Implicit Phenomena in Short-Answer Scoring Data

Marie Bexte, Andrea Horbach and Torsten Zesch
Language Technology Lab, University of Duisburg-Essen, Duisburg, Germany
(firstname.lastname)@uni-due.de

Abstract

Short-answer scoring is the task of assessing the correctness of a short text given as response to a question that can come from a variety of educational scenarios. As only content, not form, is important, the exact wording including the explicitness of an answer should not matter. However, many state-of-the-art scoring models heavily rely on lexical information, be it word embeddings in a neural network or n-grams in an SVM. Thus, the exact wording of an answer might very well make a difference. We therefore quantify to what extent implicit language phenomena occur in short answer datasets and examine the influence they have on automatic scoring performance. We find that the level of implicitness depends on the individual question, and that some phenomena are very frequent. Resolving implicit wording to explicit formulations indeed tends to improve automatic scoring performance.

1 Introduction

Automatic short answer scoring is an application area of natural language processing where short free-form answers written by students in an educational context are automatically scored based on the correctness of their content. They occur for example in science education (Nielsen et al., 2008; Dzikovska et al., 2010), but also in foreign language learning to measure reading (Bailey and Meurers, 2008; Meurers et al., 2011) or listening comprehension (Horbach et al., 2014).

In such a scoring task, answers are graded based on their content alone - in comparison to essay scoring (Attali and Burstein, 2006) where also linguistic form is taken into consideration. Thus, judging whether an answer is correct or not may require the resolution of a number of implicit language phenomena as a form of normalization. Figure 1 shows two answers that express the same content, but with differing levels of explicitness. How the content is expressed on the surface does not matter for the score.

Implicit:
3 is the perfect amount,
2 is not enough,
3 is too many.

Explicit:
3 *scoops* is the perfect amount *of fertilizer*,
because 2 *scoops* is not enough,
but 3 *scoops* is too many.

Figure 1: Two (made-up) answers to the same prompt demonstrating how one can say the same thing with different levels of explicitness.

In fact, the two answers in the example should be treated in the same way regardless of their explicitness. The only relevant criterion should be whether they convey the right content and thus show that the learner understood the concepts. While humans often effortlessly resolve implicit phenomena, automatic resolution of many of these phenomena is not trivial. However, we argue that resolution of implicitness is a kind of normalization step that can help to improve automatic scoring performance.

Most work on automatic short-answer scoring does not actively resolve most implicit phenomena. However, the c-rater system performs pronoun resolution (Leacock and Chodorow, 2003), but they do not report the impact of that single component. Banjade et al. (2015) perform implicit resolution of coreferences between entities in learner answers and entities in the question and similarly target ellipses resolution, where part of the question is implied in the learner answer, both by aligning concepts from the learner answer to the question. They

Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language (UnImplicit 2021), pages 11–19
Bangkok, Thailand (online), August 5, 2021.

report a positive influence on overall scoring performance. Another notable exception is information structure, i.e. whether the answer repeats parts of the question as researched through focus annotations by Ziai and Meurers (2014). They report only a minor effect on automatic scoring performance.

In this paper, we analyse which implicit phenomena occur in short answer scoring datasets. We then analyze the impact of implicit language on automatic scoring performance.

2 Implicit Language in Learner Answers

There are a number of linguistic phenomena that pertain to the implicitness of language and are especially relevant for learner answers. In the following, we describe the ones we considered as candidates for our analysis.

Coreference Coreference describes the phenomenon that the same entity is referred to several times throughout a text, often using different referring expressions (see (Mitkov, 2014)). The most prototypical example of pronominal reference is shown in Example 1, where *they* at the beginning of the second sentence refers to the same entity as *pandas* in the first sentence.

– Pandas live in China. They eat bamboo.
– Pandas live in China. **Pandas** eat bamboo.

Example 1: Coreference

Bridging Anaphora The relationship between an anaphor and its antecedent may be indirect, constituting the special case of bridging anaphora (Clark, 1975). Take for example the statement shown in Example 2. While this can be understood from the context of the first sentence, it is left implicit that the second sentence refers to the fur *of the panda*.

– The panda is ill. The fur is dull.
– The panda is ill. The fur **of the panda** is dull.

Example 2: Bridging

Ellipsis An ellipsis is the omission of content that can be derived from context (see Example 3). There, the second sentence does not explicitly state that koalas are *highly specialized*, too, which can however be gathered from the first sentence.

– Pandas are highly specialized. Koalas are, too.
– Pandas are highly specialized. Koalas are **highly specialized**, too.

Example 3: Ellipsis

Numeric Terms In numeric expressions, the head word, i.e. usually the measurement unit, can often be left out. In cases with parallelism to a previous sentence this is a sub-type of an ellipsis, in others it is not (Elazar and Goldberg, 2019). Example 4 shows an instance of the latter case, where the implication is that this sentence talks about age, indicated by the use of *turn* in front of *30*. Instead of saying that pandas *turn 30 years* old, this is shortened to saying that they *turn 30*.

– Pandas turn 30 in the wild.
– Pandas turn 30 **years** in the wild.

Example 4: Numeric Terms

Information Structure Another specific subcase of ellipses that is particularly important in a question and answer scenario is information structure (Krifka and Musan, 2012), i.e. the distinction whether the answer repeats given information from the question. Given the question that is shown in Example 5, *bamboo* is the focus of the answer, that actually answers the question. Focus has been automatically annotated for short answer data, although focus-based feature made only a minor difference in scoring performance (Ziai and Meurers, 2018).

– *What do pandas eat?* Bamboo.
– *What do pandas eat?* **Pandas eat** bamboo.

Example 5: Information Structure

Presupposition A presupposition (see Example 6) is a precondition that has to be fulfilled for a sentence to be true or false (Strawson, 1950). The statement *pandas no longer eat bamboo* presupposes that pandas used to eat bamboo, which then makes it a valid statement to say that they no longer do.

– Pandas no longer eat bamboo.
– **Pandas used to eat bamboo.** Pandas no longer eat bamboo.

Example 6: Presupposition

Restrictive vs. Non-restrictive Remarks Any appositional adjective and any relative clause (Fabb, 1990) can either be restrictive, i.e. necessary for selecting the right entity out of a set of alternatives or non-restrictive. In the question

Explain how pandas in China are similar to koalas in Australia.

in China is non-restrictive (because it is not meant to differentiate between different kinds of pandas living in different parts of the world). We could think of such non-restrictive terms as the explicit version of an implicit sentence. Especially in a learner answer targeting that question the term *pandas* can be used, implicitly meaning *pandas in China*.

Implicit Discourse Relations The relation between sentences is often marked by discourse connectives. In some cases, there may be a discourse relation that is left implicit. With regard to the statement shown in Example 7, there is such a relation between the two sentences, which is an implicit *therefore*, as the reason for taking the panda to the veterinarian was its dull fur.

– The panda had dull fur. We took it to the vet.
– The panda had dull fur, **therefore** we took it to the vet.

Example 7: Implicit Discourse

3 Implicitness Annotations

Short answer-scoring datasets can include very different *prompts*, i.e. an (optional) reading text and some question the student has to answer, coming from domains such as sciences, biology, or English language arts. To cover a range of different *learner answers*, we select prompts from two short answer datasets and annotate occurrences of the implicit phenomena within the learner answers given in response to these prompts.

This procedure has three goals: First, we want to assess the frequency of these phenomena in learner data. Second, we want to evaluate the effect of implicitness on the final score an answer receives, i.e. we ask whether implicit answers are on average scored higher or lower than explicit ones by teachers. And finally, we want to know the effect of implicitness on automatic scoring performance. We investigate this third question by extracting explicit versions of the answers regarding the different phenomena from the implicit versions.

3.1 Datasets

For our annotations we needed publicly available short-answer data in English where answers are full sentences and not only single phrases like in the Powergrading dataset (Basu et al., 2013). Ideally, there should be a larger amount of answers for a single prompt so that prompt-specific models can be trained later in Section 4. (For an overview of publicly available shortanswer datasets, see Horbach and Zesch (2019).) We consider two short answer datasets in our analysis. The first one is the Student Response Analysis Corpus (SRA) of the 2013 SemEval task 7 (Dzikovska et al., 2013). It consists of data from two different sources. The *Beetle* subset has 3k student answers to 56 questions about electricity and electronics. The *SciEntsBank* subset contains 10k student answers to 197 questions about different science domains. All questions have a reference answer and (among others) 5-way labels judging the appropriateness of the student answers.

The second dataset we consider is that of the 2012 Automated Student Assessment Prize (ASAP).[1] It consists of about 2,200 student answers to each of ten science-related prompts. The answers to four of the prompts were rated on a four-point scale and the others received scores on a three-point scale.

3.2 Annotation process

Our annotation study focuses on four of the phenomena we presented in the introduction. These are coreference, bridging anaphora, ellipsis and numeric terms. We chose them as we expected them to be relatively frequent, based on a short manual inspection of the data, and because they can all be annotated following the same general schema, which we describe below. Thus, we expected that they would have a larger influence on automatic scoring performance. For each of them, we selected prompts from one of the datasets that

[1] https://www.kaggle.com/c/asap-sas

Phenomenon	Dataset	Prompt	# Answers
Coreference	ASAP	8	100
Bridging Anaphora	SRA	LF_26b2	40
Bridging Anaphora	SRA	ST_31b	40
Ellipsis	ASAP	2	100
Numeric Terms	SRA	LF_27a	40
Numeric Terms	SRA	VB_22c	40

Table 1: Prompts selected for annotation of the implicit phenomena.

seemed to contain instances of that phenomenon in larger quantities. For the ASAP data, we randomly sampled 100 of the answers to the selected prompt. As some of the SRA prompts only have 40 answers, we in these cases selected two suitable prompts to arrive at a combined amount of 80 candidate sentences. Table 1 shows the chosen prompts.

Coreference, numeric terms and bridging anaphora were all annotated following the same pattern. An occurrence of any of these phenomena is marked by annotating the span, which is then linked to the last explicit mentioning of what is necessary to resolve the phenomenon. Take for example a sentence *30 meters plus 20 is 50*. Here, both *20* and *50* would be annotated and linked back to *meters*. Ellipses were annotated in the same way, but following the convention that the token before the ellipsis was linked to what is necessary to resolve the ellipsis.

In some instances, there was no explicit mentioning of what is necessary to resolve implicit into explicit. Depending on whether this could be inferred from the context we then either directly annotated these spans with their resolved form or marked them as non-resolvable.

3.3 Annotation analysis

All answers were double-annotated by two of the authors of this paper to calculate two different measures of agreement. The first one is the **token-level agreement** on whether a token was annotated as covering the phenomenon. The other is the **antecedent agreement**, which is based on the subset of tokens where both annotators agreed that a token was part of a chain. Here, we only check those tokens that were not the first item in a coreference chain. For those, we checked whether they linked to the same antecedent.

Table 2 shows the agreement results. The κ token-level agreement ranges between .74 and .86 for all phenomena, except ellipsis where it is only .45. Ellipses seem to be hard to annotate. While both annotators found the same amount of instances, they substantially disagreed what exactly to label. One example for such a problematic instance was the sentence *Plastic A is the most stretchy* that could be either interpreted as a normal superlative or as leaving out the head (*the most stretchy plastic*).

Antecedent agreement is .90 and above for coreference, bridging and ellipsis, but lower for numeric terms with values between .51 and .7. With respect to prompt VB_22c, this arises from the fact that many answers reference numbers for which the context suggest that they represent some kind of unit of weight, but while one annotator did not find the context clues sufficient to resolve this, the other linked these numbers back to the span *mass of beans* mentioned in the prompt question. Example 8 shows the prompt and an example answer where this occurs. While both annotators agreed on the whole numbers being *scoops*, the decimal numbers created disagreement, with one annotator linking them to *mass of beans*, the other marking them as unresolvable. Without disagreement arising from this particular phenomenon, antecedent agreement increases to .81.

- **Question:**
  Describe what the graph tells you about the relationship between the number of scoops of fertilizer and the mass of beans harvested?
- **Answer:**
  It goes in a pattern like 0 is on 0.2 and like one is on 0.7 and goes from even to odd.

Example 8: Annotation of numeric terms

Table 3 shows how frequently the different phenomena occur within the prompts. As we did not curate the two sets of annotations, the reported phenomenon counts are based on the first annotator, who is the same for all of them. The most prevalent phenomenon is coreference, with 97 out of the 100 answers we annotated containing at least one instance of it. The two prompts we chose for the annotation of bridging anaphora differ in the frequency of answers with bridging, as 80% of the answers to one of the prompts contain instances of bridging, whereas just 18% of the other do. With respect to ellipsis and numeric terms we find that 40% of the answers contain ellipsis, and that 30% of the answers to VB_22c and 50% of the answers to LF_27a contain at least one unre-

Phenomenon	κ Token-level Agreement	% Antecedent Agreement
Coreference (ASAP_8)	.74	.91
Bridging Anaphora (LF_26b2)	.86	.91
Bridging Anaphora (ST_31b)	.80	1.00
Ellipsis (ASAP_2)	.45	.93
Numeric Terms (VB_22c)	.85	.51
Numeric Terms (LF_27a)	.76	.70

Table 2: Binary token-level and antecedent agreement for the annotation of the phenomena.

Phenomenon	% LA w/ phen.	ø # phen. per LA	Scores of LAs w/ phen.	Scores of LAs w/o phen.
Coreference	97	5.0		
Bridging Anaphora (LF_26b2)	80	0.9		
Bridging Anaphora (ST_31b)	18	0.2		
Ellipsis (ASAP_2)	40	0.9		
Numeric Terms (VB_22c)	30	1.2		
Numeric Terms (LF_27a)	50	1.0		

Table 3: Frequency with which the phenomena occur in the chosen prompts shown in Table 1. For the label distribution, individual labels from left to right are: 0, 1 and 2 points for Coreference, 0, 1, 2 and 3 points for Ellipsis and *contradictory*, *irrelevant*, *partially correct*, *correct* for the other phenomena.

solved numeric term. Apparently some phenomena are more frequent than others even when selecting datasets that seem most suitable for a certain phenomenon. While coreference by means of pronouns is a common phenomenon where sentences avoiding it completely would look marked, students in a school context might be less inclined to leave out, e.g., units of measurement in an exam situation.

In Table 3, we also report on the question of whether explicit or implicit answers are scored higher by humans and find mixed results.

As only three of the answers to ASAP prompt 8 did not contain coreferences, we cannot compare how the assigned labels may differ between answers with and without coreference.

In the case of bridging, the two prompts we chose also exhibit different patterns. Within the answers to prompt LF_26b2, the majority contains instances of bridging and those that do not tend to be labeled worse, most frequently as *irrelevant*. The other bridging prompt, ST_31b, contains fewer instances of bridging, and those answers that include bridging receive worse labels, most frequently *irrelevant*. Therefore, a typical answer to the LF_26b2 prompt seems to be one with bridging, with those that do not contain bridging receiving lower scores. A typical answer to the ST_31b prompt on the other hand is one without bridging, with those that do contain it getting lower scores.

For numeric terms, while answers to the VB_22 prompt that contain unresolved numeric terms generally receive good labels of either *partially correct* or *correct*, the other prompt we chose does not exhibit such a pattern. There, answers with unresolved numeric terms are equally likely labeled as *contradictory* or *correct*. We also see very similar label distributions for answers with and without ellipsis.

Overall we do not see a clear trend, which is reassuring, as teachers scoring such answers manually are probably not influenced by the presence or absence of implicit language (although of course a controlled annotation study would be needed to confirm this). In the next section, we will check whether automatic scoring models are equally unimpressed by the choice of wording in a learner answer.

4 Impact of Implicit Language on Automatic Scoring

As we have seen in our dataset analysis, there is a large variance whether learners use implicit or explicit language. However, as in content scoring, only the meaning and not the form of an answer is important, both variants should be scored by an automatic scoring model in completely the same way. Many state of the art models heavily rely on lexical information, be it word embeddings in a neural network or n-grams in an SVM. Thus, the exact wording of an answer might very well make a difference, especially if one variant is much more frequent than the other and therefore only rarely seen in the training data. To asses the extent of the influence of implicitness, we perform in this section automatic scoring experiments that control for the implicitness of our annotated phenomena in the data.

4.1 Experimental setup

For our experiments we use Weka's (Hall et al., 2009) SMO Support Vector classifier in standard configuration with the top 10,000 most frequent token uni- to trigram and the 1,000 most frequent POS uni- to trigram features, and train a separate classifier per prompt.[2] Due to the small amount of answers, we perform leave-one-out cross validation.

4.2 Controlling the amount of explicitness in the data

In order to assess the impact of implicitness, we compare two versions of the dataset, making use of our annotations. In the **baseline** condition, the training and test data is used as is. In the **explicit** condition, we use the antecedent annotations to resolve any implicit phenomena to their explicit version and then train and test on explicit answers.

Figure 2 shows examples for implicit and explicit versions of the four phenomena. For coreference, we resolve every pronoun to obtain the explicit version. For bridging and numeric terms, we add what is necessary to resolve them. In case of ellipsis, we add what was left out.

4.3 Experimental results

Table 4 shows the results of our experiments. Because the SemEval labels do not have a natural order, we report κ values for them, but QWK for the ASAP prompt. For the two ASAP prompts, we only had 100 annotated answers and hence a much smaller amount than the full set of answers that is typically used to train models on this dataset. This is reflected in a reduced performance compared to other experiments on the same dataset, but the focus of our experiments is rather to assess of the effect of making things implicitly contained in the answers explicit than to achieve the best possible performance for a prompt.

Overall, making the phenomena explicit within the answers seems to be beneficial for their automatic scoring. For coreferences and ellipsis, we see slight increases of .01 and .03 OWK, respectively. For the two bridging prompts, κ increases by .03 and .07. Regarding numeric terms, for the prompt VB_22c we see a decrease of κ of .03, but even the baseline does not do well here. The other prompt we annotated for numeric terms shows the highest increase of κ .17.

4.4 Error Analysis

One obvious question one might ask as a student being graded by such an automatic system is whether it is beneficial to use explicit or implicit wording to get a better grade. We therefore also compare the average number of points a model trained on the original data assigns to either an explicit or implicit answer. This can be seen as analogous to our analysis of whether human evaluators favor implicit or explicit answers, this time examining whether the automatic scoring model prefers one over the other.

Table 5 shows the results of this analysis. For coreference, results are mixed. While the overall average predicted score of the explicit testing data is slightly higher, there are also answers where the explicit version receives a lower score. For nine answers, the predicted score drops by an average of 1.1 points when they are made explicit, but for 14 answers the predicted score increases by an average of 1.75 points.

Within the ellipsis data, being more explicit is beneficial. There are four instances where the predicted score improves by one point, and none where

[2] We also ran experiments using a fastText classifier (Joulin et al., 2016), which was however unable to generalize from the small number of training examples.

Coreference	
Prompt:	During the story, the reader gets background information about Mr. Leonard. Explain the effect that background information has on Paul. Support your response with details from the story.
Original answer:	It motivated him , He knew what Mr. leonard meant and that gave him incintive to try harder.
Explicit Answer:	The background information motivated Paul , Paul knew what Mr. leonard meant and that gave Paul incintive to try harder.
Bridging Anaphora	
Prompt:	One function of the bess beetle's elytra (the hard, black wing set) is protection. What is another function of the elytra?
Original Answer:	To help make the strangulating sound .
Explicit Answer:	To help make the strangulating sound of the bess beetle .
Ellipsis	
Prompt:	Draw a conclusion based on the student's data.
Original Answer:	Based on student data, I noticed that the trial two (T2) plastics stretched longer then most plastics in trial one (T1).
Explicit Answer:	Based on student data, I noticed that the trial two (T2) plastics stretched longer then most plastics stretched in trial one (T1).
Numeric Terms	
Prompt:	Describe what the graph tells you about the relationship between the number of scoops of fertilizer and the mass of beans harvested?
Original Answer:	Well 3 is the perfect amount because 4 is too many 2 is not enough.
Explicit Answer:	Well 3 scoops is the perfect amount because 4 scoops is too many 2 scoops is not enough.

Figure 2: Exemplary original and explicit variants of answers.

it worsens.

While the instance count for the SemEval prompts is low, numeric terms and bridging seem to exhibit different trends. For the numeric prompts, the prediction only changes for three of the answers, with the predicted outcome always improving, twice from *contradictory* to *correct* and once from *partially correct* to *correct*. For bridging, the outcome changes for five of the answers, the predicted label once changing from *partially correct* to *correct*, but worsening in the remaining cases, three times from *correct* to *partially correct* and once from *partially correct* to *contradictory*.

Thus, our results suggest that it depends on the phenomenon whether making it explicit leads to a more favorable prediction of the model. While refraining from using an ellipsis or leaving out the head word of a numeric term seems beneficial, making bridging explicit does not lead to the model predicting a higher score.

5 Conclusions

We find that implicit language does occur frequently in short answer data and that the phenomena we focused our analysis on can reliably be annotated in learner answers, thus showing that such data is a promising source for implicit language in a relatively controlled setting. We will publish our set of annotated answers.

As we find that making the answers more explicit improves their automatic scoring, a next step would be to automatically resolve implicit language into explicit, to enable examining this effect on a larger scale. Subsequent analyses will also widen the experiments to include more different implicit phenomena and resolve more than one phenomenon in the same set of answers.

Acknowledgement. This work was supported by the DFG RTG 2535: *Knowledge- and Data-Based Personalization of Medicine at the Point of care.*

Setting	QWK		κ			
	Coreference	Ellipsis	Bridging		Numeric Terms	
	ASAP_8	ASAP_2	ST_31b	LF_26b2	LF_27a	VB_22c
Baseline	.20	.50	.69	.55	.42	.14
Explicit	.21	.53	.72	.62	.59	.11

Table 4: Automatic scoring results for the training and testing on the original data (baseline) compared to training and testing on answers that were made explicit.

Change in Prediction after Making Explicit	Number of Answers			
	Coreference	Ellipsis	Bridging	Numeric Terms
Better	14	4	1	3
Worse	9	0	4	0

Table 5: Analysis of how the predictions of a model trained on original prompt answers differ for the original answers and their explicit versions.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 107–115.

Rajendra Banjade, Vasile Rus, and Nobal Bikram Niraula. 2015. Using an implicit method for coreference resolution and ellipsis handling in automatic student answer assessment. In *The Twenty-Eighth International Flairs Conference*.

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Herbert H Clark. 1975. Bridging. In *Theoretical issues in natural language processing*.

Myroslava O Dzikovska, Johanna D Moore, Natalie Steinhauser, Gwendolyn Campbell, Elaine Farrow, and Charles B Callaway. 2010. Beetle ii: a system for tutoring and computational linguistics experimentation. In *Proceedings of the ACL 2010 System Demonstrations*, pages 13–18.

Myroslava O Dzikovska, Rodney D Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual embodiment challenge. In *Second Joint Conference on Lexical and Computational Semantics (* SEM): Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2019. Where's my head? definition, data set, and models for numeric fused-head identification and resolution. *Transactions of the Association for Computational Linguistics*, 7:519–535.

Nigel Fabb. 1990. The difference between english restrictive and nonrestrictive relative clauses. *Journal of linguistics*, 26(1):57–77.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Andrea Horbach, Alexis Palmer, and Magdalena Wolska. 2014. Finding a tradeoff between accuracy and rater's workload in grading clustered short answers. In *LREC*, pages 588–595. Citeseer.

Andrea Horbach and Torsten Zesch. 2019. The influence of variance in learner answers on automatic content scoring. In *Frontiers in Education*, volume 4, page 28. Frontiers.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Manfred Krifka and Renate Musan. 2012. Information structure: Overview and linguistic issues. *The expression of information structure*, pages 1–44.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011. Evaluating answers to reading comprehension questions in context: Results for german and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9.

Ruslan Mitkov. 2014. *Anaphora resolution*. Routledge.

Rodney D Nielsen, Wayne H Ward, James H Martin, and Martha Palmer. 2008. Annotating students' understanding of science concepts. In *LREC*. Citeseer.

Peter F Strawson. 1950. On referring. *Mind*, 59(235):320–344.

Ramon Ziai and Detmar Meurers. 2014. Focus annotation in reading comprehension data. In *LAW VIII*, page 159.

Ramon Ziai and Detmar Meurers. 2018. Automatic focus annotation: Bringing formal pragmatics alive in analyzing the information structure of authentic data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 117–128.