Who Controlled the Evidence? Question Answering for Disclosure Information Extraction

Hardy HARDY.HARDY@MCGILL.CA, HARDY@MIKROSKIL.AC.ID
McGill University, Universitas Mikroskil

Derek Ruths DEREK.RUTHS@MCGILL.CA
McGill University

Nicholas B King NICHOLAS.KING@MCGILL.CA
McGill University

Abstract

Conflict of interest (COI) disclosure statements provide rich information to support transparency and reduce bias in research. We introduce a novel task to identify relationships between sponsoring entities and the research studies they sponsor from the disclosure statement. This task is challenging due to the complexity of recognizing all potential relationship patterns and the hierarchical nature of identifying entities first and then extracting their relationships to the study. To overcome these challenges, in this paper, we also constructed a new annotated dataset and proposed a Question Answering-based method to recognize entities and extract relationships. Our method has demonstrated robustness in handling diverse relationship patterns, and it remains effective even when trained on a low-resource dataset.

Data and Code Availability We created a new annotated dataset for the conflict of interest relationship extraction task. We provide the dataset publicly, as well as all the source code for our Question Answering methods at https://github.com/networkdynamics/COIRelationExt.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

The role of sponsors in the production of health research is a perennial concern (Lo and Field, 2009; Sismondo, 2011; Bourgeois et al., 2010; Melander et al., 2003; Bekelman et al., 2003; Sismondo, 2008; Cherla et al., 2019; Krimsky, 2013; Angell, 2008; Krimsky, 2013; Baethge, 2008). A recent, comprehensive systematic review found that studies of medical devices and drugs that were sponsored by the manufacturer were more likely to produce favorable results and conclusions than studies sponsored by other sources (Lundh et al., 2017), which suggests that industry sponsors might have a biasing influence on the studies they sponsor. While the exact mechanism for this influence is unknown, as the systematic review notes "it seems most plausible that industry achieves overly positive results through a variety of biasing choices in the design, conduct, and reporting of their studies."

There is general agreement that certain types of relationships between sponsors, authors, and studies are more likely to generate biased results than others. For example, one study of orthopaedic research found that presentations authored by researchers who received royalties, held stock options, or were industry consultants or employees, was significantly more likely to report positive results (Okike et al., 2007). Similar concerns have been raised about the involvement of industry sponsors in the 'ghost management' of research studies (Sismondo, 2009, 2011). There is widespread agreement that "passive" involvement (e.g. funding, providing physical materials such as drugs or medical devices) is acceptable, while "active" involvement (e.g. participating in study design, analysis, writing, etc.) poses a serious risk of producing biased results (Booth and Detsky, 2019). As such, identifying the exact types of relationships between sponsoring entities and the research studies they sponsor is of considerable importance, both for understanding the potential for bias in particular studies and for examining the prevalence of potentially biasing relationships on a large scale.

In an effort to increase transparency and reduce bias, most biomedical journals now require authors

to specify the exact role of sponsors in study funding, design, conduct, analysis, and publication (Lo and Field, 2009). For example, the SPIRIT guidelines for clinical trials recommend that authors explicitly declare the "role of study sponsor and funders, if any, in study design; collection, management, analysis, and interpretation of data; writing of the report; and the decision to submit the report for publication, including whether they will have ultimate authority over any of these activities"[1].

Enabling a large-scale study on extracting previously mentioned relationships would require text processing and understanding capabilities, e.g. Graham et al. (2020b) uses pattern matching to extract common relationship patterns such as "is employed by". However, not all disclosure statements have discernible patterns, for example, a funding relationship is often declared explicitly as "Funding was provided by X". It can also come as an indirect relationship, for example, "Author X was supported by a grant from Y". Some even require a further comprehension of the text. For example, when an organization has supported specific activities, phrases such as this appear: "The sponsor did provide contextual information that assisted in the interpretation of the results." or "We thank the X Department for providing access to shelter databases.", or even "We acknowledge the efforts of X; Y, and Z in the creation of the ABC database." In short, it is intractable to use simple pattern matching to identify all these patterns.

In this paper, we introduce a new task called COIRelationExt where, given a disclosure statement in the form of an unstructured text, we seek to extract information about the relationships between the sponsoring industry and the study itself. To accompany the COIRelationExt task, we created a new dataset where we collect disclosure statements from various journals and then employ expert annotators to annotate the sponsoring relationships to the study. An example of the resulting annotations is shown in Figure 1.

The COIRelationExt task poses a number of challenges. As mentioned before, one challenge is that the disclosure statement text can vary greatly in format, with some of them resembling coded text. While there are guidelines for how disclosure information should be presented (e.g. SPIRIT[2]), not all authors adhere to them. The second challenge is the hierarchical nature of the task itself. We first have to identify the correct entity from the text and then for each entity identified in the disclosure statement, we need to extract all possible relationships, most of which are implicit in the text, that may exist between the sponsoring industries and the study itself. It is possible to train several models to address each relationship in addition to one model that can identify the entity of interest but that would be very time-consuming. The last challenge is the small number of annotated data due to the costly annotation process.

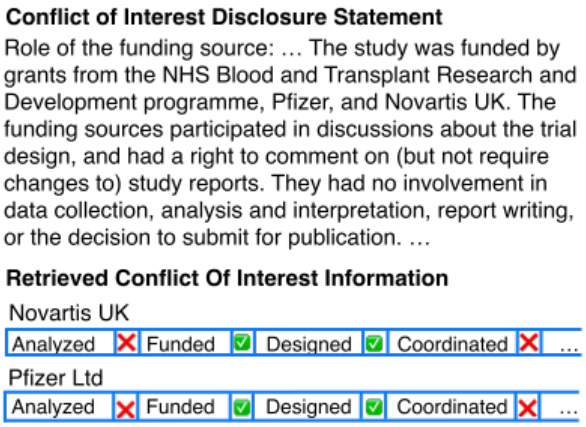

Figure 1: An example of a disclosure statement and some of its extracted COI information

To address the above challenges, we propose a novel method called Question Answering for Conflict of Interest Extraction (QA4COI) for the COIRelationExt task that is based on a Question Answering (QA) approach. QA4COI model is composed of two stages: QA for entity recognition, and QA for relationship extraction. The QA approach is an auto-regressive generative approach which is much more suitable to extract implied relationships from the disclosure statement than the pattern-matching approach. Our model also handles the lack of annotation data by utilizing the transfer learning capability that is provided by large language model (LLM) (Devlin et al., 2018; Zhang et al., 2020; Liu et al., 2019; Raffel et al., 2020). We use Flan-T5 (Chung et al., 2022) which has been fine-tuned on a large corpus including QA datasets.

1. https://www.spirit-statement.org/sponsor-and-funder/
2. https://www.spirit-statement.org/declaration-of-interests/

In summary, our contributions are threefold: (1) we introduce a new task called COIRelationExt, (2) we created a new dataset for the task, and (3) we introduce a new method (QA4COI) which is a two-stage QA approach.

2. Related Works

2.1. Automatic Disclosure Information Extraction

Studies (Okike et al., 2007; Tian et al., 2021; Graham et al., 2022) on disclosure statements have been done to analyze the extent of companies' involvement in the studies. Since manually analyzing them can be laborious and time-consuming, especially when working with larger sample sizes, there have been efforts (Graham et al., 2020a,b) to bring a machine-learning approach to these studies in order to increase the scope of the study. Graham et al. (2020b) extract author-sponsoring entities relationships using a pipeline consisting of a Named Entity Recognition (NER) model for recognizing entities, a deterministic algorithm for recognizing authors, and a pattern matching via regular expression for extracting relationships. These approaches are, however, limited to observable surface text patterns only. Similarly, Graham et al. (2020a) also uses a combination NER model and pattern-matching algorithms to extract author-sponsoring entities relationships. They, however, apply an additional entity disambiguation process to discern unique entities with the objective of building a relationship network.

2.2. Question Answering (QA)

A QA task is a type of machine comprehension task (Hermann et al., 2015) in which a model is trained to generate an answer to a question based on a given context, such as a passage of text or a set of documents. The challenge of this task is that the machine must use reasoning to find the answer based on the given context. Recent advances (Zhang et al., 2020; Lewis et al., 2019; Raffel et al., 2020) in large sequence-to-sequence (seq2seq) pre-trained language models have enabled breakthroughs on many seq2seq tasks, including the QA task. Furthermore, research (Chung et al., 2022; Ouyang et al., 2022) in scaling these tasks and fine-tuning large pre-trained language models (Brown et al., 2020; Raffel et al., 2020; Chowdhery et al., 2022) on many downstream tasks, including QA tasks, have enabled transfer learning on low-resource QA dataset.

3. COIRelationExt Dataset

The COIRelationExt dataset consists of disclosure statements and accompanying expert annotations that we collect from the year 2004 to 2014. The annotation is a set of binary labels that indicate the relationships between the sponsoring industries and the study containing the disclosure statement.

We employed four expert annotators to read the disclosure statement and extract relationships between the sponsoring entities and the study itself. At the outset, we conducted an open coding exercise to consolidate comprehensive categories of relationships. We identified the following 11 types: *analyze*, *collected data*, *coordinated*, *designated*, *funded*, *participated in*, *reviewed*, *supplied*, *supplied data*, *supported*, and *wrote*. These labels are created based on the ICMJE recommendations[3] and the AMA styles. Our expert annotators labeled all relationships in our dataset using a binary *yes* or *no* for every 11 types. In addition to identifying the authors, the annotation process aimed to discern the sponsorship entities from the text, which can be a challenging task as the name of the sponsor is often intermingled with the names of the authors and other non-relevant entities, e.g. the word *NIH* can be initials of an author, name of a grant or name of a government organization. Due to time and cost constraints, we utilize a combination of double and single annotation procedures. As a result, we are able to obtain 11,299 distinct articles that have been annotated. Of these articles, 3,449 have achieved complete agreement between both annotators. Any inconsistencies or disagreements were carefully scrutinized and addressed by the study authors, ensuring the highest level of accuracy and reliability in our findings.

We split the 11,299 articles into three sets (training, validation, and test) based on the number of authors as papers with more authors typically have longer disclosure statements. By observation, it has been noted that the level of external engagement tends to increase proportionally with the number of authors involved. We divide all our articles into two groups: articles with three or fewer authors[4], and articles with four or more authors. We then distribute

3. https://www.icmje.org/icmje-recommendations.pdf
4. Statistics on a larger unannotated dataset shows that 60.6% have three or fewer authors

Table 1: Comparison of disclosure datasets with respect to the size of training, validation, test set, and average article length (in terms of tokens).

Split set	# articles	avg # tokens
Train	7,861	221.630
Validation	1,684	228.62
Test	1,684	229.53

Table 2: Percentages of each class within the training set

Classes	% of dataset
Supply Data	1
Write	1
Collect Data	2
Participate	7
Coordinate	3
Review	3
Analyze	2
fund	62
supply	5
design	2
support	11

each group into three sets with the ratio 7:1.5:1.5 for training, validation, and test sets respectively. Finally, we obtain train/validation/test sets with the size of 7,861/1,684/1,684 articles. Table 1 shows the size of our training, validation, and test set. Further statistics of these sets can be seen in the Appendix.

Table 2 shows the percentages of each class within the training set. The distribution of class is not balanced which reflects the condition in the real world, e.g. the *fund* class occurs in the majority of samples as declaring financial COI is the norm in practice. The *write*, *supply data*, *collect data*, *analyze*, and *design* classes, on the other hand, are very rare in our training set.

4. QA4COI: Question Answering (QA) for Conflict-of-Interest Extraction

In this section, we explain the Question Answering for Conflict-of-Interest Extraction (QA4COI) model. First, we formally define the general task of QA and our model. Given a question and its context, we first concatenate both of them into a single text, T, of length n and then embed it as $\mathbf{X}_{1:n}$. We use Flan-T5 (Chung et al., 2022) that is based on an encoder-decoder Transformers (Vaswani et al., 2017) to map this latent input into an output $\mathbf{Y}_{1:m}$ which serves as the answer to the question. The encoder part of Transformers, $f_{\theta_{enc}}$, has n layers (denoted L_i^e for layer i), that transform this input sequentially: $X^{L^e+1} = L_i^e(X^L)$. The decoder part of Transformers which has the same stack as the encoder models the conditional probability of the target output denoted by $p_{\theta_{dec}}(\mathbf{Y_{1:m}}|\mathbf{X_{1:n}})$. Further details of Transformers can be found in the work of Vaswani et al. (2017).

QA4COI is composed of two similar QA models that work sequentially. The first model is the QA model for Entity Recognition and the second model is the QA model for Relationship Information Extraction. Figure 2 shows the whole pipeline of our model.

4.1. QA for Entity Recognition

The first stage of QA4COI discovers entities of interest in the disclosure statement. Our entities of interest are companies or organizations that sponsor the study. In Figure 1, the entities of interest are *Novatis UK* and *Pfizer Ltd.*

Instead of fine-tuning on existing Named Entity Recognition (NER) model, which may have pre-trained labels that are different from our dataset, we built a QA model that extracted the entities of interest by treating the task of NER as machine comprehension. Specifically, we pose the input as a question "What organizations are involved in the study? context: {context}" where the context is the disclosure statement.

For our model, we use pre-trained Flan-T5 (Chung et al., 2022) which is an extended T5 model (Raffel et al., 2020) that is further fine-tuned on hundreds of downstream tasks including QA tasks. We seek to leverage the knowledge gained from similar tasks to compensate for our limited training dataset.

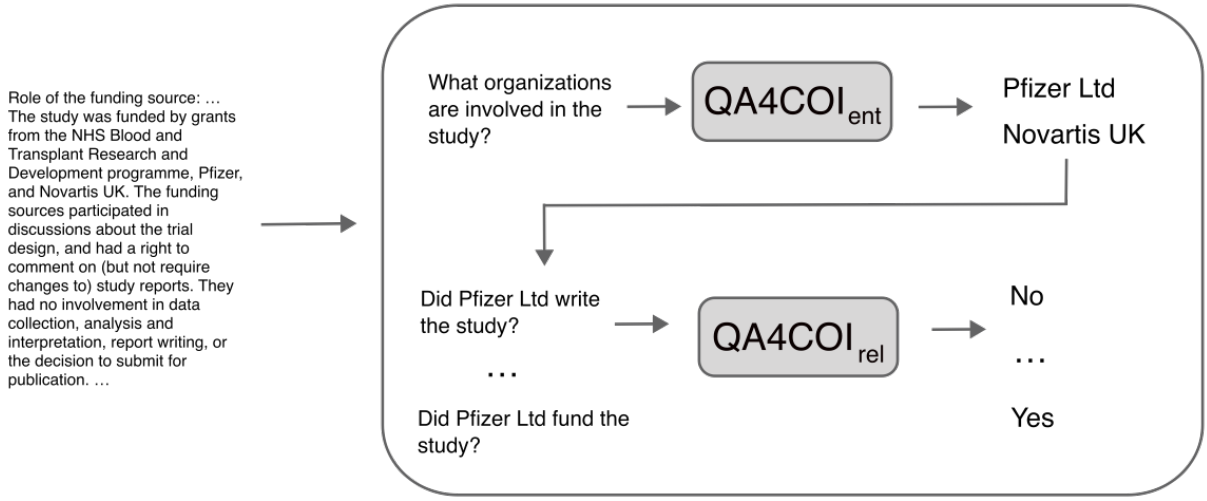

Figure 2: In the QA4COI pipeline, the $\text{QA4COI}_{\text{ent}}$ model first extracts entities from the disclosure statement using a Question Answering format and then the $\text{QA4COI}_{\text{rel}}$ extracts 11 entity-study relationship types from the same statement using predicted entities

For the baseline, we use an off-the-shelf NER model called Spacy[5]. Our aim was to assess how well typical NER systems would perform on our dataset. Since Spacy model would label all entity types, we filter out all entities that are not of the *ORG* (organization) type.

Given that our model is a generative one: producing text as output, whereas the baseline is a classifier that outputs a set of tags, we have to use an evaluation approach that is suitable for both types of models. We first convert the output of the NER model into a list of entities to match the output of the QA model. We then calculate the F_1, *precision* and *recall* scores. The numerator (*COR*) of our *precision* and *recall* is the number of match entities in the strict setting, i.e: exactly the same tokens, which is shown in the following equations:

$$\text{Precision}_{\text{ent}} = \frac{\text{COR}}{\text{TP} + \text{FP}} \quad (1)$$

$$\text{Recall}_{\text{ent}} = \frac{\text{COR}}{\text{TP} + \text{FN}} \quad (2)$$

We use the strict setting as there are many entities that share similar tokens, for example: *National Institutes of Health* and *National Institute of Aging*. While the strict setting reduces false positives, it increases the false negative as well if the model misses even one token. To reduce this, we normalize all tokens before evaluation by removing the English article words such as *the*.

4.2. QA for Relationship Information Extraction

The second stage of QA4COI is to identify the relationships between recognized entities of interest from the previous stage with the study itself. In total, we seek to identify 11 binary relationships. Identifying these relationships comes with varying difficulties as previously mentioned in the introduction sections. Moreover, a typical classification model would require us to build 11 models for each type of relationship (which is impossible as well as some classes are very rare).

Our proposed QA model can solve the above issue as the model relies on text comprehension instead of pattern matching. Moreover, the model is robust to the huge variability of the text syntactic variants, even including those that have grammatical or spelling mistakes. Using the extracted entities of interest from the previous stage, we pose the problem as a QA form, for example: "question: did

5. https://spacy.io/

{entity name} write the study? context: {context}" with the context being the same disclosure statement. The answer to that question will be a binary label of *yes* and *no*. Finally, the QA format enables us to build a single model for all types of relationships.

Since we are the first ones to approach this task, there are no comparable baselines in the literature. Therefore, our method serves as the strong baseline for future work on the QA4COI task. We, however, created two dummy baselines for qualitative assessment purposes. The first one is the random baseline where we randomly assign *yes* or *no* class to each of the 11 relationships for all the oracle entities found in the text. The second one is the text matching baseline, where we assign *yes* if the class name occurs in the text, and assign *no* if otherwise.

For evaluation, we measure the coarse classes: *yes* and *no*, and also the fine classes which are the 11 relationship classes. We use F_1, *precision*, and *recall* scores for each coarse and fine class.

We add a *blank* class to the coarse classes to accommodate missing entities (false negative) that the previous stage (QA entity recognition) misses. For example, the previous stage produces only two entities along with their relationships while the annotation has three entities recognized. The missing entity and all its relationships will be assigned as *blank* class to differentiate it from *no* class. This way we don't inflate the result of the *no* class due to the previous stage error.

5. Experiments and Results

In this section, we explain our experiments and results.

5.1. Setup and Preprocessing

All of our experiments are run on a single GPU GeForce RTX 3090 with 24 GB RAM. Due to limited computing power and memory size, we can only use the base model of Flan-T5[6]. We tune the hyperparameter using the validation dataset and tested the final model on the test dataset.

5.2. QA for Entity Recognition

We convert our annotated dataset, which is presented in table format as shown in Figure 1, into a question-answering format. Each disclosure statement is used as the context for the question, and the annotated industry entities are used as the answer, separated by a specific symbol. We train our model using the Flan-T5 base model, running the training process three times with different random seeds. We reported the mean and std deviation results for our model in Table 3. For Spacy NER we only run one time as we don't train the model and only use it for inference only.

We show that the F_1, precision, and recall scores of our model ($\text{QA4COI}_{\text{ent}}$) are significantly higher than the baseline Spacy NER in both the precision and recall scores. Our model also has low variance as shown by the insignificantly small standard deviations.

Table 3: F_1, precision and recall scores on test and validation splits for our model ($\text{QA4COI}_{\text{ent}}$) and the baseline Spacy NER

Model	F1	Precision	Recall
Spacy NER	0.59	0.51	0.71
$\text{QA4COI}_{\text{ent}}$	$0.82_{\pm 0.00}$	$0.85_{\pm 0.00}$	$0.79_{\pm 0.00}$

To understand the results difference in Table 3, we look into the performance of both models in different sizes of entities. We first calculate the average length of annotated entities in each article and then we distribute each article to non-uniform bins accordingly. In each bin, we run both models and evaluate them in isolation. Figure 3 shows the histogram of each bin with respect to the F_1 score of each model. Investigating Figure 3 closely shows two trends. The first one is that Spacy NER performs badly on short entities and on long entities. The second one is that our model performs almost uniformly on different lengths of entities. An example output comparison between our model and the baseline can be seen in the Appendix.

In the case of short entities, Spacy NER often mislabeled part of the names, e.g. the tokens *Novartis, UK* and *Novartis UK* should be classified as the same type, meanwhile, for the Spacy NER model the first one is classified as *organization* and *country*, while the second one is classified as *organization*. Another issue is that the disclosure statement has many potentially false positive *organization* entities such as the name of the grant, the author's name that is shortened, and others.

6. https://huggingface.co/google/flan-t5-base

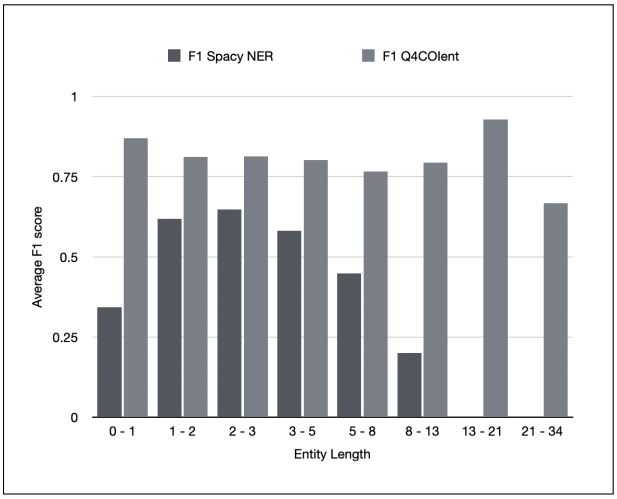

Figure 3: A histogram depicted the F_1 score for different length of entities for Spacy NER and QA4COI predictions

In the case of long entities, Spacy NER often breaks them into two or more entities, e.g. *Claude D. Pepper Older Americans Independence Center at Yale University School of Medicine* which should be tagged as a single *organization* entity but Spacy tagged as two separate entities: *Claude D. Pepper Older Americans Independence Center* and *Yale University School of Medicine*. All of these issues lower the F_1 scores and would certainly cascade errors into the next stage. Our QA model is, however, robust to those issues.

5.3. QA for Relationship Information Extraction

Given the identified entities of interest in a disclosure statement, we form 11 questions for each one and group all of them in the dataset for training. We also trained an oracle version where we use the gold standard entities to form the question. The oracle version is created to measure the performance of this stage without being penalized by the performance from the previous stage.

Table 4 shows all model results of the *yes* and *no* coarse classes. In the table, we see that the *no* class has almost a perfect result in the oracle version. However, this is due to the largely imbalanced class samples. For the *yes* class result, our model obtains 76% F_1 score but there is still a lot of room for improvement in the entity recognition stage as the oracle version obtains 85% F_1 score. The text-matching model performs much better than the random model which means that there are observable indicators in the surface text albeit very minimal.

Table 4: Models' performance on Relationship Information Extraction on three metrics (F_1, precision and recall)

Model	F1	Precision	Recall
Yes Class			
Random	$0.13_{\pm 0.01}$	$0.07_{\pm 0.01}$	$0.49_{\pm 0.01}$
Text-Matching	0.29	0.19	0.62
QA4COI	$0.76_{\pm 0.01}$	$0.87_{\pm 0.01}$	$0.68_{\pm 0.01}$
Oracle	$0.85_{\pm 0.01}$	$0.85_{\pm 0.00}$	$0.85_{\pm 0.01}$
No Class			
Random	$0.65_{\pm 0.01}$	$0.92_{\pm 0.00}$	$0.50_{\pm 0.01}$
Text-Matching	0.86	0.96	0.78
QA4COI	$0.86_{\pm 0.00}$	$0.99_{\pm 0.00}$	$0.76_{\pm 0.00}$
Oracle	$0.99_{\pm 0.00}$	$0.99_{\pm 0.00}$	$0.99_{\pm 0.00}$

Figure 4 depicts all models' performance on the fine-grained classes. The random model reflects the number of samples in each class, the higher the number of samples the closer the number to 0.5 (as shown by the *fund* class which has 0.5 F_1 score). The correlation between QA4COI and the random model is 63.66%, which means that the number of samples in the training data (see Table 2) affects the performance of QA4COI. The text-matching model has better performance than the random model and scores highly on *fund* class. This means that there are observable surface indicators that can help the model in making a good prediction. In many classes that don't have many surface indicators as shown by the text-matching model results, our model still shows over 0.5 F_1 score. This means that our model successfully captures the implied meaning from the text.

These results point to two major issues that pose a challenge to our method (and likely will pose a challenge to any approach):

Few samples, many forms. The *supply data*, *write*, *collect data*, *analyze*, *coordinate*, *review*, and *design* classes' samples are lower than 5% percent in the dataset. This, in turn, affects the performance of the QA4COI model. With fewer samples, the model doesn't have exposure to many variants of the word-

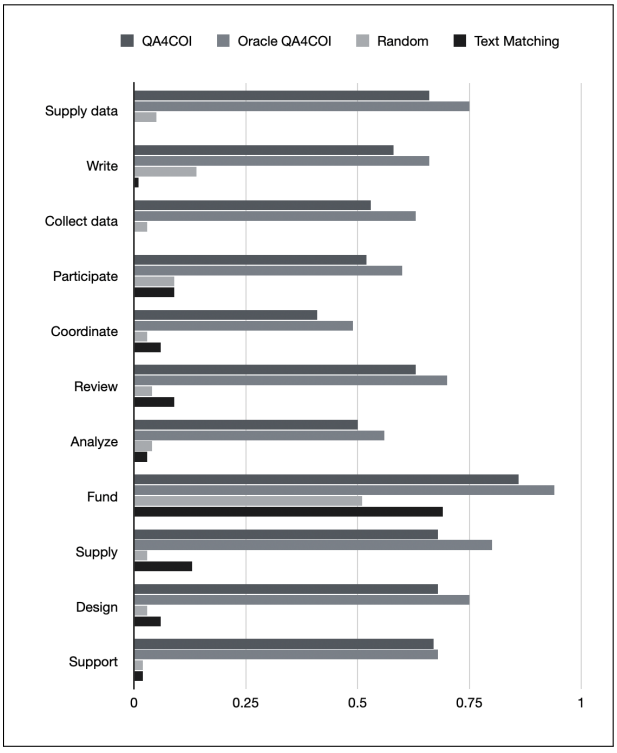

Figure 4: A histogram bar chart depicted the F_1 score for each fine-grained class

ing. For example, the word *analyze* occurs in some texts, e.g. "Author A was supported by grant Z from X, where the study was conceived, coordinated, and analyzed", but it is a rare case. The language of *analyze* can come in many different forms, e.g. "interpreting", "studying", and "analyzing". Classes with a high number of samples, e.g. *fund* gain the highest performance among others.

Relationships that pertain to name entities/actions/resources. The *supply data* and *collect data* classes are difficult to capture due to requisite specificity in the text for proper extraction. This specificity could pertain to specific resources to be collected or actions to be taken, which further exacerbates the complexity of the task. As a result of these inherent difficulties, text-matching models failed to attain any F_1 score in this instance.

Despite these issues, we observed that our QA4COI model was able to overcome the challenges associated with extracting author-study relationship information from the disclosure statement. It demonstrated robustness in handling various wording variants, despite being trained on a low-resource dataset. Additionally, our model was capable of capturing implicit meanings from the unstructured text, further showcasing its proficiency in this task.

6. Conclusion

In this paper, we presented a new task for extracting conflict of interest relationships called COIRelationExt. In order to approach this problem, we contributed a new dataset for the task and presented a novel method (QA4COI). Our method's performance establishes it both as a viable solution for this problem and a strong baseline against which future methods can be measured. More broadly, given the importance of transparency and accountability within research, our hope is that this work lays the foundation for progress on this important task.

Acknowledgements

We would like to acknowledge the work of Matthew Gittings and Sarah Berry in the preparation of the original data from which the dataset used in this paper was draw. This work was funded by a grant from the Government of Canada New Frontiers in Research Fund.

References

Marcia Angell. Industry-sponsored clinical research: a broken system. *Jama*, 300(9):1069–1071, 2008.

Christopher Baethge. Transparent texts: authors of scientific articles often have conflicts of interest. it is important for these to be communicated transparently to the readers. *Deutsches Arzteblatt International*, 105(40):675, 2008.

Justin E Bekelman, Yan Li, and Cary P Gross. Scope and impact of financial conflicts of interest in biomedical research: a systematic review. *Jama*, 289(4):454–465, 2003.

Christopher M Booth and Allan S Detsky. From the $80 hamburger to managing conflicts of interest with the pharmaceutical industry. *Bmj*, 365, 2019.

Florence T Bourgeois, Srinivas Murthy, and Kenneth D Mandl. Outcome reporting among drug

trials registered in clinicaltrials. gov. *Annals of internal medicine*, 153(3):158–166, 2010.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Deepa V Cherla, Cristina P Viso, Julie L Holihan, Karla Bernardi, Maya L Moses, Krislynn M Mueck, Oscar A Olavarria, Juan R Flores-Gonzalez, Courtney J Balentine, Tien C Ko, et al. The effect of financial conflict of interest, disclosure status, and relevance on medical research from the united states. *Journal of general internal medicine*, 34(3): 429–434, 2019.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

S Scott Graham, Zoltan P Majdik, and Dave Clark. Methods for extracting relational data from unstructured texts prior to network visualization in humanities research. *Journal of Open Humanities Data*, 6(1), 2020a.

S Scott Graham, Zoltan P Majdik, Dave Clark, Molly M Kessler, and Tristin Brynn Hooker. Relationships among commercial practices and author conflicts of interest in biomedical publishing. *Plos one*, 15(7):e0236166, 2020b.

S Scott Graham, Nandini Sharma, Martha S Karnes, Zoltan P Majdik, Joshua B Barbour, and Justin F Rousseau. A content analysis of self-reported financial relationships in biomedical research. *AJOB Empirical Bioethics*, pages 1–8, 2022.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28, 2015.

Sheldon Krimsky. Do financial conflicts of interest bias research? an inquiry into the "funding effect" hypothesis. *Science, Technology, & Human Values*, 38(4):566–587, 2013.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

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

Bernard Lo and Marilyn J Field. *Conflict of Interest in Medical Research, Education, and Practice*. National Academies Press (US), 2009.

Andreas Lundh, Joel Lexchin, Barbara Mintzes, Jeppe B Schroll, and Lisa Bero. Industry sponsorship and research outcome. *Cochrane database of systematic reviews*, (2), 2017.

Hans Melander, Jane Ahlqvist-Rastad, Gertie Meijer, and Björn Beermann. Evidence b (i) ased medicine—selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *Bmj*, 326(7400):1171–1173, 2003.

Kanu Okike, Mininder S Kocher, Charles T Mehlman, and Mohit Bhandari. Conflict of interest in orthopaedic research: an association between findings and funding in scientific presentations. *JBJS*, 89(3):608–613, 2007.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140): 1–67, 2020.

Sergio Sismondo. Pharmaceutical company funding and its consequences: a qualitative systematic review. *Contemporary clinical trials*, 29(2):109–113, 2008.

Sergio Sismondo. Ghosts in the machine: publication planning in the medical sciences. *Social Studies of Science*, 39(2):171–198, 2009.

Sergio Sismondo. Corporate disguises in medical science: Dodging the interest repertoire. *Bulletin of Science, Technology & Society*, 31(6):482–492, 2011.

Tina Tian, Anand Y Shah, Jeremy Darling, Charlie Zaepfel, Abhishek Chatterjee, Mark Iafrati, and Payam Salehi. Assessment of self-reported financial conflicts of interest in vascular surgery studies. *Journal of Vascular Surgery*, 74(6):2047–2053, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

Appendix A. Dataset Composition

Table 5: The composition of each split with regard to the number of authors per article. The average, maximum, and minimum number of authors per article are represented by the mean, max, and min, respectively.

Split set	<= 3	>3	Mean	Max	Min
Train	774	7087	11.08	1941	1
Test	165	1519	11.63	460	1
Valid	166	1518	10.81	440	1

Appendix B. Output Example

Table 6: An example output between Spacy and Flan T5 models. Each identified entity is separated by the '—' symbol

Context	From San Francisco Veterans Affairs Medical Center and University of California, San Francisco, San Francisco, California. Grant Support: Dr. Walter is a recipient of the Veterans Administration Research Career Development Award from the Division of Health Services Research and Development. Dr. Covinsky was supported in part by an independent investigator award from the Agency for Healthcare Research and Quality (K02HS000006-02) and is a Paul Beeson Faculty Scholar in Aging Research. Potential Financial Conflicts of Interest: None disclosed.
Gold Annotation	Paul Beeson Faculty Scholar in Aging Research — Veterans Administration Research Career Development Award from the Division of Health Services Research and Development — independent investigator award from the Agency for Healthcare Research and Quality
Spacy NER	the Division of Health Services Research and Development — University of California, — San Francisco Veterans Affairs Medical Center—the Agency for Healthcare Research and Quality
$QA4COI_{ent}$	Agency for Healthcare Research and Quality — Paul Beeson Faculty Scholar in Aging Research — Veterans Administration Research Career Development Award from the Division of Health Services Research and Development