do' as a main verb (2%, 41/1707).

One potential reason for the differences in results between the test data and the out-of-domain data is that revision types are distributed differently as well. In fact, the edits of sentences that require clarification in the out-of-domain data almost always involve the insertion of an adverbial/adjectival modifier or an optional complement (82%, 18/22).

Insights from Participants. In addition to our observations, the system descriptions also report a number of interesting findings. For instance, **Ruby et al.** found that pronouns requiring replacement are often denoting a generic referent or a type of individual, rather than a specific entity. Based on this observation, they perform several experiments in which they first identify pronouns that should potentially be revised and then they combine representations of the identified pronouns with a sentence-level system to generate predictions.

A more technically motivated approach is taken by **Wiriyathammabhum**, who build on the observation that the distribution of sentence labels (requiring revision or not) is generally unbalanced and that revised versions of sentences that required clarification may be viewed as instances of sentences that do not require further clarification.

Both participants discuss interesting approaches to the shared task and show interim results on the training/development sets. For details, we refer the interested reader to the system description papers (Wiriyathammabhum, 2021; Ruby et al., 2021).

5 Conclusions

Two teams participated in our shared task on predicting the need for clarifications, with the top performing system achieving an accuracy of 68.4%. Perhaps unsurprisingly, the main takeaway from both systems is that transformer-based models pose a strong baseline for future work.

Linguistic insights. An analysis of the different types of needed clarifications showed that certain revision requirements are more difficult to predict than others. For example, we found edits that introduce potentially subtle and context-specific shades of meaning much more difficult to predict than cases where generic pronouns are resolved. Nonetheless, we find that the best system is able to predict the need for clarification across all types with an accuracy higher than expected by chance. We take this as a promising result and as motivation for future work on this task.

Open questions. A number of unanswered questions remain: for example, we have not investigated what is a realistic upper bound for the discussed task. We did find that annotators are generally able to identify which of two versions of a sentence is revised/better and they generally achieve high agreement. However, it still remains unclear under which conditions a revision is seen as mandatory. It also remains unclear to what extent the selected revision types actually reflect general clarification needs in a representative way.

In a preliminary study, we originally assumed that revisions of board game manuals could provide us with useful information about when clarifications are necessary. However, we found the application of syntactic rules for finding such revisions to be of limited use. Our annotation further showed that people also have difficulty distinguishing old game instructions from revised ones. It is quite likely that some texts are simply too specific for annotators (and computational models) as they require too much specialized knowledge.

Lessons learned. From our results, we draw the following conclusions for future tasks: a focus on instructions on everyday situations as described in wikiHow is generally desirable to enable a distinction between clarification needs due to implicit and underspecified language on the one hand and clarification needs due to lack of familiarity or specialized knowledge on the other hand. To better understand different needs for clarification, it will also be necessary to consider additional types of revisions in the future. Lastly, more context should be considered, both on the methods side as well as with regard to the data itself, in order to be able to better identify subtle clarification requirements.

We are already implementing some of these lessons in a follow-up task that will take place as part of SemEval-2022. In that task, the focus will be on sentences that require clarification and systems will need to predict which of multiple possible changes represent plausible clarifications.

Acknowledgments

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

References

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France. European Language Resources Association.

Talita Anthonio and Michael Roth. 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). International Committee on Computational Linguistics.

Irshad Bhat, Talita Anthonio, and Michael Roth. 2020. Towards modeling revision requirements in wikiHow instructions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8407–8414, Online. Association for Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking self-attention: Towards interpretability in neural parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 731–742, Online. Association for Computational Linguistics.

Ahmed Ruby, Christian Hardmeier, and Sara Stymne. 2021. A mention-based system for revision requirements detection. In *Proceedings of the First Workshop on Understanding Implicit and Underspecified Language*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Peratham Wiriyathammabhum. 2021. TTCB System description to a shared task on implicit and underspecified language 2021. In *Proceedings of the First Workshop on Understanding Implicit and Underspecified Language*.