HHS Public Access

Author manuscript

Stud Health Technol Inform. Author manuscript; available in PMC 2018 March 20.

Published in final edited form as:

Stud Health Technol Inform. 2015 ; 216: 539–543.

Assessing the Need of Discourse-Level Analysis in Identifying Evidence of Drug-Disease Relations in Scientific Literature

Majid Rastegar-Mojarad[a,b], **Ravikumar Komandur Elayavilli**[a], **Dingcheng Li**[a], and **Hongfang Liu**[a]

[a]Biomedical Statistics & Informatics, Mayo Clinic, Rochester, MN, USA

[b]University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Abstract

Relation extraction typically involves the extraction of relations between two or more entities occurring within a single or multiple sentences. In this study, we investigated the significance of extracting information from multiple sentences specifically in the context of drug-disease relation discovery. We used multiple resources such as Semantic Medline, a literature based resource, and Medline search (for filtering spurious results) and inferred 8,772 potential drug-disease pairs. Our analysis revealed that 6,450 (73.5%) of the 8,772 potential drug-disease relations did not occur in a single sentence. Moreover, only 537 of the drug-disease pairs matched the curated gold standard in Comparative Toxicogenomics Database (CTD), a trusted resource for drug-disease relations. Among the 537, nearly 75% (407) of the drug-disease pairs occur in multiple sentences. Our analysis revealed that the drug-disease pairs inferred from Semantic Medline or retrieved from CTD could be extracted from multiple sentences in the literature. This highlights the significance of the need of discourse-level analysis in extracting the relations from biomedical literature.

Keywords

Relation extraction; Discourse-level analysis; Literature-based discovery; Semantic Medline

Introduction

Information extraction (IE) aims to automatically extract information from text. To understand and extract information from text, IE systems should look at the text as whole and not only sentences separately. Researchers in the discourse domain agree that usually sentences/clauses are not understood in isolation [1]. Researchers in IE have studied discourse-level analysis [2] in applications such as question answering and dialogue generation. Of course, discourse-level analysis is built on top of sentence-level analysis and its performance somehow depends on sentence level. One of the well-studied parts of discourse-level analysis is identifying relation between sentences (clauses), called discourse relation. These types of relations could be contrast or explanation-evidence [1].

Address for correspondence: The correspondence for this study is Majid Rastegar-Mojarad. Mojarad.Majid@mayo.edu.

Two main tasks in IE are named entity recognition (NER) and relation extraction. In relation extraction, IE systems identify the relation between two or more entities where the entities could be located in one or multiple sentences. Identifying the latter type of relations needs discourse-level analysis such as coreference resolution which can be challenging.

The goal of this paper is to asess the need of discourse-level analysis for relation extraction and indirectly evaluate the amount of information that IE systems are missing if they just focus on sentence level. First, to perform this experiment, Semantic Medline [3] is used to generate a list of potential literature-based discoveries. The potential discoveries are limited to drug-disease relations as potential drug repositioning candidates. Then, we implement two kinds of searches on top of Medline abstracts to find any evidence of these discoveries, 1) Sentence-level search: we looked for pairs in the sentence 2) Discourse-level search: the whole abstract was searched. Comparing the results of these searches indirectly evaluates the amount of relations that IE systems can extract from sentence-level v.s. discourse-level analysis.

Background

Relation Extraction

The IE pipeline usually begins with NER which has been very well studied and various types of methods such as dictionary-based, rule-based, and machine learning algorithms have been applied to NER [4], [5]. After NER, the main task in IE is identifying relations among the entities. Relations could be binary or they could involve more than two entities. Most of relation extraction systems focuson relations in a single sentence [6]. It is imperative for a relation extraction system to extend beyond clauses and sentences and handle complex discourse-level analysis.

The importance of discourse-level analysis is well known and the analysis has been studied in several applications such as: question-answering systems [7]–[9], automatic dialogue generation [10], etc. Besides, there are several studies on extracting discourse-level relations. Marcu and Echihabi [1] developed an unsupervised approach to classify four types of relations between sentences/clauses. The goal here was extracting the relation between two or more entities that are mentioned in multiple sentences. Bach and Badaskar [6] reviewed some of these types of relation extraction systems and their applications. The methods for extracting such relations, often range from a simple distance based criteria to a more complex one that employs statistic and linear algebraic approaches [11] to extract explicit and implicit semantic relations from text.

In the biomedical domain, researchers developed various IE systems to extract different types of biomedical relations [12]–[16]. Quan et al [13] proposed two systems, unsupervised and semi-supervised to extract protein-protein interactions and gene-suicide associations. Their systems employed dependency parsing. Bundschus et al [14] developed biomedical relation extraction using conditional random fields. Their system identified relations between diseases and treatments; also relations between genes and diseases. All these systems just focused on relations within a single sentence and often ignored relations involving entities across sentences. In general, extracting relations that involve entities

mentioned in two separate sentences is a complex one and requires special NLP techniques such as coreference resolution and complex semantic analysis.

Literature-based discovery

Literature-based discovery (LBD) aims to find a connection/correlation between concepts using scientific literature. Therefore, LBD is one kind of special relation detection we are interested in. Many LBD systems have been developed in the biomedical domain to generate new hypotheses that potentially could lead to new discoveries. Swanson [17] first introduced this approach and applied it to find a correlation between migraine and magnesium. After that, a number of studies followed his approach and had interesting discoveries. The important part of LBD is how to decide that two concepts are correlated. The most commonly used approaches are co-occurrence analysis [18], Association Rules [19], TF-IDF, Z-Score, and Mutual Information Measure [20]. Yetisgen-Yildiz and Pratt [20] briefly discussed these approaches and meanwhile Andronis et al. [21] reviewed literature mining systems which identify potential drug repurposing candidates. Another approach to identify correlated concepts is using semantic relations [22]. Hristovski et al. proposed [22] using semantic predications to enhance LBDs. In this study, semantic predications are used to create a list of potential LBDs (drug-disease pairs). The generated LBDs in this study are potential drug repurposing candidates.

Semantic Medline, semantically enhanced with predicates extracted by SemRep [23] from Medline titles and abstracts, contains approximately 70 million semantic predications. Predications are triplets of *Subject, Predicate*, and *Object* where the subject and object are biomedical entities (drug, gene or disease in our study), and predicate shows the type of relation between the entities such as inhibits, interaction with, associated with, etc. The predications are stored in a relational database called Semantic Medline Database (SemMedDB). It has been used in many studies to facilitate knowledge discovery [24], [25].

In this study, we plan to assess the need of going beyond sentences for extracting relations. Our goal in this study is to highlight the need and asess the amount of relations, relation extraction systems that are potentially missing if they just focus on sentence-level analysis. For this purpose, following Swanson's model, a list of potential drug-disease relations is generated from literature. The extracted LBDs are assumed as relations between drugs and diseases that could be mentioned implicitly or explicitly in literature. Two kinds of searches are conducted to identify any evidence of these relations in the sentence level or discourse level of Medline abstracts.

Methods

Our study contains two steps: 1) generating a list of drug-disease pairs based on LBDs and 2) using the discoveries to evaluate the drug-disease relation extraction by comparing the extractions from a single sentence to the one from multiple sentences.

Generating LBDs

In the first step, we generated a list of potential relations between drugs and diseases. We followed Swanson's model [17] to generate a list of LBDs. According to Swanson's model

if one scientific study notes a correlation between concept A (Starting concept) and concept B (Linking concept), and another study mentions a correlation between concept B and concept C (Target concept), then there might be a correlation between concept A and C. In our study, drug is the starting concept and disease the target concept, gene serves as a conceptual link between the two leading to the generation of a list of potential drug-disease relations.

In contrast to commonly used approaches such as document-level or sentence-level co-occurrence of entities or concepts to generate LBDs, we used semantic predications as evidence for correlation between entities [22]. For example, from the following two predications from SemMedDB:

- Flecainide (Drug) *INTERACTS WITH* SCN5A (Gene)
- SCN5A (Gene) *ASSOCIATED WITH* Heart Failure (Disease)

The system generates *Flecainide-Heart* Failure as a potential new drug-disease pair. Figure 1 shows the architecture of our LBD process.

There are two major reasons for our choice of using semantic predication to generate LBDs 1) semantic predication takes biological meaning into consideration 2) the semantic type of the interaction and the contextual information about the interaction such as *NEGATION* allow us to filter unnecessary predications: *NEGATIVE TREATS*, *NEGATIVE ASSOCIATED WITH*, etc. At this point, we do not consider any threshold based on a frequency measure to further eliminate drug-disease pairs. We believe that using a frequency measure will eliminate potentially novel relations from our initial list.

In order to further refine the drug-disease pairs to be relevant for LBD, we used Timeline profiles to narrow down the drug-disease pairs with literature evidence. We only considered drug-disease pairs where the first literature evidence appeared after their related drug-gene and gene-disease pairs. Equation (1) clarified our Timeline profile:

$$Max\left(\frac{Min\left(Y_{sp}\left(Drug-Gene\right)\right)}{Min\left(Y_{sp}\left(Gene-Disease\right)\right)}\right) \leq Min\left(Y_{c}\left(Drug-Disease\right)\right) \quad (1)$$

where Y_{sp}(Drug-Gene) indicates the publications' date of all Medline abstracts that contain at least one semantic predication between Drug-Gene. For Drug-Disease, we considered publications that contain at least one co-occurrence of the entities. For example, assume two studies in 2003 and 2005 reported an association between Drug A and Gene B; and three studies in 2006, 2008, and 2009 mentioned an association between Gene B and Disease C. The left hand side of the equation would be Max (Min (2003, 2005), Min(2006, 2008, 2009)) = 2006. So if Drug A and Disease C appeared together in any publication before 2006, we do not consider Drug A-Disease C as a potential discovery.

Evaluating sentence and discourse-level relation extraction

After generating LBDs, we used the drug-disease pairs identified earlier to evaluate their co-occurrence in literature, thereby estimating the need for discourse-level analysis in relation extraction. So, we have a list of drug-disease relations that potentially could be mentioned in the scientific literature. In the second step, we searched Medline abstracts to find any co-occurrence (evidence) of drug and disease pairs. We categorized the drug-disease pairs with at least one evidence in Medline, into two groups, i) sentence-level relation and ii) discourse-level relation. For a given pair, if the drug and the disease appeared in a single sentence in at least one of the abstracts, we classified it as a sentence-level relation, otherwise we considered it as a discourse-level relation.

In order to assess the true validity of the drug-disease pairs we used the Comparative Toxicogenomics Database (CTD) [26], manually curated biological relations as our reference standard. CTD contains annotations of biological relations from various categories such as chemical-disease, gene-disease, drug-gene associations along with the corresponding PubMed citation. We compared the drug-disease pairs identified by the system against the chemical-disease associations in CTD. We considered our pair to be valid only if there was a match in the PubMed citation in addition to the drug-disease pairs.

Results

Retrieval of LBD relations

In order to generate potential LBD pairs, we started with 1,710 approved drugs from DrugBank [27]. We extracted 4,096 drug-gene (A-B) unique pairs where the drugs were restricted to our chosen list of 1,710 along with the semantic predicates from SemMedDB. For all the genes mentioned in A-B pairs we further retrieved 2,741 gene-disease (B-C) relations from SemMedDB. With gene being the common link between Drug-Gene and Gene-Disease, we inferred 71,842 drug-disease (A-C) relations. Further analysis revealed that only a small fraction of (118) of the 71,842 drug-disease pairs had an overlap in terms of abstracts from which the drug-gene (A-B) and gene disease (B-C) pairs were identified. We also found 14,451 drug-disease pairs catalogued in SemMedDB since they co-occurred in the same sentence. We have 57,391 drug-disease pairs also indirectly inferred through the Swanson's model of LBD and not present in SemMedDB. The results of our study is illustrated in Figure 2.

Comparison of sentence level and discourse level

We further attempted to assess the level of textual extraction required to identify the drug-disease relations (A-C pairs). We searched Medline using "Drug AND Disease" as the query, to find any evidence of the identified pairs in the first step. We found only 37,719 (52.05%) of the 71,842 drug-disease pairs with at least one literature evidence.

Timeline analysis (Eq 1) further narrowed down the number of drug-disease pairs to 8,772 (23.25%) from 37,719. 6,450 drug-disease relation pairs (73.52%) out of 8,772 identified earlier transcend sentence boundaries, demanding the requirement of discourse-level analysis for textual extractions.

We performed additional analysis on the remaining 2,322 drug-disease relational pairs, which had at least one literature evidence as sentence-level co-occurrence. For these 2,322 pairs, we found 89,805 literature evidences, which indicated that there were more than one literature evidence for each pair. Further composite analysis revealed that there were far more literature evidences across sentences than from a single sentence as shown in Figure 3.

We further carried out an assessment to validate the 8, 322 drug-disease pairs against a curated resource (CTD). We found that only 537 (6.4%) of them matched the gold standard. The low match was due to the fact that in addition to the match in the drug and disease names we also considered an agreement in the cited literature. Further analysis revealed that only 130 (24.20%) of the 537 that matched the gold standard occurred in a single sentence while the rest (75.80%) appeared across different sentences. If we ignored the match in the literature citation then we found 17,094 drug-disease pairs in CTD, which was significantly higher than the earlier one.

Discourse-level analysis may impact Time lag of LBD

We performed another interesting analysis to study whether discourse-level analysis would have a positive impact on the time lag due to the reporting of causal pairs (drug-gene/gene-disease) and the appearance of drug-disease pairs in the scientific literature. Figure 4 plots the cumulative percentage which represents a cluster of drug-disease pairs observed at a time zone at both sentence level and discourse level. The trend shows that performing discourse-level analysis would significantly advance the identification of the drug candidate for a specific disease.

Discussion

In this paper, we attempted to assess the need for performing discourse-level analysis to extract biological relations that might potentially lead to serendipitous discovery. We found that biological relations quite often transcend clausal sentences and hence demand mechanisms to connect information across such boundaries. Using drug-disease relation discovery as an example, the description of 23,268 potential drug-disease relations across sentence boundaries against the 14,451 relations within sentences indicates the importance of a discourse-level analysis requirement to extract these relations from biomedical literature.

Across all experiments we observed a consistent trend that drug-disease relations that occurred across sentences outnumbered the ones within sentences. Our evaluation against a curated resource such as CTD also reinforces this fact. Results from two experiments need attention 1) Frequency of drug-disease relation co-occurrence at sentence level against the discourse level (Figure 3) was skewed towards the latter than the former. 2) The time lag analysis where the discourse-level analysis would have significantly reduced the time lag between the scientific reporting of causal pairs (drug-gene/gene-disease) and drug-disease relational pairs. This shows the need for advance discourse-level analysis approaches to extract information from literature in time and its ability to hasten the pace of discovery.

Another significant observation is the amount of potential false-positive drug-disease relations identified through literature mining. We observed a substantial reduction in the number of drug-disease relations when we compared the literature based drug-disease pairs with that of curated resource CTD. The reduction among the discourse-level pairs is far greater (3.3 times) when compared to the ones from the sentence level (2.5 times). As a note of caution discourse-level analysis might potentially extract more false positives, provided one does not explore sophisticated linguistic/statistical approaches to handle them.

As a final note we would like to mention that the way we generate the seed pairs of drug-disease relations in this study, coupled with our timeline based restriction analysis reinforces that the goal of our study is centered on LBD of novel relations. However we also understand the limitations of this study. Simple sentence-level or document-level co-occurrence information does not imply any biological relation between entities. As pointed out earlier, simple discourse-level analysis has the danger of extracting more false positives. In certain tangential areas of research such as drug repurposing certain techniques [28], [29] were explored to reduce the false positives.

Conclusion

In this paper, we investigated the extent of the need of discourse-level analysis for drug-disease relation extraction from biomedical literature. We used Semantic Medline to extract LBDs and then, based on co-occurrence analysis, we collected any evidence of the discoveries in Medline abstracts. We categorized the evidence into two categories, sentence level and discourse level. From subsequent analysis we infer that there is a potential to miss more than 70% of drug-disease relations when we extract information from the sentence level. This clearly demonstrates the need for deeper discourse-level analysis, which may translate to significant improvement in the state of the art of NLP techniques.

In the near future, we plan to explore much more sophisticated NLP approaches, a significant departure from the co-occurrence based extraction which may lead to significant improvement in the state of the art.

Acknowledgments

This study was made possible by the National Institutes of Health grant R01 GM102282-01.

References

1. Marcu D, Echihabi A. An Unsupervised Approach to Recognizing Discourse Relations. 2002
2. Johanna PW-H, Moore D. Discourse in Computational Linguistics and Artificial Intelligence.
3. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics. Dec; 2012 28(23):3158–3160. [PubMed: 23044550]
4. Goulart RRV, Strube de Lima VL, Xavier CC. A systematic review of named entity recognition in biomedical texts. J Braz Comput Soc. Mar; 2011 17(2):103–116.
5. Ghiasvand O. Disease Name Extraction from Clinical Text Using Conditional Random Fields. Theses and Dissertations. May.2014
6. Bach N, Badaskar S. A Review of Relation Extraction.

7. Sun M, Chai JY. Discourse processing for context question answering based on linguistic knowledge. Knowledge-Based Systems. Aug; 2007 20(6):511–526.
8. Kim S, Bracewell R, Wallace K. From discourse analysis to answering design questions. Proceedings," of the Workshop on the Application of Language and Semantic Technologies to support Knowledge Management Processes. 2004:43–49.
9. Suzan Verberne LB. Evaluating discourse-based answer extraction for why-question answering. 2007
10. Prendinger H, Piwek P, Ishizuka M. Automatic Generation of Multi-Modal Dialogue from Text Based on Discourse Structure Analysis," in. International Conference on Semantic Computing 2007. 2007:27–36. ICSC 2007.
11. Zahedi M, Kahani M. SREC: Discourse-level semantic relation extraction from text. Neural Comput & Applic. Nov; 2013 23(6):1573–1582.
12. Qian L, Zhou G. Tree kernel-based protein–protein interaction extraction from biomedical literature. Journal of Biomedical Informatics. Jun; 2012 45(3):535–543. [PubMed: 22388011]
13. Quan C, Wang M, Ren F. An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature. PLoS ONE. Jul.2014 9(7):e102039. [PubMed: 25036529]
14. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics. Apr.2008 9(1):207. [PubMed: 18433469]
15. Rastegar-Mojarad M. Extraction and Classification of Drug-Drug Interaction from Biomedical Text Using a Two-Stage Classifier. Theses and Dissertations. Dec.2013
16. Mehrabi S, Krishnan A, Tinsley E, Sligh J, Crohn N, Bush H, Depasquale J, Bandos J, Palakal M. Event Causality Identification Using Conditional Random Field in Geriatric Care Domain," in. 2013 12th International Conference on Machine Learning and Applications (ICMLA). 2013; 1:339–343.
17. Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med. 1988; 31(4):526–557. [PubMed: 3075738]
18. Weeber M, Klein H, W LT, Berg J, Has DRS. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migrainemagnesium discoveries. J Am Soc Inf Sci Tech. 2001:548–557.
19. Hristovski D, Peterlin B, Dzeroski S. Literature-based Discovery Support System and Its Application to Disease Gene Identification. Proc AMIA Symp. 2001:928.
20. Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. J Biomed Inform. Aug; 2009 42(4):633–643. [PubMed: 19124086]
21. Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A. Literature mining, ontologies and information visualization for drug repurposing. Brief Bioinformatics. Jul; 2011 12(4):357–368. [PubMed: 21712342]
22. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting Semantic Relations for Literature-Based Discovery. AMIA Annu Symp Proc. 2006; 2006:349–353.
23. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. Dec; 2003 36(6):462–477. [PubMed: 14759819]
24. Rastegar-Mojarad M, Li D, Liu H. Operationalizing Semantic Medline for meeting the information needs at point of care. presented at the AMIA Clinical Research Informatics Summit. 2015
25. Moosavinasab S, Rastegar-Mojarad M, Liu hongfang, Jonnalagadda S. Towards Transforming Expert-based Content to Evidence-based Content. presented at the AMIA Clinical Research Informatics Summit. 2014
26. Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Res. Oct.2014
27. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. Jan; 2011 39(Database issue):D1035–1041. [PubMed: 21059682]

28. Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, Dai Q, Levy M, Shah A, Han X, Ruan X, Jiang M, Li Y, Julien JS, Warner J, Friedman C, Roden DM, Denny JC. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J Am Med Inform Assoc. Jul.2014

29. Rastegar-Mojarad M, Ye Z, Kolesar JM, Hebbring SJ, Lin SM. Opportunities for drug repositioning from phenome-wide association studies. Nat Biotech. Apr; 2015 33(4):342–345.

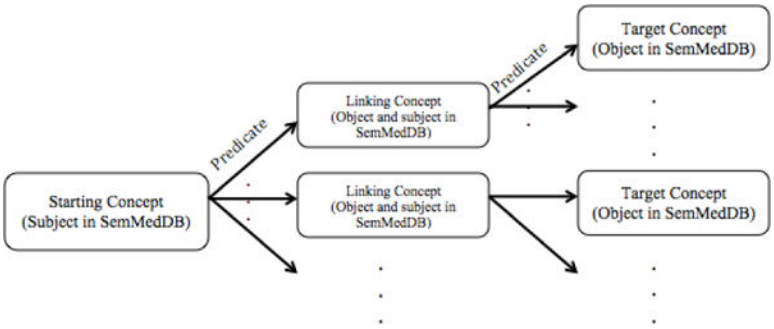

Figure 1.

Architecture of our LBD system. In this system, the starting concept is drug, the linking concept is gene, and the target is disease that leads to drug-disease discoveries. Our system uses Semantic predications as evidence of correlation between the concepts.

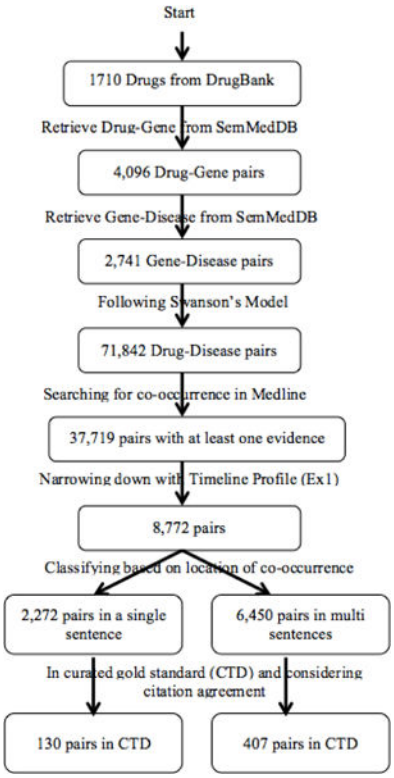

Figure 2.
Results of our study

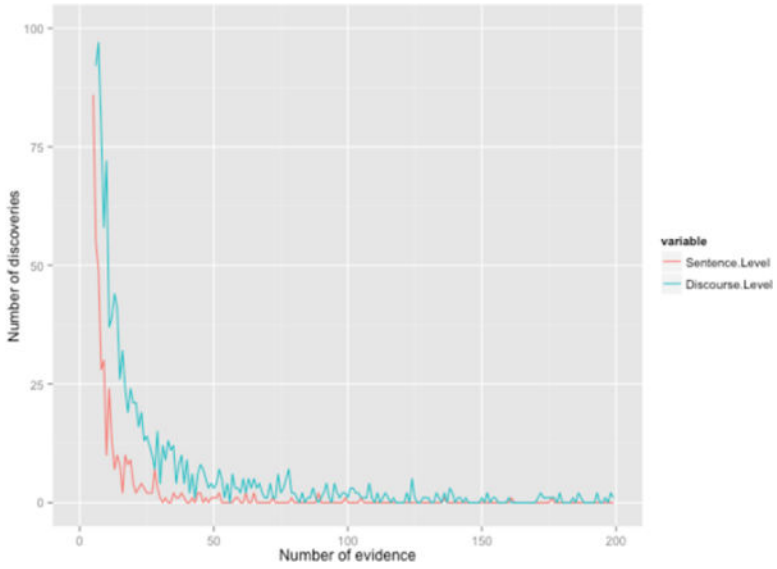

Figure 3.

Comparison of frequencies of drug-disease relations' co-occurrence in a single sentence versus multiple sentences

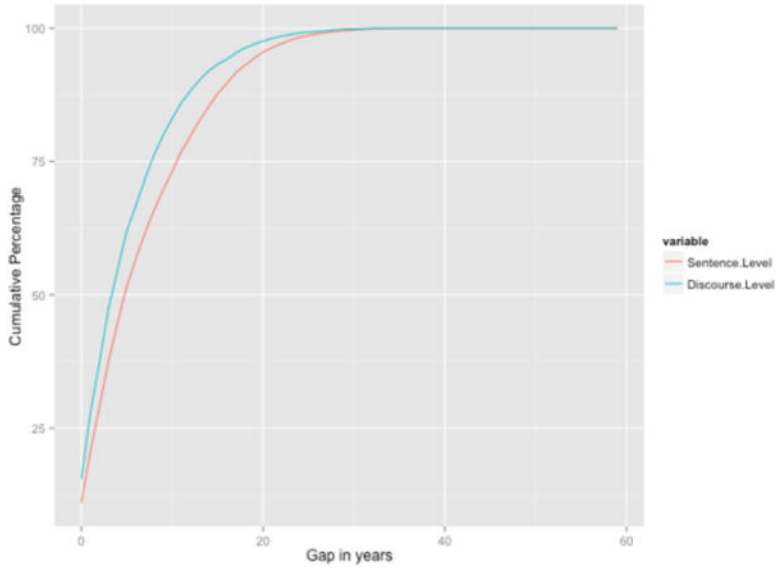

Figure 4.

Comparing the time gap between the first co-occurrence of discovery and the causal pairs (Sentence level versus discourse level)