What Can We Learn from Noun Substitutions in Revision Histories?

Talita Rani Anthonio **Michael Roth**

Institute for Natural Language Processing
University of Stuttgart
{anthonta,rothml}@ims.uni-stuttgart.de

Abstract

In community-edited resources such as wikiHow, sentences are subject to revisions on a daily basis. Recent work has shown that resulting improvements over time can be modelled computationally, assuming that each revision contributes to the improvement. We take a closer look at a subset of such revisions, for which we attempt to improve a computational model and validate in how far the assumption that 'revised means better' actually holds. The subset of revisions considered here are noun substitutions, which often involve interesting semantic relations, including synonymy, antonymy and hypernymy. Despite the high semantic relatedness, we find that a supervised classifier can distinguish the revised version of a sentence from an original version with an accuracy close to 70%, when taking context into account. In a human annotation study, we observe that annotators identify the revised sentence as the 'better version' with similar performance. Our analysis reveals a fair agreement among annotators when a revision improves fluency. In contrast, noun substitutions that involve common lexical-semantic relationships are often perceived as being equally good or tend to cause disagreements. While these findings are also reflected in classification scores, a comparison of results shows that our model fails in cases where humans can resort to factual knowledge or intuitions about the required level of specificity.

1 Introduction

Instructional texts often go through multiple revisions. For their final form, the author has to ensure that it is clear in which sequence described actions need to be performed and in what manner they have to be executed. A collection of revisions of instructional texts can be found, for instance, on the community-edited platform wikiHow. In previous work using wikiHow revision histories, we showed that improvements in texts can be modeled computationally under the assumption that revised sentences are better than their predecessors (Anthonio et al., 2020). This assumption seems intuitive for edits that correct or clarify parts of a sentence (e.g., in terms of grammar/factual mistakes). In the analysis, however, we also found a number of edits that involve paraphrases or more subtle differences in terms of specificity/genericity. In these cases, it remains unclear in how far edits actually represent improvements.

This problem notably exists also in other work on textual revisions which aim to develop a computational model, including studies on Wikipedia (Bronner and Monz, 2012; Daxenberger and Gurevych, 2012; Faruqui et al., 2018) and ArgRewrite (Zhang and Litman, 2015). In particular, these studies neglect two important aspects: First, they merely categorize edits by how they change a sentence (e.g., information deletion/addition, link creation/deletion), without specifying if or in how far the change represents an actual improvement. Secondly, they provide limited or no insight about how the content of a revised sentence is semantically related to the content of the original sentence. Such information could facilitate the understanding of why a change is perceived as a potential improvement. For instance, it might be that a sentence is perceived as better because it uses more specific words (e.g., *book/textbook*, *bean/soybeans*, *work/homework*) or that it fits better into the context in terms of textual cohesion.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1359–1370
Barcelona, Spain (Online), December 8-13, 2020

How to Prepare for a Banking Job

Internships are a great way to get your feet wet and apply what you've learned in school to the working world (. . .) Employers will be glad to know that you made an effort to get some experience before applying for the job.

Take advantage of your summers off from school to do an internship. [original]

Take advantage of your summers off from university to do an internship. [revised]

Many banks offer programs to undergraduate students.

Figure 1: Example of an original–revised sentence pair (shown in bold) from the text *How to Prepare for a Banking job*, given the original context. The underlined term 'school', meant in a general sense, was replaced during an edit by the term 'university', which is more specific in the given context.

With this paper, we aim to fill the forementioned gaps. First, we shed light on the semantic relationships that hold between words in an original sentence and corresponding substitutions that appear in a revised sentence. Second, we investigate the extent to which a revised sentence is actually perceived as better. We approach these points by conducting three types of studies. In the first, we develop a supervised classifier that distinguishes the original and revised version of a sentence. We extend this classifier to take into account the original context (illustrated in Figure 1), making it possible to assess in how far factors related to cohesion are relevant in this distinction. In the second study, we collect annotations to verify in how far a revised sentence is actually perceived as better than its original counterpart within the original context. Finally, we compare model predictions and human annotations to analyze the extent to which our model reflects improvements perceived by humans.

For our studies, we use a set of original–revised sentence pairs from wikiHowtoImprove (Anthonio et al., 2020). The most frequent type of edit operation (substitution/deletion/insertion) in these sentence pairs are substitutions and the most frequent word category affected by this operation are nouns. Therefore, we focus our analysis on sentence pairs in which (1) one or multiple nouns are changed as a result of an edit and (2) the length of the original and revised sentence remains identical. We argue that this subset is well suited for computational and linguistic analyses as we can directly examine lexical-semantic relationships that hold between substituted nouns as well as factors related to lexical cohesion in context.

In summary, the main contributions of this paper are:

- We examine the lexical-semantic relationship between the words in the original sentence and their corresponding substitutions in the revised sentence through automatic labeling (§3).
- We develop a computational system that can identify the revised version of a sentence, while taking into account contextual features (§4).
- We evaluate the extent to which a revised sentence is being perceived as being better by humans, especially for those sentences that are not changed to correct the fluency of the text (§5).
- We compare human and machine performance for identifying the revised sentence (i.e., the sentence that is presumably 'better') and explore how the performances vary in relation to the lexical-semantic differences between the original and revised sentence (§6).

2 Related Work

There is only one related study using wikiHow (Anthonio et al., 2020), which we built upon (see Section 3). Therefore, we discuss studies within two related lines of research in this section: studies on revisions in Wikipedia (Section 2.1) and on revisions in persuasive texts (Section 2.2).

2.1 Revisions in Wikipedia

Within NLP, various applications have been developed that build upon Wikipedia revision history. Examples are preposition error correction (Cahill et al., 2013), sentence compression (Nelken and Yamangil, 2008) simplification (Woodsend and Lapata, 2011; Yatskar et al., 2010), bias detection (Recasens et al., 2013), sentence paraphrasing (Max and Wisniewski, 2010), textual entailment (Zanzotto and Pennacchiotti, 2010; Cabrio et al., 2012) and information retrieval (Aji et al., 2010; Nunes et al., 2011).

Revisions in Wikipedia have also been used to gain more theoretical knowledge on why and how texts are edited collaboratively (Bronner and Monz, 2012; Liu and Ram, 2011; Yang et al., 2017; Daxenberger and Gurevych, 2012; Faruqui et al., 2018). The majority of these studies categorize edits based on two aspects. First, whether the edits are domain-specific or not. Second, whether the edits affect the meaning of the text (text-base) or not (surface-changes). These edit categories were introduced in Faigley and Witte (1981) and extended in Jones (2008). In general, text-base edits are further subcategorized by their syntactic operation, such as deletion or modification. In this categorization, questions such as "why did authors decide to change the sentence" and "what is the effect of the change" remain mostly unanswered.

To fill this theoretical gap, Yang et al. (2017) established a 13-category taxonomy of the semantic intentions behind edits in Wikipedia, which includes categories such as *fact update, point of view, verification* and *refactoring*. They additionally investigated how various types of edits predict the retention of newcomers and changes in the article's quality using a classification model.

Our work is similar to Yang et al. (2017) in that we shed more light on those sentences that have been revised for other reasons than improving the fluency. In particular, we pinpoint the lexical-semantic relationship between nouns that have changed. Our study differs from previous work in that we investigate the effectiveness of the edits through human annotations and that we work with instructional texts instead of Wikipedia articles.

2.2 Revisions in Persuasive Texts

Another genre that has been used to examine revision histories are essays, such as those in the ArgRewrite corpus (Zhang et al., 2017). Similar as in Wikipedia, there is a body of research available devoted to the classification and analysis of edit types (Zhang and Litman, 2015; Zhang and Litman, 2016; Afrin and Litman, 2018). A closely related study to ours is the one from Afrin and Litman (2018), which aimed to develop a computational model that could predict, given the original and revised sentence, if the revised sentence is better than the original. For this purpose, Afrin and Litman (2018) showed the annotators the original and revised sentence, and asked them to label the revised sentence as *Better* or *Not Better*. The authors developed a small set of guidelines that the annotators had to use, and relied on the annotators' judgement for cases not covered by those guidelines. They let the annotators know the identity of the revised and original sentence, which might have introduced an annotation bias. The annotators obtained a slight agreement when using all results ($\kappa = 0.201$) and a fair agreement using the majority voting out of 7 annotators ($\kappa = 0.263$), where κ is Fleiss' Kappa (1971).

In the present work, we use a similar experimental set-up as Afrin and Litman (2018). We ask annotors to decide, given the original and revised sentence, which one is better than the other. Yet, our study differs in three aspects. First, we ask annotators to rely on their intuition to determine which sentence is better instead of providing detailed guidelines. Second, we hide the identity of the revised and unrevised sentence to prevent annotation bias. Third, we allow annotators to label either sentence as being better or indicate that both are equally good.

Another closely related study using ArgRewrite (Zhang et al., 2017) is the work of Zhang and Litman (2016). The authors propose to leverage context-based methods to enhance the classification of revision purposes in the ArgRewrite corpus. This is motivated by the fact that prior work, such as the studies described in Section 2.1, neglect contextual information or use relatively shallow features such as position. The authors focus specifically on cohesion and coherence related features and therefore hypothesize that the revision type affects the cohesion and coherence. The revision types considered here were introduced in Zhang and Litman (2015), containing surface-edits such as *conventions, organization* and text-base edits such as *claims, warrant, evidence* and *general*. To consider the cohesion and coherence, Zhang and

Litman (2016) compute the similarity between the two sentences before and after the revised sentence. The features significantly improved the performance of their classifier, compared to the baseline.

3 Data Description

As a starting point for our studies, we use *wikiHowToImprove* (Anthonio et al., 2020), a corpus of around 2.7 million sentences and their revisions. In the following, we use the term *revised version* and variable S_r to refer to a new version of a sentence, resulting from one or multiple edits. For the first version of a sentence, we use the term *base version* and variable S_b. In this work, we only use the base and last revised version of a sentence, dropping all intermediate versions. We further restrict our studies on a subset of the data by considering only pairwise versions of a sentence such that: (1) S_b and S_r have the same number of tokens ($N = 710\,353$), (2) at least one noun in S_b has been replaced by another noun in S_r, i.e., noun-to-noun substitutions and (3) the only edits are noun-to-noun substitutions.[1] This subset contains 261 325 noun-to-noun substitutions, spread over 240 157 pairs. Since nouns can be used to refer to entities, concepts, and events, the resulting set allows us to assess a broad range of lexical-semantic phenomena that differ in S_b and S_r. Furthermore, we suspect that the usage of nouns in a given sentence is related to the usage of nouns in the surrounding sentences, which makes it possible to study the selected subset as cases where edits may have been triggered by context.

In practice, our studies focus on sentence-level edits, that is, we distinguish only the base and revised sentence while keeping the context (i.e., the surrounding sentences) the same. Nonetheless, we provide the context of the base sentence during annotation and we derive contextual features from it for classification.

We characterize the pairs $\langle S_b, S_r \rangle$ by the lexical relationships that hold between the nouns $\{N_b^1, N_b^2 ... N_b^n\}$ in S_b and their corresponding substitutions $\{N_r^1, N_r^2 ... N_r^n\}$ in S_r. Note that we can easily identify the substitutions by their position in the sentence. For our characterization, we use the Paraphrase Database (PPDB) (Pavlick et al., 2015), which contains 100 million paraphrases. Specifically, each entry in the database contains a label that describes the entailment relationship between the phrases: *Equivalent* (synonyms), *Independent* (not related), *Exclusion* (antonym), *OtherRelated* (related but not by entailment), *ReverseEntailment* (hypernym $\rightarrow$ hyponym) and *ForwardEntailment* (hyponym$\rightarrow$ hypernym). The PPDB contains lexical, phrasal and syntactic paraphrases and contains different sizes. We use the biggest set (XXXL, 7 million instances) of lexical paraphrases (single word to single word)[2] to label as many noun-to-noun substitutions as possible. We find that 55 945 out of 261 325 noun-to-noun substitutions (21%) in the 240 157 pairs occur in the PPDB. The substitutions are distributed over 54 182 out of 240 157 pairs (23%), which indicates a fair coverage. The frequency distribution of the noun modifications (~~base formatted in strikethrough~~/ revised underlined) is as follows:

- Independent ($N = 24\,373$): ~~ingredients~~/food, ~~story~~/novel, ~~days~~/time
- Equivalence ($N = 14\,308$): ~~percentage~~/%, ~~kid~~/child, ~~realisation~~/realization, ~~maths~~/mathematics
- OtherRelated ($N = 11\,289$): ~~click~~/right-click, ~~week~~/hour, ~~hand~~/hands
- Forward entailment ($N = 2\,771$): ~~woman~~/person, ~~car~~/vehicle, ~~army~~/military
- Reverse entailment ($N = 2\,686$): ~~work~~/homework, ~~plants~~/flowers, ~~eat~~/kitten
- Exclusion ($N = 518$): ~~right~~/left, ~~high~~/low, ~~outside~~/inside, ~~red~~/blue.

In the remainder of this paper, we continue experimenting with the 54 182 pairs. We work with those cases to restrict our analyses and experiments to substitutions containing valid English words that are semantically related. Yet, this restriction does not eliminate all fluency-related errors, as some can be

[1] Despite our efforts, there are a approx. 15% cases in the dataset which have other changes than nouns. This can be caused by POS-tagging errors or by other pre-processing errors. We manually identified such cases in a posthoc analysis and verified that they had no noticable effect on the findings of our studies.

[2] `http://paraphrase.org/#/download`

caused by wrong inflections (e.g., a ~~students~~/student). We do not apply any additional filters as we aim to cover as many interesting cases as possible in our study.

4 Classification Experiments

In this section, we describe a set of classification experiments that we conducted using the 54 182 pairs described in Section 3. For these experiments, we divide the data by article into a train ($N = 44\,147, \sim$80%), development ($N = 5\,116, \sim$10%) and test set ($N = 4\,919, \sim$10%).[3] We address two settings. Firstly, we train a classifier to identify for a sentence if it belongs to the class of revised versions. We call this setting *sentence in isolation*. Our hypothesis is that revised versions can be distinguished from base versions by modeling differences in noun choice within the sentence. Secondly, we develop classifiers that also take the base context into account. That is, the classifier can make use of features of the sentence to be classified as well as features derived from the sentences surrounding the sentence in its base version. For this setting, we extract the sentences before and after S_b using a window of five sentences. We denote the previous sentences as C_p and the following sentences as C_f. We use the best classifier from the first setting as a baseline for the second setting, henceforth *sentence in context*, and our hypothesis is that features derived from the sentence context can improve the identification of revised versions by means of modeling factors such as cohesive ties, which are known to contribute to the readability of a text (Mahlberg, 2006; Pitler and Nenkova, 2008).

4.1 Experimental Setup

Model. In both settings, we use a bidirectional long short-term memory network (BiLSTM) as our classification model (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997). The input layer of the BiLSTM is initialized with pre-trained word embeddings. In the sentence in isolation setting, we experiment with various types of embeddings, including FastText (Bojanowski et al., 2017), GloVe (Pennington et al., 2014), BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). Furthermore, we use two stacked BiLSTMS and experiment with dimensions for each BiLSTM of 256 and 512. We concatenate the last forward and backward hidden layers to obtain a final representation, which is fed into a linear layer that outputs a single classification score. Between the stacked BiLSTMs and on the final representation, we apply Dropout (Srivastava et al., 2014) with a probability of 0.5. At training time, we use the labels 0 and 1 for S_b and S_r, respectively, and we compute the loss using Binary Cross Entropy Loss[4]. At evaluation time, we use the sigmoid function to obtain and compare the output values for the base and revised sentence.

For the sentence in context setting, we use the classifier with the best embeddings as our baseline. We take context into account by concatenating context-based features to the final hidden representation of the BiLSM. For normalization, we use batch normalization. Given the previous and following context C_p and C_f, respectively, and a sentence S in its base or revised version (S_b vs. S_r), we experiment with the following features:

- Discourse Markers: this feature represents a count of discourse markers in the tuple $\langle C_p, S, C_f \rangle$.[5]
- Type–Token Ratio: the ratio of word types (i.e., unique tokens) by word tokens in the tuple $\langle C_p, S, C_f \rangle$.
- Noun Overlap: the frequency of the base/revised noun(s) in the tuple $\langle C_p, S, C_f \rangle$.
- Left Similarity: this feature measures the cosine similarity between the vector representations of the last sentence in C_p and S. We compute each sentence-level representation as a component-wise product over the GloVe embeddings of each word in the sentence.

[3] We use the article-based split released with wikiHowToImprove, which is available here: `https://github.com/irshadbhat/wikiHowToImprove/tree/master/data`

[4] For computing the loss, we use the PyTorch implementation `BCEWithLogitsLoss`, which first projects the output of the linear layer onto a range of $[0, 1]$ by applying the sigmoid function (see `https://pytorch.org/docs/stable/nn.functional.html`)

[5] Using the explicit connectives occurring in the Penn Discourse Treebank 2.0 (Prasad et al., 2008), as listed in the manual `https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf`

- Right Similarity: same as left similarity but using the first sentence of C_f.

We selected those features because they can be used, to a certain extent, as indicators of the cohesion and coherence of a text (Crossley et al., 2016; Zhang and Litman, 2016). We suppose that the revised version of a sentence may exhibit higher cohesion and coherence in context than the base version, and that the features can capture this effect.

Training and evaluation. All models are trained on the train set and evaluated on the development set after each epoch of training. The baseline is trained for a total of 30 epochs. All other models are trained for at most 30 epochs, with early stopping applied if there is no improvement within 10 epochs. On the test set, we only evaluate our baseline and the best models using context features. Regarding the evaluation measure, we follow the same approach as Anthonio et al. (2020): Given a pair of sentence versions $\langle S_b, S_r \rangle$, we treat the prediction of the classifier as correct if the score for the base sentence S_b is smaller than the score for the revised sentence S_r, and report the average accuracy over all predictions.

4.2 Results

We show the results of the BiLSTM classifiers in Table 1. All classifiers achieve an overall accuracy score between 65% and 70%. The best classifier in the sentence in isolation setting uses RoBERTa embeddings and scores an accuracy of 68.26% and 66.33% on the development and test set, respectively. Note that this score is much higher than the expected accuracy of a random classifier (50%), confirming our hypothesis that the differences in noun choice can be modeled computationally. Two of our context-based features lead to improvements over this baseline: the addition of the Type–Token Ratio scores an accuracy of 69.33% on the development set and 66.64% on the test set; the addition of the Noun Overlap feature yields an accuracy of 69.55% on the development set and 68.92% on the test set. This confirms our hypothesis that features derived from context can improve the modeling of noun usage within a base or revised sentence.

(a)

S **in isolation** (dev)	$h = 256$	$h = 512$
GloVe embeddings	64.50	66.52
RoBERTa embeddings	**68.26**	68.00

(b)

S **in context** (dev)	$h = 256$	$h = 512$
baseline + Discourse Markers	68.43	67.88
baseline + Type–Token Ratio	68.45	69.33
baseline + Noun Overlap	69.53	**69.55**
baseline + Left Similarity	67.69	67.68
baseline + Right Similarity	66.47	66.36

(c)

S **in isolation** (test)	
best baseline	66.33
S **in context** (test)	
baseline + Type–Token Ratio	66.64
baseline + Noun Overlap	**68.92**

Table 1: Accuracy results (in percentage) of (a) different baseline configurations, (b) context features on the development set, and (c) the best models for sentence in isolation / in context on the test set.

5 Annotation Study

Previous work introduced the assumption that revisions represent improvements that can be made to a base version of a sentence. In our classification experiments, we relied on the observation that the noun choices in the base and revised version of a sentence can be distinguished systematically, thereby making implicit use of this assumption as well. In this annotation experiment, we test in how far this assumption actually reflects human judgements. For this purpose, we ask human annotators to identify which of two versions of a sentence they perceive as better or if both versions are equally good in their opinion. We present each pair of sentence versions within the base context C_p and C_f. Each pair consists of a sentence in its base version and in its last revised version. For the annotation, we randomize the ordering of both versions in order to hide the identity from the annotators. This procedure allows us to verify in

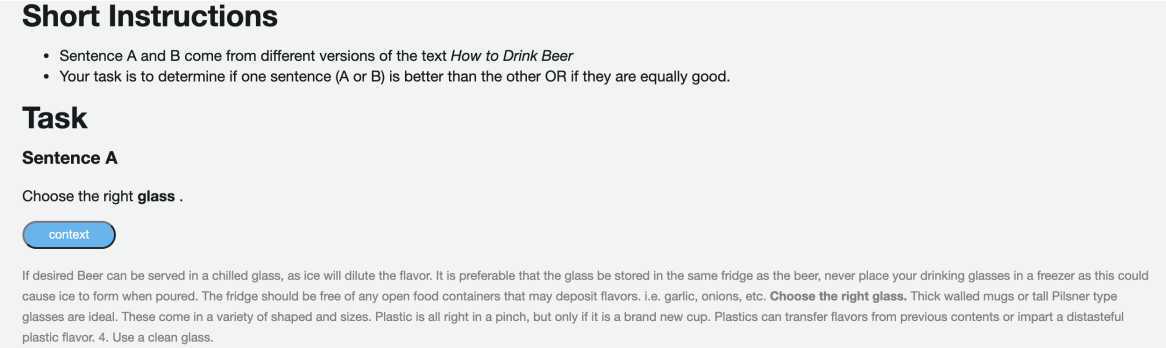

Figure 2: A screenshot of a part of the interface that we used in the annotation study.

an unbiased way whether annotators indeed label the revised sentence as being better or not. In addition, the annotation study provides us valuable insights regarding the difficulty of the task at hand.

5.1 Experimental Set-up

We use a subset of the dataset described in Section 3. In particular, we select a random sample of 499 base–revised sentence pairs from the development set that we used in the classification study. We show a part of the interface that we used to collect the annotations in Figure 2. The interface displayed two sentences to the annotators, Sentence A and Sentence B. When the annotators clicked on the button "context", the context of the sentence from the base version was displayed below the button. Thus, it was optional for the annotators to consider the context. Since the purpose of the annotation study is to find out which sentence is better, we find it acceptable if the annotators manage to perform the task without looking at the context.

Furthermore, on the bottom of the interface, we posed the question: "Which sentence is better?". Using the form, the annotators could answer: Sentence A, Equally Good, or Sentence B. For our analyses, we convert indicated preferences to "revised is better" and "base is better" labels, based on the hidden identity of a selected sentence. In addition to collecting individual annotations, the annotators had to write a short report that explains how they made their decisions. We used this information in our analyses to understand how the annotators define "better" and "equally good".

We asked two annotators to participate in our study. Both are graduate students with expertise in (computational) linguistics. We did not provide any instructions apart from the short info shown in Figure 2, such that annotators can decide for themselves what makes a sentence better than the other. Of course, keeping the task intuitive comes at a cost of additional disagreements in cases where intuitions differ. However, the goal of the annotation study was merely to find out whether revisions are perceived to be better than their previous versions, rather than providing a gold standard for further experiments. Therefore, we decided to omit additional annotation steps, such as refining guidelines or adjudicating cases of disagreement.

5.2 Results

	Annotator 2			
Annotator 1	Revised is better	Equally good	Base is better	**Total**
Revised is better	153	65	7	225
Equally good	35	142	30	207
Base is better	16	32	19	67
Total	204	239	56	499

Table 2: Confusion Matrix comparing the labels given by Annotator 1 and 2.

We present the confusion matrix of the two annotators in Table 2. It shows that the annotators marked for most pairs the revised sentence as the better sentence. Since they labeled an almost similar number of pairs as equally good, we conclude that the assumption "revised is better" holds depending on the characteristics of the substitutions. Overall, the percentage of overlapping judgements is 62.93% ($N = 314$) and the chance-corrected agreement score is $\kappa = 0.384$ (Cohen, 1960). This is viewed as a fair agreement (McHugh, 2012) and higher than the scores obtained in related studies (see Section 2).

Revised is better. As shown in Table 2, there are 153 pairs where both annotators marked the revised sentence as better. Most of them ($N = 88, 58\%$) contain a fluency-related error in the base sentence resolved in the revised version. The errors include spelling errors, grammatical errors caused by wrong inflection and incorrect usage of certain words, such as *number/amount*. There are two prominent characteristics in the remaining cases. Firstly, a revised sentence is perceived as better when it contains a more specific noun than the base sentence. This point was addressed by both annotators in their report. Most noun modifications in these pairs are tagged with an Independent relation in the PPDB, because they lack a direct hypernym/hyponym relationship (e.g., ~~*one*~~*/option*, ~~*stuff*~~*/content*, and ~~*things*~~*/ foods*). This seems to be a caveat of the labels provided by the PPDB. Secondly, the revised sentence is also perceived as better when it contains a more commonly used noun or yields a more commonly used phrase than used in the base sentence (e.g., ~~*cooking*~~*/baking powder,* ~~*shadow*~~*/eyeshadow*). For some of these instances, common knowledge about the topic could also have played a role (e.g., ~~*Microsoft*~~*/Windows PC*).

Equally Good. The annotators marked 142 sentence pairs as equally good. Most modifications were synonym replacements (e.g., ~~*pictures*~~*/photos*), most of which are tagged with the Equivalent relation in the PPDB. In addition, we found 9 sentence pairs with antonym substitutions that were perceived as equally good (e.g., ~~*left*~~*/right*). Finally, a number of equally good sentences contain nouns that only vary due to geographical spelling variations ($N = 26, 18.31\%$). Both annotators mentioned this point in their report.

Base is better. As shown in Table 2, there are only 19 pairs where both annotators marked the base sentence as better. In four cases, the base sentence was clearly better because of a fluency-related error introduced in the revised sentence. For the remaining cases, it is unclear why the annotators selected the base sentence as better. Based on the low number of cases labeled as "base is better", we conclude that when the assumption 'revised is better' does not hold, it is likely that base and revised sentence are equally good.

Disagreements. In 185 out of 499 pairs, the annotators provided different labels. As illustrated in Table 2, a major set of disagreements is based on that Annotator 1 marked more cases as revised is better ($N = 225, 45.09\%$) and Annotator 2 labeled more pairs as equally good instead ($N = 239, 47.90\%$). One cause of this difference could be that Annotator 1 only selected the sentence with the more specific noun as being better when this was appropriate within the topic using the title of the text. The pairs with overlapping judgements do not rule out that Annotator 2 used a similar consideration, but contextual aspects were not mentioned in her report. Other disagreements are caused by differences in common knowledge, intuitions (e.g., selecting the pair with the most common word) and annotation inconsistencies. The latter is particularly visible in the pairs labeled by one annotator as "base is better". Finally, some disagreements may have been caused by differences in usage of context, since viewing the context was optional.

6 Analysis

In this section, we compare classifier predictions and human judgements. As discussed in Section 4, our best classifier can distinguish the base and revised version of a sentence with an accuracy of close to 70%. In Section 5, we have seen that annotators do not always perceive the revised sentence as the better version of a sentence. Thus, we examine in how far classifier predictions overlap with human judgements. For this analysis, we use the 314 pairs of sentences from the development set that were annotated identically by both annotators. For a direct comparison, we post-process the model predictions to also include a category "equally good", which comprises all cases where the model's output scores for both versions of a sentence correspond to the same label (e.g., score$(S_b) = 0.6$ and score$(S_r) = 0.7$). In

the remaining cases, we treat the model's prediction as "revised is better" if score(S_r) >score(S_b) and "base is better" if score(S_b) >score(S_r).

6.1 Results

	Classifier			
Annotators	Revised is better	Equally good	Base is better	**Total**
Revised is better	59	83	11	153
Equally good	59	59	24	142
Base is better	4	13	2	19
Total	122	155	37	314

Table 3: Confusion matrix for the cases where both annotators agreed compared to classifier predictions.

Correct predictions. The classifier's correct identification of a revised sentence overlaps with the annotator's judgement that the revised sentence is better in 59 cases. In 46 (77,96%) of those pairs, a fluency-related error in the base sentence was resolved in the revised sentence. These errors include typos ($N = 31$), grammatical mistakes ($N = 8$), a combination of both ($N = 2$) and incorrect word usages ($N = 5$). Most of the remaining pairs involved the PPDB relation Independent ($N = 8$, e.g., ~~*food*~~/*dish*, ~~*dirt*~~/*soil*). Thus, the only improvements (according to the annotators) that are systematically identified by the classifier are cases in which the revised sentence has a higher fluency. Among the 59 cases that the classifier and annotators identified as equally good, we find that the noun substitutions frequently involve the PPDB relations *Equivalence* ($N = 19$, e.g., ~~*bike*~~/*bicycle*, ~~*kids*~~/*children*, ~~pictures~~/photos), *Independent* ($N = 18$, e.g., ~~*drawings*~~/*pictures*) and *Exclusion* ($N = 4$, e.g., ~~*left*~~/*right*).

Incorrect predictions. There are 83 pairs where the annotators marked the revised sentence as being better, while the classifier labeled both sentences as equally good. We noticed that a substantial number of those pairs contained grammatical errors ($N = 13$) or mispelled words ($N = 22$) in the base sentence which were solved in the revised sentence. Yet, the errors do not systematically differ from the correctly identified pairs that we mentioned in the previous paragraph. Therefore, it is difficult to pinpoint why those cases where incorrectly predicted by the model. In the remaining 48 cases, the disagreements seem to be caused by factors related to linguistic specificity, commonality, and factual knowledge. For instance, there are several cases where the annotators choose a sentence with the more specific noun as being better ($N = 7$) or the sentence with the more commonly used word ($N = 9$). The former includes noun substitutions not directly related by hypernymy, as addressed in Section 5. Consequently, these cases can be hard to recognize for the classifier. Moreover, there are several cases where annotators presumably used common sense or factual knowledge ($N = 16$) to determine the better sentence. Two examples are: (1) *dpi is dots per* ~~*minutes*~~/*inch* and (2) *the user*/~~*viewer*~~ *need to protect her/himself from software threats*. Finally, the annotators made effective use of neighbouring words in a sentence to determine the correct noun, for example: "*The NFPA says to remember the* ~~*word*~~/*acronym PASS*" or took the context/topic into account if necessary.

7 Conclusion

In this work, we examined a subset of revisions in wikiHow, namely substitutions that involve lexically related nouns, in order to assess the role of lexical semantics in annotating and modeling revisions. In our computational experiment, we found that differences between base and revised sentence can be modeled with high accuracy ($> 65\%$) and that context-based features regarding cohesion and coherence can further improve this result. In an annotation study, we verified that humans identify the revised sentence as better when it improves the fluency. However, when pairs with noun substitutions are closely related, such as by synonymy or antonymy, annotators tend to label both sentences as equally good. In some cases, our annotators had to rely on their intuitions and background knowledge to select whether

one sentence is better than another. Yet, we have also seen a high number of disagreements, showing that the assumption "revised is better" does not always apply or is at best subjective in some cases.

In a final study, we observed that model predictions are largely in line with human annotations when the revised sentence improves the fluency by correcting mispelled words. In contrast, we found our best model to struggle with pairs that involve noun substitutions with high semantic relatedness and cases where humans can resort to background knowledge or intuitions. In future work, we plan to extend our model and analyses to take into account further information, including measures and features regarding ambiguity and document structure.

Acknowledgements

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

References

Tazin Afrin and Diane Litman. 2018. Annotation and classification of sentence-level revision improvement. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–246, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ablimit Aji, Yu Wang, Eugene Agichtein, and Evgeniy Gabrilovich. 2010. Using the past to score the present: Extending term weighting models through revision history analysis. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 629–638, New York, NY, USA. ACM.

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, May. European Language Resources Association.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Amit Bronner and Christof Monz. 2012. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 356–366, Stroudsburg, PA, USA. Association for Computational Linguistics.

Elena Cabrio, Bernardo Magnini, and Angelina Ivanova. 2012. Extracting context-rich entailment rules from wikipedia revision history. In *Proceedings of the 3rd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources and Their Applications to NLP*, People's Web '12, pages 34–43, USA. Association for Computational Linguistics.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. 2013. Robust systems for preposition error correction using Wikipedia revisions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 507–517, Atlanta, Georgia, June. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4):1227–1237.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proceedings of COLING 2012*, pages 711–726, Mumbai, India, December. The COLING 2012 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College Composition and Communication*, 32(4):400–414.

Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. WikiAtomicEdits: A multilingual corpus of Wikipedia edits for modeling language and discourse. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 305–315, Brussels, Belgium, October-November. Association for Computational Linguistics.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80, 12.

John Jones. 2008. Patterns of revision in online writing. a study of wikipedia's featured articles. *Written Communication*, 25:262–289, 04.

Jun Liu and Sudha Ram. 2011. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manage. Inf. Syst.*, 2(2):11:1–11:23, July.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Michaela Mahlberg. 2006. Lexical cohesion: Corpus linguistic theory and its application in english language teaching. *International Journal of Corpus Linguistics*, 11:363–383, 01.

Aurélien Max and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

M. L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22:276–282.

Rani Nelken and Elif Yamangil. 2008. Mining wikipedia's article revision history for training computational linguistics algorithms. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 31–36.

Sérgio Nunes, Cristina Ribeiro, and Gabriel David. 2011. Term weighting based on document revision history. *JASIST*, 62:2471–2478, 12.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015. Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1512–1522, Beijing, China, July. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii, October. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria, August. Association for Computational Linguistics.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, November.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, Copenhagen, Denmark, September. Association for Computational Linguistics.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368, Los Angeles, California, June. Association for Computational Linguistics.

Fabio Massimo Zanzotto and Marco Pennacchiotti. 2010. Expanding textual entailment corpora fromWikipedia using co-training. In *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 28–36, Beijing, China, August. Coling 2010 Organizing Committee.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–143, Denver, Colorado, June. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2016. Using context to predict the purpose of argumentative writing revisions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1424–1430, San Diego, California, June. Association for Computational Linguistics.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578, Vancouver, Canada, July. Association for Computational Linguistics.