Extracting Gene-Disease Relations from Text to Support Biomarker Discovery

Paul Thompson Sophia Ananiadou
School of Computer Science, University of Manchester, UK
{paul.thompson, sophia.ananiadou} @manchester.ac.uk

ABSTRACT

The biomedical literature constitutes a rich source of evidence to support the discovery of biomarkers. However, locating evidence in huge volumes of text can be difficult, as typical keyword queries cannot account for the meaning and structure of text. Text mining (TM) methods carry out automated semantic analysis of documents, to facilitate structured searching that can more precisely match users' information needs. We describe our TM approach to the detection of sentence-level associations between genes and diseases, as a first step towards developing a sophisticated search system targeted at locating biomarker evidence in the literature. We vary the sophistication of our detection methodology according to sentence complexity, using either co-occurring mentions of genes and diseases, or linguistic patterns obtained using evidence from approximately 1 million biomedical abstracts. We demonstrate that this method can detect associations more successfully than applying a single technique, with an accuracy that compares highly favourably to related efforts. We also show that the identified relations can complement those detected using alternative approaches.

KEYWORDS

Text mining; gene-disease relations; biomarkers; dependency grammar

ACM Reference format:

Paul Thompson and Sophia Ananiadou. 2017. SIG Proceedings Paper in Word Format. In *Proceedings of DH'17, July 02-05, London, United Kingdom*, 10 pages. DOI: http://dx.doi.org/10.1145/3079452.3079472

1 INTRODUCTION

Over recent years, there has been an increasing trend towards stratified medicine, in which specific therapies are targeted towards patients with certain characteristics. This approach is dependent upon the development *biomarker* tests, to determine which people will respond best to which therapies. Biomarkers may be defined as "objective indications of medical state observed from outside the patient – which can be measured accurately and reproducibly" [1]. Laboratory-based biomarkers build up a biological blueprint of patients, which can determine whether they possess a specific variant of a gene, their levels of gene expression, etc. Such information can be used in diagnosing a disease or to predict response to different treatments, etc.

Scientific advances have led to a deluge in the availability of both biomedical data and publications aiming to contextualise and interpret this data. The biomedical literature thus represents a hugely valuable repository of information which, when used effectively in combination with other biomedical data, can provide vital evidence to support the discovery of biomarkers and the development of associated tests. However, the sheer size of the literature can make locating supporting evidence akin to finding a needle in a haystack.

Keyword querying facilities provided for large repositories of scientific publications are typically poorly aligned to the needs of the researcher, whose aim is generally to locate different pieces of *knowledge*, e.g., evidence of associations between genes and diseases. However, queries involving individual words and phrases are not sufficiently expressive to represent knowledge requirements. For example, genes and diseases are *concepts*, each of which may be described in text using a variety of different words or phrases, which can be difficult to predict or enumerate. A simple example is *rheumatoid arthritis*, which can also be referred to using the abbreviation *RA*. Particularly for genes, however, there may be a very large number of possible variations. Additionally, a given word or phrase could refer to multiple concepts (e.g., *RA* can also be an abbreviation for *retinoic acid* and *radial artery*, amongst others). Furthermore, keyword queries are not sufficiently powerful to isolate documents that specifically mention an *association* between *rheumatoid arthritis* and one or more (unspecified) genes.

The inability of keyword queries to take into account the *meaning* and *structure* of text means that, on one hand, they usually retrieve many irrelevant results, whilst on the other hand, certain documents that *are* relevant may fail to be retrieved. Accordingly, it can be extremely difficult to exploit the rich knowledge available in literature to its full potential, and much valuable information may remain "locked away" and undiscovered, which can hinder scientific progress [2].

Text mining (TM) methods offer a solution, by carrying out automated analyses of huge collections of documents, to detect and structure various aspects of their meaning or *semantics*. Using these analyses, it is possible to develop systems providing

DH '17, July 02-05, 2017, London, United Kingdom

ACM ISBN 978-1-4503-5249-9/17/07.
http://dx.doi.org/10.1145/3079452.3079472

structured, knowledge-centric search facilities, making it far easier to tap into the information encoded in vast repositories.

The mature nature of many TM techniques means that it is already feasible to develop systems that allow searching at the level of concepts, rather than words. Automatic *named entity recognition (NER)* tools identify and categorise words and phrases in text that denote concepts of interest. The results of NER can allow sense-based restrictions to be placed on queries such that, e.g., only documents in which *RA* corresponds to a disease will be retrieved. Automatic *normalisation* of recognised named entities (NEs) aims to identify all ways in which a given concept could be mentioned, allowing a search for *interleukin 2* to automatically retrieve documents containing synonyms and abbreviations such as *IL2*, *IL-2*, *T-cell growth factor*, *TCGF*, *lymphokine*, etc. Various high performance tools can recognise and/or normalise various concepts of relevance to biomarkers, e.g., genes [3, 4], diseases [5], anatomical entities [6], chemicals/drugs [7, 8] and gene variants [9]. Further processing can identify associations or *relationships* that hold between the recognised/normalised concepts, which can facilitate the development of powerful systems that allow search for many different relations over huge numbers of documents (e.g., [10]).

In this paper, we describe our novel approach to extracting relationships between genes and diseases described within single sentences, as a step towards developing a search system that will allow such relations to be located and filtered according contextual and interpretative information, and make it possible to find answers to complex queries, e.g., *In which population subgroups is Gene X is a putative biomarker for Disease Y?*

Our approach uses extraction methodologies of differing levels of sophistication, according to the complexity of sentences. We show that in simpler sentences, it can be assumed that genes and diseases are related if they are mentioned together. However, linguistic patterns generated using evidence from approximately 1 million biomedical abstracts about how relationships are typically described are used to "disentangle" individual associations between multiple genes and diseases in complex sentences. We demonstrate that this combined approach is more successful than using a single method to identify all relations. We also show that our method can extract relations with an accuracy that compares highly favourably to related efforts, and that the types of evidence found using our approach complement those extracted using other methods.

In the remainder of this paper, section 2 outlines related research into relation extraction, while section 3 introduces the main characteristics of our approach. Section 4 describes the annotated corpus used for evaluation, while section 5 explains and reports the results of our relation extraction experiments. Based on these, section 6 proposes an optimal method for gene-disease relation extraction and compares our final results with those of related efforts. Section 7 describes the application of our method to a larger data set, and shows how its output can complement that of a related method. Finally, in section 8, we provide some conclusions and outline our planned future work.

2 RELATED WORK

The identification of relationships between various types of concepts in biomedical texts, including genes and diseases, has formed the focus of research efforts of varying sophistication, ranging from those which detect general associations, to those which categorise the nature of the relationship, e.g., according to whether it occurs due to altered expression or mutation of the gene [11, 12], and even to those which assign "interpretative" information about the relation, e.g. whether it constitutes a hypothesis or experimental results, etc. [13, 14].

The methodology employed and the nature/complexity of the information extracted can depend on the types of supporting resources available. Approaches can be coarsely divided into *supervised* and *unsupervised* methods. Supervised approaches use a sample set (or *corpus*) of documents, in which relations have been manually marked-up (or *annotated*) by domain experts. Machine learning (ML) methods then learn characteristics of the sample documents that will allow the relations of interest to be detected automatically in unseen texts. Whilst ML methods can be trained to recognise and categorise complex relationships in biomedical text [15-17], the production of enough high-quality, expert-annotated text can be complex, time-consuming and expensive. ML methods can also be highly sensitive to the specific features of the text on which they are trained. This can reduce their portability, unless training corpora are very heterogeneous in terms of subject areas, article types, etc.

Various annotated corpora include gene-disease relationships, some of which have been used for ML purposes [11, 12, 18-21]. However, most such corpora suffer from drawbacks, limiting their suitability for training systems that can accurately detect relationships in diverse types of literature articles. These include a lack of public availability of the annotated data (e.g., [20]); the annotations not being linked to specific text spans in documents (e.g., [22-24]); the number of annotated relations being very small (e.g., [25-27]); and the corpus not being representative of how information is expressed in biomedical literature [11].

In contrast to supervised methods, *unsupervised* methods are not reliant on the availability of annotated corpora. Although they can be less suited to detecting detailed and complex relationships, at least without large amounts of manual effort, they can be very powerful in detecting simpler relationships that exploit general features of text. By using more general features, unsupervised methods are often less bound than supervised methods to specific text types, and hence they can be more readily applied to documents that have varying characteristics.

A simple example of an unsupervised method is one that assumes that possible relationships exist between *all* genes and diseases mentioned together in the same document; the probability of an association becomes much greater if there are many documents that mention the gene and disease [28]. The powerful nature of calculating such co-occurrence statistics over large document collections as a means to find associations has been exploited in a number of semantically-enhanced search engines (e.g., [29, 30]). Additional techniques to improve the results include automatically classifying documents according to

whether they concern topics of interest [31-33], or by identifying documents that contain additional keywords denoting relationships of interest (e.g., *alteration, association*) [34-36].

Other methods only consider relations within single sentence, given the higher probability of associations in this context [37], possibly accompanied by keywords belonging to relation-denoting semantic classes (e.g., *association)* [38]. However, simple co-occurrence of genes, diseases and other keywords is not sufficient to detect relationships accurately in more complex sentences, e.g., *The BCAP31 gene is located between SLC6A8, associated with* ***X-linked creatine transporter deficiency****, and ABCD1, associated with* ***X-linked adrenoleukodystrophy****.* A co-occurrence-based approach would identify 6 pairwise associations between the underlined genes and emboldened diseases. However, only 2 associations are actually mentioned.

To better handle such cases, sentence structure can be exploited to identify valid relationships, e.g., using sets of hand-crafted linguistic patterns [26, 39-42], possibly supported by ontological information [43]. However, the diverse ways of describing relationships makes it impossible to develop patterns that can account for all cases. Domain-adapted syntactic parsers, (e.g., [44, 45]) analyse the "deeper" syntactic structure of sentences, in order to identify consistent grammatical relationships between words and phrases, regardless of the exact sentence structure. For example, in the sentence *PKLR and NOS1AP are reported to be strongly associated with* ***type 2 diabetes****,* parsers can identify the relationships between *PKLR, NOS1AP* and *type 2 diabetes,* via their grammatical links with the verb *associate,* regardless of intervening verbs and adverbs, etc.

Grammatical information has previously been exploited to recognise gene-disease and pharmacogenomic relationships ([46, 47], possibly further filtered by imposing limits on the distance between the gene and disease [37], or by applying ML to determine the most likely relationship-denoting grammatical patterns [18]. The latter approach (BeFree) has been applied to a large collection of MEDLINE abstracts, resulting in the detection of over 300,000 relations, which have subsequently been integrated within DisGeNET [10, 48], a large repository of gene-disease relations with an associated search interface, combining expert curated information with text-mined data.

Since a purely grammar-driven approach may identify fewer relations than using co-occurrence [46], another approach used specific sentence characteristics to determine whether a co-occurrence or grammar-based relation extraction strategy should be employed [37]. However, such a strategy has not been applied to the detection of gene-disease associations.

3 OUR APPROACH

In common with several studies mentioned above, our aim has been to extract general gene-disease relations within single sentences. The defining features of our approach are as follows:

- We have explored how the performance of different extraction techniques (co-occurrence or grammar-driven) varies according to the *complexity* of the sentence.
- Our *selective* use of co-occurrence is aimed at maximising the number of relations identified, without sacrificing the increase in accuracy that can be achieved by applying grammar-driven extraction to more complex sentences.
- We use an *unsupervised* method to generate relationship-denoting grammatical patterns, using evidence from approximately 1 million MEDLINE abstracts.
- We explore various ways to *refine* and *filter* these patterns, in order to achieve maximum relation extraction accuracy.

4 EVALUATION CORPUS

A "gold-standard" human-annotated corpus, marked up with genes, diseases and relationships between them was needed to allow us to evaluate the relationships identified by our methods against the human-identified relationships.

EU-ADR [25] is one of the few suitable corpora, consisting of complete, randomly selected abstracts, which reduces bias towards specific subject areas, and provides evidence of how relations are expressed in different sentence types. Annotations consist of diseases, genes and associated variants, with 265 relationships annotated between them. The association of annotations to text spans is also advantageous.

However, upon a close examination of the corpus, we found a number of inconsistencies, i.e., certain types of information are annotated in some sentences, but not in others. Table 1 exemplifies the 4 main types of inconsistencies that we found.

We decided to resolve these issues by adding additional relation annotations (bringing the total number to 477), given our interest in detecting *all* potential relations between genes and diseases, regardless of how they are described (e.g., as full forms or abbreviations) and of their intended interpretation, e.g., whether the relationship is a subject of investigation, a definite experimental observation, a tentative analysis, etc.

Annotation quality was maintained in the augmented corpus by adding new relationships only when existing relation annotations provided sufficient evidence of the validity of the new relationship. Specific guidelines used included:

- Missed relations involving variants were only added if a relationship involving the associated gene and the same (or related) disease was already annotated in the abstract.
- Missed relations were added where the sentence structure provided reliable evidence of the validity of the relation, according to its relationship with an existing annotated relation (e.g., where a gene is in a list with another gene for which a relationship has already been annotated).
- Missed relations were annotated in contexts that had been annotated inconsistently, in cases where the sentence structure makes the gene-disease association very clear, i.e.: as a subject of investigation: *The aim of the present study was to investigate relationships between single nucleotide polymorphisms* ***(SNPs) in the human SLC12A3 gene*** *and* ***essential hypertension*** *(EH) in Japanese* or in a title: *Three new* ***BLM gene mutations*** *associated with* ***Bloom syndrome****.*

Table 1: Inconsistencies in the EU-ADR corpus – blue=diseases; green=genes/variants annotated as related to the disease; red=genes/variants NOT annotated as related to the disease

Category	Example 1		Example 2	
	Sentence	Comment	Sentence	Comment
Genes/ variants	*The haplotypes constructed from the three SNPs (rs3840846, rs3826047 and rs3743500, in order) in the 5'-upstream of NPTN showed a significant association with schizophrenia*	Four associations annotated involving **BOTH** the SNPs and gene	*Ala394Thr polymorphism in the clock gene NPAS2: a circadian modifier for the risk of non-Hodgkin's lymphoma.*	**ONLY** associations between disease and gene annotated. Relationship involving the *Ala394Thr* polymorphism **NOT** annotated.
Titles	*Association of variants of the interleukin-23 receptor gene with susceptibility to pediatric Crohn's disease.*	Association between gene and disease **IS** annotated	*Malic enzyme 2 and susceptibility to psychosis and mania.*	**NO** associations annotated (2 associations missed)
Investigative sentences	*The study investigated the possible association of NRG3 gene and schizophrenia in a Han Chinese population.*	Association between gene and disease **IS** annotated	*The authors investigated the correlation between the presence of the rs42524 polymorphism in COL1A2 and the occurrence of sporadic IAs in Chinese patients*	**NO** associations annotated (2 associations missed)
Full forms and abbreviations	*Fetal haemoglobin (HbF) level modifies the clinical severity of HBB disorders*	Associations annotated between the disease and **BOTH** the full form and abbreviation of the gene	*Recently an association was shown between the single nucleotide polymorphism (SNP), rs11209026, within the interleukin-23 receptor (IL23R) locus and Crohn's disease (CD).*	Only **SOME** relationships annotated. Missing relationships between *IL23R* and *CD*, and between *interleukin-23 receptor* and both short and long forms of the disease.

5 RELATION EXTRACTION EXPERIMENTS

Using the augmented EU-ADR corpus for evaluation, we assessed the accuracy of various approaches to extracting gene-disease relations. We divided sentences containing at least one gene and one disease mention into 4 different categories, allowing us to assess how different methods perform when applied to sentences of varying complexity. We characterise "complexity" according to the number of gene and disease mentions in a sentence. The 4 sentence categories are as follows:

- **Single-both (SB)** – single gene and single disease mention
- **Single-gene (SG)** – single gene and multiple diseases
- **Single-disease (SD)** – single disease and multiple genes
- **Multiple-both (MB)** – multiple gene and disease mentions

We treat annotations of both genes and their variants as a single class, which we refer to as "gene". This is because common types of patterns are often used to denote associations involving both concept categories, and we wanted to make maximum use of all annotated relationships in evaluating our methods. We report our results using the following measures:

- **(P)recision** – The proportion of relations recognised by the method that are actually correct, according to comparison with annotations in the augmented EU-ADR corpus.
- **(R)ecall** – The proportion of all correct relations (according to the augmented EU-ADR) recognised by the method.
- **(F)-score** – The harmonic mean of precision and recall, providing a single overall measure of performance.

5.1 Baseline Methods

We began by performing 2 sets of simple "baseline" experiments, as a point of comparison for more complex methods:

- Sentence-based co-occurrence (i.e., every gene mentioned in the sentence is considered to be related to every disease).
- Unrestricted dependency paths (i.e., genes and diseases are only considered to be related if they are connected via grammatical relations).

A grammatical dependency analysis of the sentence *Many childhood gliomas are associated with activation of BRAF* is shown in Fig. 1. The analysis takes the form of a tree, whose root is the main verb in the sentence, i.e., *associated*. Each "node" in the tree corresponds to a word and each "branch" connects two words, and has a label indicating the type of grammatical relationship that holds between them. For example, since the sentence is in the passive voice, *gliomas* is identified as the "head" of a noun phrase that corresponds to the passive subject (*nsubjpass*) of the verb *associated*. The word *gliomas* is itself part

of the compound noun *childhood gliomas;* the fact that *childhood* modifies *gliomas* is denoted by the *nn* label.

Within the tree structure, we can trace a *path* connecting *BRAF* with *gliomas*. Working upwards from both the gene and the disease nodes, the routes meet at a common node, and thus a *dependency path* can be traced between them. In the example shown in Fig. 1, the path is *BRAF* → *activation* → ***associated*** → *gliomas,* in which *associated* constitutes the common node.

For the unrestricted dependency baseline method, we assume that *any* path connecting a gene and a disease denotes an association. We used the recently released MEDLINE dependency analyses [49], which apply the BLLIP constituent parser [50], with a biomedical model [45]. We use the conversions to the collapsed Stanford dependency scheme [51].

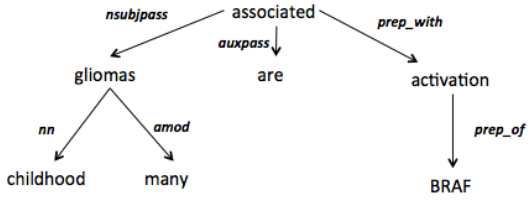

Figure 1: Example of dependency parsing

Table 2 shows the baseline method results. In line with previous findings (e.g., [20, 37]), simple co-occurrence achieves high overall results. However, this general result hides the considerable discrepancies in precision between sentences of different complexities, e.g., the difference of almost 20% between SB and MB sentences. This provides strong motivation to consider alternative approaches for more complex sentences. Unrestricted dependency paths appear to offer few advantages over simple co-occurrence. The complex nature of dependency trees and possible paths through them may mean that there is no strong association between some connected words.

Table 2: Baseline approaches for different sentence types

		Co-occurrence			Dependency path		
	# rels	P	R	F	P	R	F
SB	116	90.5	100	95.0	90.3	97.1	93.6
SG	66	83.3	100	90.1	82.8	96.3	89.1
SD	249	79.9	100	88.8	81.4	99.5	89.6
MB	164	71.2	100	83.4	71.6	98.3	82.3
Total	477	80.2	100	89.0	80.1	98.3	88.6

5.2 Collecting Textual Evidence about Relations

Since other studies have reported positive results using dependency relations, we investigated imposing restrictions on which dependency paths constitute valid relations. We obtained evidence about how gene-disease relationships are described within approximately 1 million MEDLINE abstracts from 2015. We used the DisGeNET database [10] of gene-disease relations to obtain gene-disease pairs for which evidence would be sought in the MEDLINE abstracts. We used 2 sets of pairs:

- *All* gene-disease pairs listed in DisGeNET, obtained using a manual and TM methods (approx. 430,000 pairs)
- *Manually curated* pairs only (approx. 33,000 pairs)

Whilst the manually curated pairs are expected to be of high quality, the use of text-mined pairs provides scope for collecting additional evidence about how relations are described in text. In DisGeNET, each gene-disease pair is represented as a pair of concept identifiers, referring to entries in domain specific databases, i.e., NCBI Entrez Gene [52] for genes and the UMLS Metathesaurus [53] for diseases.

To find textual evidence about how all DisGeNET gene-disease relations are described in the 2015 abstracts, we made use of the gene and disease mentions automatically recognised in MEDLINE abstracts [49]. We subsequently used Pubtator mappings [54] to associate each of these mentions with a suitable database identifier. For each DisGeNET gene-disease pair, we could then collect all sentences from the abstracts in which the gene and disease are mentioned together.

5.3 Restriction Based on Common Nodes

In Fig. 1, the common node in the path between the gene and the disease is *associated*, which is the type of relation-indicating word that has been used as a filter for gene-disease relations in several previous studies. Therefore, we investigated whether relation extraction performance could be improved by placing restrictions on which words can appear at the common node.

We collected a list of all words (3,392 in total) appearing at the common node of all of textual mentions of DisGeNET gene-disease relations that occur in the 2015 abstracts. We then evaluated the effect of requiring that the common node should correspond to one of these words in order to be classified as a valid association (see Table 3). Given the large number of words that appear rarely as the common node (over 2,000 of the words occur 5 times or less), we also assessed the impact of applying a frequency threshold to filter out rare words.

For all categories of sentences, the use of *all words* from 2015 abstracts results in a modest increase in precision compared to the baseline methods, demonstrating a slight filtering ability. Although recall is quite high in most cases (suggesting that most relations are described using fairly fixed vocabulary), it is still lower than the baselines, and a small improvement in F-Score (0.5) over either of the baselines is only observable for SG sentences. Frequency-based filtering of common node words results in further small increases in precision for SD and MB sentences, but a corresponding significant drop in recall.

Table 3: Common node dependency path restriction

	All words			Words with frequency > 5		
	P	R	F	P	R	F
SB	90.7	93.3	92.0	89.6	81.9	85.6
SG	85.4	96.3	90.6	83.6	83.6	83.6
SD	83.8	96.0	89.5	85.0	91.5	88.1
MB	71.8	90.7	80.1	73.4	83.0	78.1
Total	82.1	94.1	87.7	82.7	86.3	84.5

5.4 Pattern-Based Dependency Restrictions

We next collected all of the unique dependency *patterns* that connect DisGeNET gene-disease pairs in sentences from the 2015 abstracts. We represent the path pattern between the gene and the disease in Fig 1. as follows:

prep_of ➔*prep_with* ➔*VBD* ←*nsubjpass*

Starting from the left, the pattern consists of the labels on the route through the dependency tree leading from the gene to the common node; the labels on the route from the disease start from the right of the pattern. The common node is represented by its part-of-speech, in this case a past tense verb (VBD).

We collected separate sets of patterns for gene-disease pairs originating from both DisGeNET lists (i.e., all pairs and curated pairs). The counts of unique patterns were as follows:

- All DisGeNET gene-disease pairs – 70,229 patterns
- Curated DisGeNET gene-disease pairs – 21,105 patterns

We subsequently applied these patterns to the augmented EU-ADR corpus, extracting relations only in cases where the dependency path between the gene and the disease matched one of the extracted patterns. The results are shown in Table 4.

Table 4: Evidence-based dependency pattern extraction

	Patterns from **all** DisGeNET pairs			Patterns from **curated** DisGeNET pairs		
	P	R	F	P	R	F
SB	94.1	60.9	74.0	97.9	44.8	61.4
SG	81.8	32.7	46.8	88.8	29.1	43.8
SD	81.2	41.2	54.7	82.8	31.7	45.8
MB	**94.8**	46.6	62.5	**94.4**	43.2	59.2
Total	88.0	45.9	60.3	90.3	37.1	52.6

The precision values obtained are mostly higher than when common node restrictions are used, demonstrating a superior filtering ability of the patterns. Most striking is the precision increase of around 23% for MB sentences, compared to the baselines, showing that the patterns are particularly effective in separating out the individual relations expressed in these complex sentences. The slightly higher precision values when patterns are generated from curated relations highlight the possible advantages of using these as a starting point.

The major disadvantage of using fine-grained dependency patterns to restrict relation extraction is the very low recall. The varying path lengths, using combinations of the 50+ different dependency labels in the Stanford scheme, result in an immense potential number of unique patterns, which cannot be accounted for even when using evidence from 1 million abstracts.

5.5 Generalising Dependency Patterns

To address the high variability of unique dependency patterns using fine-grained labels, we generalised the patterns, to allow them to cover a wider range of cases, and hopefully to increase recall. We performed 2 levels of generalisation, which we term *simple* generalisation and *hierarchy-driven* generalisation. For simple generalisation, we applied the following steps:

- In relation labels that are "specialised" with specific words, these specific words were removed (e.g., *conj_and* is generalised to *conj*, *prep_in* is generalised to *prep*, etc.)
- Any identical, consecutive labels in the path are collapsed into a single label (e.g., *prep*➔*prep* is collapsed to *prep*).
- Three character part-of-speech tags at common nodes were generalised to two-character tags (e.g., *VBD, VBP* and *VBZ* are all different forms of verbs, which we generalise to *VB*).

Hierarchy-driven generalisation exploits the hierarchical structure of Stanford dependency labels. We focus specifically on generalising the *argument* and *modifier* branches of the hierarchy, some examples of which are provided in Table 5.

Table 5: Argument and modifier examples

Type	Description	Examples
Argument	Completes the meaning of a verb	**Subject** - ***ChIP sequencing*** *reveals novel binding targets*
		Direct Object - *CC-122 binds* ***CRBN***
		Adjectival complement - *Jugular venous thrombosis could be* ***secondary to malignancy***
Modifier	Describes a phrase to make it more specific	**Adjectival modifier** - ***infective*** *endocarditis*
		Adverbial modifier - ***aberrantly*** *upregulated*
		Prepositional modifier - *19% of ILI patients had died* ***from melanoma***

In addition to the steps of simple generalisation, hierarchy-driven generalisation collapses all labels falling under a particular branch to a single label. Our experiments (results shown in Table 6) collapsed different combinations of hierarchy branches, and at different levels of granularity. The following abbreviations refer to the different levels of generalisation:

- **MOD** – modifier labels (approx. 20) generalised to MOD
- **ARG** - argument labels (approx. 15) generalised to ARG
- **SUBJ** - the 4 labels falling under the subject category (a sub-category of ARG) generalised to SUBJ
- **OBJ** - the 3 labels falling under the object category (a sub-category of ARG) generalised to OBJ

All types generalisation have a positive impact on recall, which generally increases with the degree of generalisation. However, increased recall comes at the cost of decreased precision - the more general the patterns become, the less discriminative they appear to be in terms of identifying valid gene-disease relations *only*. Although precision is least affected by *simple* and *ARG* generalisation (the latter suggesting that there is less variation in fine-grained argument relations compared to fine-grained modifier relations), the lower recall levels compared to other generalisations are problematic. In contrast, all cases involving MOD generalisation achieve far better recall rates (up to 25% for MB sentences). Indeed, the most extreme generalisations (MOD-ARG) achieve recall that is almost the same as for unrestricted dependency paths, providing

Table 6: Results of applying different pattern generalisation approaches

		Simple			ARG			MOD			MOD-ARG			MOD-SUBJ-OBJ		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
All	SB	91.3	89.5	90.4	91.6	93.3	92.5	90.2	96.2	93.1	90.3	97.1	93.6	90.2	96.2	93.1
	SG	87.8	78.2	82.7	86.0	78.1	81.9	82.8	96.4	89.1	82.8	96.4	89.1	82.8	96.4	89.1
	SD	81.5	75.4	78.3	80.1	78.9	79.5	81.1	92.5	86.4	81.6	96.0	88.2	81.1	93.0	86.7
	MB	86.0	67.8	75.8	81.3	73.7	77.3	75.3	98.3	85.3	75.3	98.3	85.3	75.3	98.3	85.3
	Tot	85.5	76.9	81.5	83.7	80.7	82.2	81.5	95.2	87.8	81.8	96.7	88.7	81.5	95.4	87.9
	# Patts	31,946			22,532			8,465			3,088			7,194		
Curat.	SB	91.5	81.9	86.4	92.1	88.6	90.3	90.0	94.2	92.1	90.3	97.1	93.6	90.0	94.2	92.0
	SG	86.4	69.0	76.8	86.7	70.9	78.0	82.8	96.4	89.1	82.8	96.4	89.1	82.8	96.4	89.1
	SD	83.8	70.4	76.5	81.4	74.9	78.0	81.6	88.9	85.1	81.0	92.0	86.1	81.9	91.0	86.2
	MB	85.6	62.5	74.0	80.1	72.0	76.2	78.8	91.5	84.7	75.5	98.3	85.2	**80.0**	**98.3**	**88.2**
	Tot	86.4	71.5	78.2	84.3	76.7	80.3	82.8	91.6	87.0	81.5	95.2	87.8	83.1	94.1	88.3
	# Patts	11,169			8,569			3,331			1,435			2,904		

strong evidence that generalised patterns can identify nearly all gene-disease relations.

The problems in balancing precision and recall mean that there appear to be no clear advantages in using any kind of restricted dependency paths over the simpler baseline approaches for SB, SD and SG sentences. Hence, we have not carried out any further experiments using these sentences.

However, the benefits of using restricted dependency paths for MB sentences are clear. The highest precision of 80% for MB sentences, obtained using MOD-SUBJ-OBJ generalisation of patterns generated using curated gene-disease pairs, represents an increase in precision over the baselines of over 8%, without any decrease in recall. Accordingly, the F-Score for such relations is increased by around 5%. This result also provides further evidence of the higher quality nature of the patterns generated from curated pairs, especially as we converge towards a smaller set of general patterns, as well as suggesting that a slightly finer-grained generalisation of ARG labels, which distinguishes subjects from objects, is advantageous.

5.6 Filtering Generalised Patterns

We investigated whether filtering of the best performing generalised patterns (i.e., MOD-SUBJ-OBJ patterns generated from curated gene-disease pairs) could help to isolate the most reliable relation-denoting patterns, and hence improve extraction accuracy. We applied three different filtering techniques, and assessed their ability to improve relation extraction performance in MB sentences (as shown in Table 7):

- **Pattern-based frequency filtering (PBF)** – only those patterns occurring above a certain frequency threshold in the 2015 abstract set are considered to denote valid relations.
- **Relation-based frequency filtering (RBF)** – only patterns occurring between gene-disease pairs that are mentioned in the 2015 abstract set above a certain threshold are considered to denote valid relations.
- **Path length filtering (PL)** – only patterns whose length is below a certain threshold are considered to denote valid relations.

Table 7: Effects of filtering generalised dependency patterns

Filtering method	P	R	F	# Patts
PBF >= 5	82.7	97.4	89.5	435
PBF >= 10	82.4	87.3	84.8	226
RBF>=25	80.6	**98.3**	88.5	1,912
RBF>=50	81.7	**98.3**	89.2	1,578
RBF >=100	81.7	**98.3**	89.2	1,252
PL <= 6	**86.3**	85.6	85.9	1,840
PL <= 7	83.3	97.5	**89.8**	2,420
PL <= 8	80.5	**98.3**	88.5	2,725

All filtering techniques improve upon the 80% precision achieved for the unfiltered patterns, with the highest precision being achieved using PL<=6. However, a large drop in recall compared to the unfiltered patterns shows that many valid paths are longer than this. Although using PL<=7 appears to be the optimal filtering technique, several other filtering techniques offer comparable levels of performance. It is interesting to note that PBF significantly reduces the number of patterns used and yet, when using patterns that occur 5 or more times, performance still remains high. This suggests that PBF is the most effective filter for high quality patterns. Whilst RBF achieves slightly lower precision, the higher recall compensates for this. Most filtering techniques have little negative effect on recall, showing that they can effectively remove patterns not needed for accurate relation extraction.

Based on the positive results obtained for individual filtering techniques, we combined the best performing settings, to try to further boost performance. The results are shown in Table 8.

Table 8: Effects of combining filtering techniques

Filtering methods	P	R	F	# Patts
PL <= 7, PBF > =5	84.4	96.6	90.1	**430**
RBF>=100, PBF >=5	**84.7**	79.7	82.1	169
PL <= 7, RBF >= 100	84.6	**97.5**	**90.6**	1,105

Whilst the combination of RBF and PBF filtering appears to remove too many potentially useful patterns, the importance of PL as a filter is further reinforced through its ability to slightly refine the sets of patterns obtained through either RBF or PBF, with only minimal loss of recall. Whilst best the filtering technique combines PL and RBF, the combination of PL with PBF achieves comparable performance, using less than half the patterns. Thus, high performance can be achieved with only 430 generalised patterns (compared to the original 21,105 fine grained patterns).

6 COMBINING EXTRACTION APPROACHES

In Table 9, we show the results of applying our proposed "optimal" approach for relation extraction to the augmented EU-ADR corpus, and compare this to the co-occurrence baseline. For SB, SG and SD sentences, we use this same baseline, since none of the experiments involving dependency paths resulted in any significant improvements in extraction performance. However, for MB sentences, we use the best performing filtered, generalised dependency patterns, as these achieve a 13% increase in precision and an 8.3% increase in F-score over the baseline for MB sentences, leading to a 3.4% increase in overall precision when considering the corpus as a whole.

Table 9: Combining co-occurrence and dependency patterns

	Final method			Baseline		
	P	R	F	P	R	F
SB	90.5	100	95.0	90.5	100	95.0
SG	83.3	100	90.1	83.3	100	90.1
SD	79.9	100	88.8	79.9	100	88.8
MB	84.6	97.6	90.6	71.6	98.3	82.3
Total	83.6	99.4	90.8	80.2	100	89.0

Although we can contrast our results with those obtained in previous studies, the use of different evaluation corpora in each case makes a direct comparison impossible. Whilst 94% precision is reported for co-occurrence based extraction in [20], their randomly selected sentences may not fully account for sentences of different complexities. In contrast, the 75.9% precision and 84.1% F-score achieved using co-occurrence extraction in [37] seem to better reflect the difficulties in achieving high accuracy using this simple method in isolation. A similar level of overall performance was achieved using the ML-driven dependency approach in [18], i.e., P 75.1%, R 97.7%, F 84.6%. Whilst the recall is similar to our dependency-based experiments, our evidence-based dependency paths appear to achieve greater precision. Overall, the evidence suggests that our selective combination of dependency patterns with co-occurrence offers advantages over methods that apply only a single approach.

7 LARGE-SCALE RELATION EXTRACTION

To assess the performance of our optimal extraction strategy on a much larger collection of documents, we compared our results to those obtained by BeFree [18] on approximately 74,000 abstracts from 2015. We obtained a file of relations extracted by BeFree from the DisGeNET website, which includes the PMIDs of all abstracts containing evidence for each gene-disease relation, along with one sentence from each abstract that contains the relation, in which the exact text spans corresponding to the gene and disease mentions are identified.

We wanted to be able to compare the performance of BeFree with our own approach on a common set of abstracts and ideally, using a common set of recognised gene and disease mentions, which would help to avoid any potential bias introduced by different NER methods. We retrieved the same set of around 74,000 abstracts processed by BeFree by using the same PUBMED query provided on the DisGeNET website used to obtain a set of abstracts focussed on human diseases and their associated genes for subsequent processing by BeFree, i.e.,

("Psychiatry and Psychology Category"[Mesh] AND "genetics"[Subheading]) OR ("Diseases Category"[Mesh] AND "genetics"[Subheading]) AND (hasabstract[text] AND "humans"[MeSH Terms] AND English[lang])

It was more challenging to try to ensure that our method had access to the same set of gene and disease mentions used as the starting point for BeFree relation extraction. This is because the full set of gene and disease NER results obtained from the BioNER module [55] of BeFree for all abstracts is not made freely available. Rather, the BeFree relations file only shows NER results for *selected* sentences of certain abstracts. We therefore attempted to approximate the output of BioNER, by taking all gene and disease mentions for each abstract provided in [49], and filtering out any mentions whose span did not correspond exactly to one of the BioNER-recognised spans within the evidence sentences of the BeFree relations file.

We then applied our relation extraction approach to mentions of the 4,065 genes and 2,763 diseases that remained after applying the above NE filtering step to the 74,000 abstracts. Our comparison (see Table 10) covers both the number of unique associations (i.e., gene-disease pairs) identified, and the number of abstracts in which *evidence* was found for these associations.

The overall number of relations recognised by each method is very similar. However, the degree of overlap shows that, whilst many of the same relations are detected by both methods, a significant proportion of relations is only recognised by one or other of the methods. This provides evidence that the different

Table 10: Comparison of BeFree with our method

Comparison type	Method	Count
Gene-disease associations detected	BeFree	12018
	Our method	**12130**
	Overlap	8957
Pieces of evidence detected	BeFree	22785
	Our method	**28704**
	Overlap	15926

approaches produce results that can complement each other. Indeed, a closer analysis of the non-overlapping relations found by our method reveals that 1,392 are potentially *novel* relations that do not appear in the DisGeNET database at all, while 1,586 correspond to relations that *are* listed in DisGeNET, but which were not found by BeFree in the 2015 abstract subset. Examples of relevant sentences containing novel relations identified by our method demonstrate its ability to recognise valid relations:

- *Our data establish **YB-1** as a critical regulator of hypoxia-inducible factor 1alpha (HIF1alpha) expression in **sarcoma** cells.*
- ***Phosphoglycerate dehydrogenase** is likely to be associated with tumorigenesis and may be a potential prognostic marker for **CIN** progression*

The fact that the above sentences, and many other sentences denoting novel relations detected by our method, mention only a single gene and disease (and hence are detected using co-occurrence), reinforces the advantages of using co-occurrence as well as dependency relations. On the other hand, of the 44,802 sentences within the abstract subset set that contain at least one gene and one disease, 13,447 (i.e., 30%) contain both multiple gene and multiple disease mentions, highlighting the importance of the dependency-based approach.

Our use of restricted dependency patterns can filter out some incorrect relations that are recognised by BeFree, as in the following example, for which BeFree incorrectly detects a relation between *lung tumors* and *JAK2*:

*Some of these mutant genes (such as BAG6, SPEN and WISP3) are recognized as major cancer players in **lung tumors**; others have been previously identified in other human cancers (**JAK2**, TCEB3C, NELFE, TAF1B, EBLN2).*

On the other hand, the fact that BeFree can complement the results of our method is evidenced in its ability to recognise certain relationships in sentences that have complex or unusual structures, which are not detected by our patterns, e.g.,

*Silent information regulator-2 (**Sir-2**) proteins, or sirtuins, are a highly conserved protein family of histone deacetylases that promote longevity by mediating many of the beneficial effects of calorie restriction which extends life span and reduces the incidence of cancer, **cardiovascular disease** (CVD), and diabetes*

Another important finding from our comparison is that, although the two methods detect roughly equal numbers of unique associations, our approach can detect significantly more *evidence* for these associations in different abstracts. This is important for our own ultimate aims, since we want to collect as much evidence as possible for each relation, to allow us to detect different types of interpretative and contextual information provided in different sentences that mention the relation. In terms of interpretative information, our future work will allow relations to be filtered according to whether they represent factual knowledge or experimental analyses etc., or to allow "tracking" of associations over time, e.g., to examine transitions from initial hypotheses to accepted associations that are backed by experimental evidence. Contextual information will include details such as risk factors (e.g., smoking) that interact with or impact upon gene-disease associations, or the specific population subgroups in which an association has been found to occur. In this respect, our method can be valuable in allowing additional details to be found. For example, whilst there are 26 sentences from abstracts in DisGeNET providing evidence of associations between *cagA* and *gastritis*, none mention the specific link with Iranian children, found in an additional sentence retrieved by our method, i.e., *vacAs1 and **cagA** are associated with more severe **gastric inflammation** in **Iranian children**.*

8 CONCLUSIONS

In this paper, we have described our approach to detecting associations between genes and diseases mentioned in literature, as a first step towards developing a sophisticated search system to facilitate the efficient location of various types of evidence relating to biomarkers. Our novel technique combines relation extraction methods of varying sophistication, according to the complexity of sentences, and our use of evidence-based dependency patterns, which have been carefully generalised and filtered to obtain maximum accuracy, can improve extraction precision in sentences containing multiple gene and disease mentions by 13% compared to simple co-occurrences, with minimal loss of recall. Comparison of the output of our method with that of a related method (BeFree) on a large dataset revealed that our approach can identify many potentially novel gene-disease relationships, and is particularly effective in identifying large amounts of supporting textual evidence.

As future work, we intend to recognise mentions of further types of concepts and develop methods to link them with the recognised gene-disease relationships, in order to construct more complex, structured representations of biomarker-related knowledge that can be queried in complex ways or used to populate biomarker databases. We will also extend upon previous work (e.g., [56]) to allow various types of interpretative information about relations to be detected automatically.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

ACKNOWLEDGEMENTS

This work has been supported by the EPSRC and MRC (MMPathIC project, Grant. No. MR/N00583X/1)

REFERENCES

[1] Strimbu, K. and J.A. Tavel, *What are biomarkers?* Current Opinion in HIV and AIDS, 2010. **5**(6): p. 463.

[2] Deyati, A., E. Younesi, M. Hofmann-Apitius, and N. Novac, *Challenges and opportunities for oncology biomarker discovery.* Drug discovery today, 2013. **18**(13): p. 614-624.

[3] Campos, D., S. Matos, and J.L. Oliveira, *Gimli: open source and high-performance biomedical name recognition.* BMC bioinformatics, 2013. **14**(1): p. 1.

[4] Wei, C.-H., H.-Y. Kao, and Z. Lu, *GNormPlus: an integrative approach for tagging genes, gene families, and protein domains.* BioMed research international, 2015. **2015**.

[5] Bhasuran, B., G. Murugesan, S. Abdulkadhar, and J. Natarajan, *Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases.* Journal of Biomedical Informatics, 2016. **64**: p. 1-9.

[6] Pyysalo, S. and S. Ananiadou, *Anatomical entity mention recognition at literature scale*. Bioinformatics, 2014. **30**(6): p. 868-875.

[7] Krallinger, M., et al., *CHEMDNER: The drugs and chemical names extraction challenge*. Journal of cheminformatics, 2015. **7**(1): p. 1.

[8] Leaman, R., C.-H. Wei, and Z. Lu, *tmChem: a high performance approach for chemical named entity recognition and normalization*. Journal of cheminformatics, 2015. **7**(1): p. 1.

[9] Wei, C.-H., B.R. Harris, H.-Y. Kao, and Z. Lu, *tmVar: a text mining approach for extracting sequence variants in biomedical literature*. Bioinformatics, 2013: p. btt156.

[10] Piñero, J., et al., *DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants*. Nucleic Acids Research, 2016: p. gkw943.

[11] Bundschus, M., M. Dejori, M. Stetter, V. Tresp, and H.P. Kriegel, *Extraction of semantic biomedical relations from text using conditional random fields*. BMC Bioinformatics, 2008. **9**: p. 207.

[12] Chun, H.W., et al., *Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts*. BMC Bioinformatics, 2006. **7 Suppl 3**: p. S4.

[13] Islam, M.T., M. Shaikh, A. Nayak, and S. Ranganathan. *Biomarker information extraction tool (BIET) development using natural language processing and machine learning*. in *Proceedings of the International Conference and Workshop on Emerging Trends in Technology*. 2010. ACM.

[14] Malhotra, A., E. Younesi, S. Bagewadi, and M. Hofmann-Apitius, *Linking hypothetical knowledge patterns to disease molecular signatures for biomarker discovery in Alzheimer's disease*. Genome medicine, 2014. **6**(11): p. 97.

[15] Nedellec, C., *Learning Language in Logic - Genic Interaction Extraction Challenge*, in *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, J. Cussens and C. Nedellec, Editors. 2005. p. 31--37.

[16] Kim, J.-D., T. Ohta, S. Pyysalo, Y. Kano, and J.i. Tsujii, *Extracting Bio-molecular Event From Literature—The BioNLP'09 Shared Task*. Computational Intelligence, 2011. **27**(4): p. 513-540.

[17] Kim, J.D., et al., *Overview of BioNLP Shared Task 2011*, in *Proceedings of BioNLP Shared Task 2011 Workshop*. 2011. p. 1-6.

[18] Bravo, À., J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L.I. Furlong, *Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research*. BMC bioinformatics, 2015. **16**(1): p. 55.

[19] Verspoor, K.M., G.E. Heo, K.Y. Kang, and M. Song, *Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts*. BMC medical informatics and decision making, 2016. **16**(1): p. 68.

[20] Chun, H.W., et al., *Extraction of gene-disease relations from Medline using domain dictionaries and machine learning*. Pac Symp Biocomput, 2006: p. 4-15.

[21] Singhal, A., M. Simmons, and Z. Lu, *Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature*. Journal of the American Medical Informatics Association, 2016. **23**(4): p. 766-772.

[22] Lee, K., et al., *BRONCO: Biomedical entity Relation ONcology COrpus for extracting gene-variant-disease-drug relations*. Database: the journal of biological databases and curation, 2016. **2016**.

[23] Craven, M. and J. Kumlien. *Constructing biological knowledge bases by extracting information from text sources*. in *ISMB*. 1999.

[24] Doughty, E., et al., *Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature*. Bioinformatics, 2011. **27**(3): p. 408-415.

[25] Van Mulligen, E.M., et al., *The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships*. Journal of biomedical informatics, 2012. **45**(5): p. 879-884.

[26] Mahmood, A.A., T.-J. Wu, R. Mazumder, and K. Vijay-Shanker, *Dimex: A text mining system for mutation-disease association extraction*. PloS one, 2016. **11**(4): p. e0152725.

[27] Verspoor, K., et al., *Annotating the biomedical literature for the human variome*. Database, 2013. **2013**: p. bat019.

[28] Chang, J.T. and R.B. Altman, *Extracting and characterizing gene–drug relationships from the literature*. Pharmacogenetics and Genomics, 2004. **14**(9): p. 577-586.

[29] Nobata, C., et al., *Kleio: a knowledge-enriched information retrieval system for biology*, in *Proceedings of the 31st Annual International ACM SIGIR*. 2008: Singapore. p. 787-788.

[30] Tsuruoka, Y., J. Tsujii, and S. Ananiadou, *FACTA: a text search engine for finding associated biomedical concepts*. Bioinformatics, 2008. **24**(21): p. 2559-60.

[31] Ongenaert, M., et al., *PubMeth: a cancer methylation database combining text-mining and expert annotation*. Nucleic acids research, 2008. **36**(suppl 1): p. D842-D846.

[32] Younesi, E., et al., *Mining biomarker information in biomedical literature*. BMC medical informatics and decision making, 2012. **12**(1): p. 148.

[33] Korhonen, A., et al., *Text mining for literature review and knowledge discovery in cancer risk assessment and research*. PloS one, 2012. **7**(4): p. e33427.

[34] Cheng, D., et al., *PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites*. Nucleic acids research, 2008. **36**(suppl 2): p. W399-W405.

[35] Liu, Y., Y. Liang, and D. Wishart, *PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more*. Nucleic acids research, 2015. **43**(W1): p. W535-W542.

[36] Tsuruoka, Y., M. Miwa, K. Hamamoto, J.i. Tsujii, and S. Ananiadou, *Discovering and visualizing indirect associations between biomedical concepts*. Bioinformatics, 2011. **27**(13): p. i111-i119.

[37] Hakenberg, J., et al., *A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions*. Journal of biomedical informatics, 2012. **45**(5): p. 842-850.

[38] Garten, Y. and R.B. Altman, *Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text*. BMC bioinformatics, 2009. **10**(2): p. S6.

[39] Plake, C., T. Schiemann, M. Pankalla, J. Hakenberg, and U. Leser, *AliBaba: PubMed as a graph*. Bioinformatics, 2006. **22**(19): p. 2444-2445.

[40] Rindflesch, T.C., B. Libbus, D. Hristovski, A.R. Aronson, and H. Kilicoglu. *Semantic relations asserting the etiology of genetic diseases*. in *AMIA*. 2003.

[41] Masseroli, M., H. Kilicoglu, F.-M. Lang, and T.C. Rindflesch, *Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease*. BMC bioinformatics, 2006. **7**(1): p. 291.

[42] Greco, I., et al., *Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation*. Journal of translational medicine, 2012. **10**(1): p. 217.

[43] Moreno, A., D. Isern, and A.C.L. Fuentes, *Ontology-based information extraction of regulatory networks from scientific articles with case studies for Escherichia coli*. Expert Systems with Applications, 2013. **40**(8): p. 3266-3281.

[44] Hara, T., Y. Miyao, and J.-i. Tsujii, *Evaluating the impact of re-training a lexical disambiguation model on domain adaptation of an HPSG parser*, in *Trends in Parsing Technology*. 2010, Springer. p. 257-275.

[45] McClosky, D., *Any domain parsing: automatic domain adaptation for natural language parsing*. 2010.

[46] Ravikumar, K.E., K.B. Wagholikar, D. Li, J.-P. Kocher, and H. Liu, *Text mining facilitates database curation-extraction of mutation-disease associations from Bio-medical literature*. BMC bioinformatics, 2015. **16**(1): p. 185.

[47] Coulet, A., N.H. Shah, Y. Garten, M. Musen, and R.B. Altman, *Using text to build semantic networks for pharmacogenomics*. Journal of biomedical informatics, 2010. **43**(6): p. 1009-1019.

[48] Piñero, J., et al., *DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes*. Database, 2015. **2015**: p. bav028.

[49] Hakala, K., S. Kaewphan, T. Salakoski, and F. Ginter, *Syntactic analyses and named entity recognition for PubMed and PubMed Central—up-to-the-minute*. ACL 2016, 2016: p. 102.

[50] Charniak, E. and M. Johnson. *Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking*. in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 2005.

[51] de Marneffe, M.-C., B. MacCartney, and C.D. Manning. *Generating Typed Dependency Parses from Phrase Structure Parses*. in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. 2006.

[52] Maglott, D., J. Ostell, K.D. Pruitt, and T. Tatusova, *Entrez Gene: gene-centered information at NCBI*. Nucleic Acids Research, 2011. **39**(suppl 1): p. D52-D57.

[53] Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology*. Nucleic Acids Research, 2004. **32**: p. 267-270.

[54] Wei, C.-H., H.-Y. Kao, and Z. Lu, *PubTator: a web-based text mining tool for assisting biocuration*. Nucleic acids research, 2013: p. gkt441.

[55] Bravo, A., M. Cases, N. Queralt-Rosinach, F. Sanz, and L. Furlong, *A knowledge-driven approach to extract disease-related biomarkers from the literature*. BioMed research international, 2014. **2014**.

[56] Miwa, M., P. Thompson, J. McNaught, D.B. Kell, and S. Ananiadou, *Extracting semantically enriched events from biomedical literature*. BMC Bioinformatics, 2012. **13**(1): p. 108.