Conflict of Interest Networks

Wing Tung (Angie) Fong
School of Information
University of Texas at Austin
angie.fong@utexas.edu

Hyojeong kim
School of Information
University of Texas at Austin
hkim92@utexas.edu

Peggy Fidelman
Department of Computer Sciences
University of Texas at Austin
peggyf@cs.utexas.edu

ABSTRACT

Transparency around conflicts of interest (COI) among re-searchers is important for maintaining the integrity of science, as well as public trust in scientific endeavors. COI disclosures are a required part of most publications, but understanding the connections between these disclosure statements and re-searchers' findings can be challenging given the enormous number of scientific publications. Here, we seek to make progress toward automatic understanding of COI statements so that network models may be built in the future to aid in the understanding of COIs' effects on scientific findings.

Keywords

natural language processing, conflict of interest, machine learning, ontology, co-clustering

1. INTRODUCTION

Conflicts of interest (COI) have been part of human beings' daily life throughout history, and are a concern in industries that affect the common good, such as law and healthcare. The definition of COI is a person or entity that has two relationships competing with each other for the person's loyalty. COI has become a hotly debated concept in the scientific literature during the last few decades in medicine too. Here, we use Stanza to access the Java Stanford CoreNLP software from Python, allowing us to apply OpenIE to extract the types of relationships that appear in COI statements submitted by authors in publications found in PubMed.

Our goal is twofold: first, to investigate whether OpenIE re-lation extraction can perform better than other NLP and non-NLP techniques used previously to automatically detect COI relations; and second, to discover new categories of relation-ships that may not be covered by an existing ontology. This study provides better performance than any NLP approach currently attempted so far. This information will be used to in-form network models to study the impact of financial conflicts on biomedical research in medicines.

ACM ISBN N/A.

DOI: N/A

2. BACKGROUND

Prior to our work, spaCy was used for entity extraction from the COI statements, and performed quite well at this after being pretrained with COI statements (which allowed it to recognize the structure of COI statements more effectively) (Graham, Majdik, Barbour & Rousseau, 2022). However, spaCy's performance at relation extraction was poor. Better performance was obtained using a set of hand-designed regular expressions, iterating exhaustively over the COI statements to find which regular expressions fit each string, and classifying the type of ICMJE conflict category expressed in the COI statement based on which regular expression(s) were found in the string. The hand-designed ontology used for this can be seen in Table 1.

3. METHOD

There are four major steps in this project, including triples extraction, data exploration of extracted relations, embedding, and clustering.

3.1 Data

Starting with 122,000 conflict of interest statements from PubMed database, we obtained and processed 40k triples of subject, relation and object from the corpus. The details are described below.

3.2 Triples Extraction

With the CoreNLP Client from the Stanza package, we man-aged to extract 40k triples of conflict of interest relationships. First, we installed a Java environment and the Stanza Package in a google colab notebook. Then, we filtered the corpus to get statements that has more than 160 characters to eliminate meaningless input. To work around the timeout constraint, we divided the corpus into a list of 5 parts, annotated them separately and assembled the output in a dataframe.

3.3 Data Exploration of Extracted Relations

OpenIE extracts all possible relation triples from the raw text of the COI statements. As an example, for the statement "X is an employee of Y", OpenIE will find the triples {X, is, an employee of Y} as well as {X, is an employee of, Y}. It would be useful to discover ways to automatically choose the triple with the most information content from among the set of all triples extracted for a given sentence; specifically, for the example above, we would like to eliminate {X, is, an employee of Y} and retain {X, is an employee of, Y}. This was the motivation for the data exploration described in this subsection.

ICMJE category	COI terms
1. All support for the present manuscript	fund, support, collaborat, research, trials, investigator
2. Grants or contracts from any entity	contract, grant
3. Royalties or licenses	royalties, license, intellectual property
4. Consulting fees	consul
5. Payment for speaking, writing, educational events	honrari, speak, talk, edit, teaching, lecture, conference, educational
6. Payment for expert testimony	witness
7. Support for attending meetings and/or travel	travel, accomodation
8. Patents planned, issued or pending	patent, inventor
9. Participation on a Data Safety or Advisory Board	advi, member, panel, board, proctor
10. Leadership or fiduciary role in other board or society	chair
11. Stock or stock options	shareholder, equity, financial interest in, stock
12. Receipt of materials, equipment, gifts, services	(none)
13. Other financial or non-financial interests	fee, remuneration, participation, manuscript-fee, reviewing
Unknown ICMJE category	own, proprietary, manager, employ, CEO, CFO, president, founde, Officer, trustee, director, partners, hired, award, associates, leader

Table 1. COI terms in the human-designed ontology. (Note that these are not always full words, but sometimes substrings designed to capture multiple forms of a word.)

Starting with the full set of relation triples extracted by Ope-nIE (exported to a 3-column csv file with the middle column containing the extracted relation in each case), we used Pan-das in an IPython notebook hosted on Google Colaboratory to find the number of unique relation strings and the number of occurrences of each. Following this, we sought patterns in the data that might help with future automatic selection of the most useful relation triple extracted from each COI sentence.

The most common relation strings tend to be among the least informative (see Figure 1). Similarly, the shortest relation strings also tend to be the least informative (see Figure 2). Relation strings shorter than three characters can likely be safely eliminated from the set of extracted relations without losing useful information. Even among the relation strings that are of length three, very few (e.g. "ran", "run", "own") are of use.

3.4 Analysis of Extracted Relations using String Similarity

To determine which extracted relation strings correspond to the known COI types defined in the existing ontology, a way to measure similarity of relation strings is necessary. A simi-larity measure is also necessary for any type of clustering that could be used to find significant groups of relations not cap-tured by the current ontology. Before moving on to semantic embeddings as a basis for judging similarity, we explored the possibility of using a simpler measure such as string similarity.

Initially, we applied an open-source fuzzy string matching package, thefuzz, on the full set of extracted relation strings. This package includes several different methods of fuzzy string matching which all include a confidence measure. It can also choose the top one or more matches to an input string from among a set of reference strings. We found that the latter was the most useful, because it is possible to provide the known set of COI terms as the set of reference strings, and then apply thefuzz's extractOne() method to all extracted relation strings to choose the closest match among COI terms for each extracted relation (along with a confidence value).

This was effective for relation strings that already contained similar substrings to one of the COI terms, but tended to give unreliable results for the rest, sometimes with a very high confidence. (As one example, for the relation string "be", it reported "member" as the match with 90% confidence.) The one advantage this conferred over simply searching for the COI terms as substrings in the extracted relation strings is that it was able to correct for small typographical errors or letter omissions (for example, "received honoraria for" was successfully matched with "honrari").

Based on this, we concluded that some semantic knowledge is necessary to get a meaningful measure of similarity and thus meaningful results when clustering the extracted relation strings.

3.5 Embedding

Knowledge graph embeddings can provide insights about relations among the entities in a knowledge graph. It is typically used to for missing link prediction, knowledge graph discovery, and entity clustering. In our case, we just want to include the semantic meanings of the relationships in the embeddings and see if it affects the result of clustering. We have used the Ampligraph library, a library built on Tensorflow with GPU support, for this task. After consolidating the triples, we trained the embeddings with the ComplEx and TransE model.

3.6 Clustering

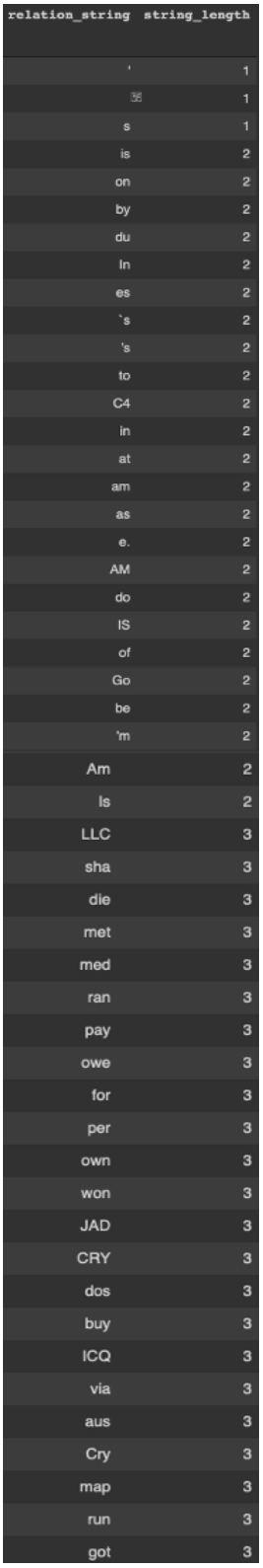
With the embeddings of the relations in the triples, we used the Ampligraph's find_clutser function with various clustering algorithms to divide the relations into clusters.

3.6.1 K-Means

With k set to 21, we are trying to mimic the ICMJE categories which has 21 human-generated categories of conflict of interest relationships. On a 2D plane, we can see that the complEx embeddings has a clearer border compared to the TransE embeddings. The results are as follow.

Relation	count
is	38086
are	32894
reports	23743
was	23544
were in	23261
has	16703
has received	15107
is in	14732
received	12915
is with	11518
were	10330
have	9697
are employees of	8140
have completed	7920
was obtained from	5927
was approved by	3926
was supported by	3650
receives	3216
is member of	2958
is employee of	2739
declare	2641
have read	2614
fees from	2592
has declared	2541
grants from	2525
is on	2401
had	2349

Figure 1. Top relation strings by frequency of occurrence. Relation strings that may contain enough information to determine ICMJE COI category are marked with a green arrow.



relation_string	string_length
'	1
[illegible]	1
s	1
is	2
on	2
by	2
du	2
In	2
es	2
`s	2
's	2
to	2
C4	2
in	2
at	2
am	2
as	2
e.	2
AM	2
do	2
IS	2
of	2
Go	2
be	2
'm	2
Am	2
Is	2
LLC	3
sha	3
die	3
met	3
med	3
ran	3
pay	3
owe	3
for	3
per	3
own	3
won	3
JAD	3
CRY	3
dos	3
buy	3
ICQ	3
via	3
aus	3
Cry	3
map	3
run	3
got	3

Figure 2. Shortest relation strings.

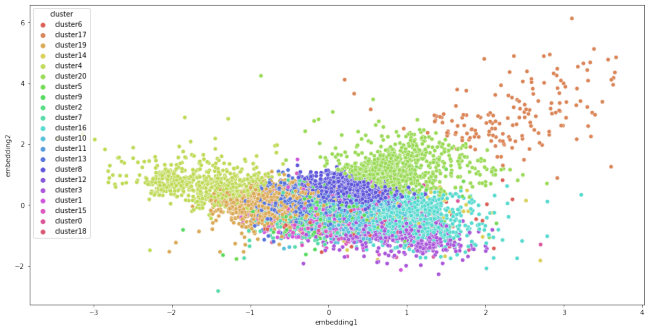


Figure 3. K-Means clustering with complEx embeddings.

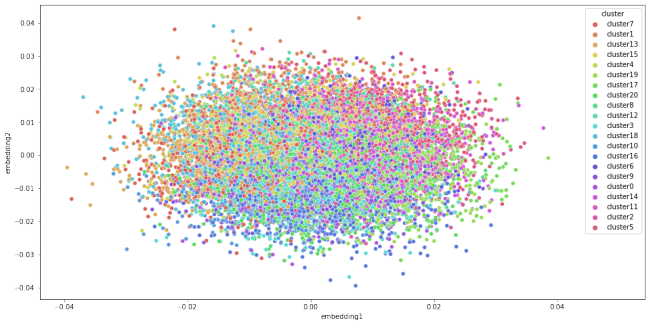


Figure 4. K-Means clustering with TransE embeddings.

3.6.2 DBSCAN

We have also tried DBSCAN to see the how the relations will be clustered naturally without specifying the number of cluster. However, the result is not so meaningful at this stage as the most number of clusters we get is just 3 on the ComplEx embeddings and no cluster on the TransE embeddings.

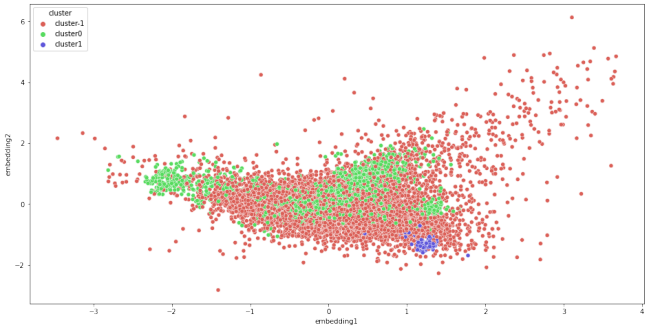


Figure 5. DBSCAN clustering with complEx embeddings.

3.7 Mapping the clusters to the ICMJE categories

As the clustering outputs from K-Means are more meaning-ful, we have chosen to map them to the ICMJE categories. After counting the number of occurrence of ICMJE category-defining-terms in each cluster, we discovered that some cat-egories appeared in some clusters a lot more frequently. For example, for category-defining-terms of IP-patent, we saw 74 counts in ComplEx cluster 13 (see Figure 7) and 25 counts in TransE cluster 7 (see Figure 8). Other categories like Owner-ship, Employment and Advisory shows similar pattern. The highlighted clusters in these categories can be a selected subset

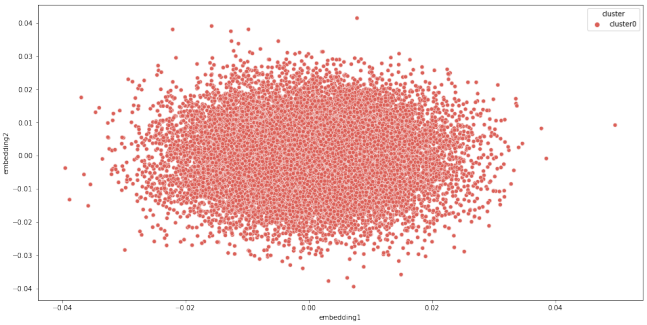


Figure 6. DBSCAN clustering with TransE embeddings.

for human-annotation that could bring a larger impact through downstream classification tasks.

We also discovered that some categories have significantly more occurrence in all the clusters and some categories have no occurrence in all the clusters. Those with no occurrences can be the focus of further research on new undiscovered categories.

4. RELATED WORK

Traditional methods on IE have focused on the use of su-pervised learning techniques such as MUC evaluation meth-ods (Meystre, Stéphane M., et al.2008), Force-directed algo-rithms(Graham, S. Scott, et al. 2021), and self - supervised learning (Santilli, Alice ML, et al. 2021). These methods were a set of rules hand-tagged training documents and learned in this manner are effective on documents similar to the set of training documents. However, Metrics perform excessively well when an object has multiple mentions, including one that is not related to the object and the force-directed algorithms required high running time. While most IE work has focused on a few relations in specific pre-selected domains, certain corpora are unlikely to be amenable to extracting a certain min-imum string length. In addition, IE requires pre-specifying a set of relations of interest and then providing training examples for each. Therefore, it was not possible to present a filtering system that automatically selects the correlation between the most useful information and the least important information. Furthermore, while existing studies have enabled generation of string sets that capture COI of ICMJE categories(Graham, S. Scott, et al. 2021), IE has not yet been attempted to ex-tract new types that have not been captured by man-generated ontologies.

A soft-co-clustering approach tried to allow a phrase to be part of multiple clusters to solve the problem of interpreting predicate phrases, depending on the context of associated sub-jects and objects (Pal, Koninika, Vinh Thinh Ho, and Gerhard Weikum.2020). These methods have certain drawbacks for ambiguous surface forms that would map to different word senses.

5. CONCLUSION

NLP is a powerful tool, but it needs lots of human attention and tuning to work reliably on a given type of data. Less tech-nologically advanced methods such as regular expressions,

cluster	stock	ownership	IP-patent	employment-management	employment	IP-royalties	research	consulting fees	advisory	speaking fees	editorship	travel fees	fee-unspecified	supplies	non-financial	expert witness	educational	fee-other	award	endowed chair	IP-other
0	0	0	0	0	0	0	4	0	41	0	0	0	2	0	0	0	0	0	0	0	0
1	2	13	33	1	7	3	132	0	7	7	1	4	20	0	0	4	5	0	1	4	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	6	4	14	11	8	1	60	0	2	3	0	6	12	0	0	0	0	0	0	5	0
4	0	0	6	5	3	1	41	0	7	2	0	2	7	0	0	0	0	0	3	3	0
5	0	0	1	1	0	0	32	0	0	3	0	2	5	0	0	0	2	0	1	0	0
6	1	0	0	0	5	0	40	0	7	0	0	0	13	0	0	0	0	0	1	1	0
7	5	3	25	5	5	7	164	0	13	15	0	9	33	0	0	1	0	0	4	2	0
8	29	11	18	16	31	11	404	0	46	37	0	46	92	0	0	1	9	0	13	9	0
9	4	4	19	12	8	4	212	0	13	14	1	6	43	0	0	4	0	0	5	1	0
10	14	11	37	34	16	12	234	0	42	21	2	18	43	0	0	7	2	0	10	12	0
11	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	28	16	74	19	14	17	436	0	59	33	4	18	68	0	0	4	10	0	12	7	0
14	0	0	1	0	0	0	37	0	0	0	0	8	19	0	0	0	0	0	0	1	0
15	2	1	5	2	0	0	53	0	1	0	0	1	9	0	0	0	0	0	1	3	0
16	17	1	22	31	11	4	148	0	21	19	0	12	38	0	0	3	0	0	5	10	0
17	5	1	0	0	5	0	61	0	2	5	0	10	18	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	13	14	35	13	6	14	262	0	30	27	0	25	57	0	0	9	6	0	10	9	0
20	28	7	4	2	10	3	259	0	23	32	0	30	97	0	0	0	0	0	2	1	0

Figure 7. ICMJE terms count in ComplEx Clusters.

cluster	stock	ownership	IP-patent	employment-management	employment	IP-royalties	research	consulting fees	advisory	speaking fees	editorship	travel fees	fee-unspecified	supplies	non-financial	expert witness	educational	fee-other	award	endowed chair	IP-other
0	10	7	12	6	4	3	139	0	17	16	1	8	24	0	0	2	4	1	3	3	0
1	2	1	13	10	4	4	117	0	14	10	1	6	29	0	0	1	1	0	1	6	0
2	1	3	9	2	3	3	105	0	6	7	0	10	23	0	0	1	2	0	1	3	0
3	13	2	12	11	3	4	98	0	15	14	0	11	37	0	0	2	0	0	5	5	0
4	8	4	13	7	5	1	118	0	21	10	1	12	28	0	0	1	1	0	0	6	0
5	6	0	16	3	8	4	108	0	8	12	0	8	16	0	0	0	0	0	5	2	0
6	5	4	21	9	9	4	123	0	16	8	0	14	30	0	0	2	2	0	6	1	0
7	9	7	25	5	6	3	140	0	13	12	0	7	38	0	0	2	1	0	3	4	0
8	9	2	16	8	4	2	96	0	16	11	1	7	18	0	0	2	0	0	1	3	0
9	8	5	14	10	7	5	113	0	14	9	0	8	28	0	0	2	2	0	3	2	0
10	14	5	12	11	4	4	151	0	12	9	0	13	34	0	0	3	0	0	4	1	0
11	5	5	13	9	9	6	123	0	17	11	1	13	21	0	0	2	3	0	5	4	0
12	9	3	16	6	7	4	104	0	12	12	0	10	22	0	0	0	2	0	1	0	0
13	7	9	16	10	12	2	151	0	28	8	0	13	27	0	0	3	3	0	2	2	0
14	7	5	8	8	9	2	116	0	12	6	0	2	31	0	0	0	1	0	5	4	0
15	7	4	14	5	6	3	114	0	11	10	1	12	26	0	0	3	1	0	6	4	0
16	12	5	15	3	3	4	130	0	17	11	0	6	27	0	0	1	3	0	4	2	0
17	5	3	9	6	10	3	113	0	10	8	1	4	29	0	0	1	2	0	2	3	0
18	8	2	9	11	1	4	143	0	22	10	0	11	29	0	0	2	1	0	3	5	0
19	3	5	17	7	5	6	131	0	11	15	0	15	29	0	0	2	4	0	3	3	0
20	6	4	19	5	10	6	142	0	21	10	1	8	30	0	0	1	1	0	4	6	0

Figure 8. ICMJE terms count in TransE Clusters.

substring matching, and data filtering will be an important part of the pipeline for an effective application of NLP to automatic extraction of entities and relations from COI statements. How-ever, with more time and attention, we believe it is possible to build upon the work we present here to develop a pipeline that involves both spaCy entity extraction and OpenIE relation extraction to reduce the load on human annotators. In the section that follows, we give specific recommendations for next steps toward that goal.

6. DISCUSSION AND FUTURE WORK

Our recommendations for future application and extension of the work we have presented here are as follows:

String similarity is of only limited use

Beyond the classification into ICJME categories that is already possible using the defined COI terms as substrings to be de-tected in extracted relations, the use of string similarity – even advanced fuzzy string matching – is likely only of use for catching cases where there are typographical errors or small differences in spelling (e.g. British English spelling versus American English spelling) between extracted relation strings and known COI terms. Although thefuzz includes a confidence measure, this confidence measure is often high even when the choice of matching strings is incorrect. It may still be possible to find a way to use extremely low confidence outputs as an indication of relation strings that are worth investigating further to determine if they are truly an unknown type of relation.

Remove stop words from extracted relation strings before further processing

This may help improve the accuracy of embedding-based clustering as well as the accuracy of string matching. It should also help to eliminate some of the less useful extracted relation strings from the set under consideration, since many of the most commonly occurring (and least informative) relation strings are stop words (e.g. "is", "are", "be").

Experiment with both smaller and larger numbers of clusters for embedding-based clustering

It may be the case that clustering will give more meaningful results with fewer or more clusters.

REFERENCES

[1] Pal, Koninika, Vinh Thinh Ho, and Gerhard Weikum. "Co-clustering triples from open information extraction." *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD.* 2020. 190-194

[2] Graham, S. Scott, et al. "A dashboard for exploring clinical trials sponsorship and potential virtual monopolies." *JAMIA open 4.4* (2021): ooab089.

[3] Graham, S.S., Majdik, Z.P, Barbour, J.B., & Rousseau, J.F. (forthcoming). "Associations between aggregate NLP-extracted conflicts of interest and adverse events by drug product." *Studies in Health Technology and Informatics.*

[4] Graham, S.S., Majdik, Z.P., Clark, D., Kessler, M.M., & Hooker, T.B. (2020). "Relationships among commercial practices and author conflicts of interest in biomedical publishing. " *PLoS ONE* 15(7), e0236166.

[5] Graham SS, Majdik ZP, Clark D. Methods for Extracting "Relational Data from Unstructured Texts Prior to Network Visualization in Humanities Research."*Journal of Open Humanities Data*. 2020;6(1):8.

[6] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: "A Python Natural Language Processing Toolkit for Many Human Languages." *In Association for Computational Linguistics (ACL) System Demonstrations.* 2020. [pdf][bib]

[7] Santilli, Alice ML, et al. "Domain adaptation and self-supervised learning for surgical margin detection." *International Journal of Computer Assisted Radiology and Surgery* 16.5 (2021): 861-869.

[8] Demner-Fushman, Dina, Wendy W. Chapman, and Clement J. McDonald. "What can natural language processing do for clinical decision support?." *Journal of biomedical informatics* 42.5 (2009): 760-772.

[9] Friedman, Carol, Thomas C. Rindflesch, and Milton Corn. "Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine." *Journal of biomedical informatics* 46.5 (2013): 765-773.

[10] Meystre, Stéphane M., et al. "Extracting information from textual documents in the electronic health record: a review of recent research." *Yearbook of medical informatics* 17.01 (2008): 128-144.