COMP90055 Computing Project Report

SurName: Teng

Given Name: Ruichen

Student Number: 678693

University Email: tengr@student.unimelb.edu.au

Name of Degree Enrolled in: Master of Information Technology — Computing

Subject Code: COMP90055 Computing Project

Total credit points for entire project: 25

Type of Project: Research project

Semester in which project commenced: Semester 1, 2015

Semester in which project is expected to complete: Semester1, 2015

Project Title: Information Extraction of biomedical relationships in published colon cancer literature

Supervisor: Prof. Karin Verspoor

Information Extraction of biomedical relationships in published colon cancer literature

Ruichen Teng

Department of Computing and Information Systems

University of Melbourne

This dissertation is submitted for the degree of

Master of Information Technology

June 2015

To my loving parents . . .

Declaration

I certify that

- this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.

- where necessary I have received clearance for this research from the University's Ethics Committee (Approval Number ....) and have submitted all required data to the Department

- the thesis is 8000 words in length (excluding text in images, table, bibliographies and appendices).

Ruichen Teng
June 2015

Acknowledgements

And I would like to acknowledge ...

Abstract

This is where you write your abstract ...

Table of contents

List of figures

List of tables

Introduction

1.1 Motivation

Text mining is the process of searching for patterns in natural language text using methods in computer science, linguistics, and statistics. Despite being unstructured and only human-understandable, text is still our primary media for exchange of information[22]. The prevalence of textual data presents a big challenge to computer-driven natural language understanding. *Information extraction*, in particular, refers to the task of acquiring organized, structured and queryable format of data from the unstructured corpus.

While text mining is widely used in areas like marketing and document verifying, it has received increased attention for its application to biomedical literatures[1, 10, 11]. This trend stems from the direct need of biomedical workers and researchers to cope with information explosion in their field. For instance, MEDLINE(Medical Literature Analysis and Retrieval System Online), the online database of United States National Library of Medicine, has accumulated nearly 0.8 million citations and 2.7 billion searches in 2014 alone[14], with total citations reaching 25 million. Within these publications there are valuable research results that should add to human knowledge. In the meantime, our primary knowledge base in life science - the biomedical databases, are still mostly being populated manually by *biocurators* - the "museum catalogers of the Internet age"[21]. They are professional scientists who read biomedical articles, record relevant data and organize them according to the biomedical database schema. The sheer volume of publications has made this process increasingly unrealistic[7].

Not only does data overload make knowledge discovery demanding, it also leads to a decline in literature quality. Nowadays biomedical workers and researchers are more prone to drawing wrong conclusions because they simply can not read all the relevant publications, among which oftentimes contradicting results are reported. Needless to say, we are

in desperate need of automatic tools for systematically analyzing documents and extracting information. In fact, it has been argued that text mining is required to improve the coverage of databases[2].

1.2 Definitions and Assumptions

Below are a list of terms which I will use throughout this thesis, detailed examples will be given in the relevant sections, but a general definition is first given here to avoid any confusions that might arise when reading the following paragraphs.

- *Relation Extraction:* In general, relation extraction refers two the process of discovering a relationship between entities in text. In the domain of natural language processing, the relation can be semantic, syntactic, etc, with semantic relations being the most important for knowledge discovery. Relations can be unary, binary and complex with complex relations sometimes intermingled with the concept of events, In this project we are mainly concerned with binary relations as shown. Relation extraction is a form of information extraction where the semantic relations between entities are extracted. Specifically, in this project we focus on the relation extraction among other informations. Biomedical relations covers a wide range of knowledge in this field.

- *Event Extraction:* to be added

1.3 Research Question

This project looks to investigate the application of information extraction system Approximate Subgraph Matching (ASM)[**?**], on relation extraction tasks, specifically regarding the relation extraction on the variome corpus. In this project we focus on the relation extraction among other informations. Biomedical relations covers a wide range of knowledge in this field. As it turns out, this is not a trivial process, The relation extraction tool has a very promising applications for researchers and medical field workers, pharmaceutical companies, and the general public.

Relation extraction can be helpful in information search, knowledge discovery and hypothesis generation.

1.4 Thesis Structure

The remainder of the thesis is organized in the following manner. In Chapter 2, we discuss the related work of different algorithms used for relation extraction tasks including pattern-based methods, co-occurrence based methods, rule-based methods and kernel based methods. In Chapter 3, we discuss the Approximate Subgraph Matching(ASM) algorithm in detail and how the system previously used for event extraction in the BioNLP Shared Task 2013 was adapted for the relation extraction task for the Variome Corpus, together with evaluation of extraction results. In Chapter 4, we conclude our work and present future directions.

Related Work

This chapter will explore the research related to this thesis.

2.1 Relation Extraction

In general, relation extraction refers two the process of discovering a relationship between entities in text. In the domain of natural language processing, the relation can be semantic, syntactic, etc, with semantic relations being the most important for knowledge discovery. Relations can be uniary, binary and complext with complex relations sometimes intermingled with the concept of events, In this project we are mainly concerned with binary relations as shown.

2.2 Named Entity Recognition

Named Entity Recognition, or NER, is the task of identifying elements in text that belong to pre-defined categories like person, organization, etc. Specifically, NER in biomedical text mining aims at identifying thing like proteins, diseases, genes, etc. There is extra difficulty for NER in the biomeidical domain mainly because of the following reasons. quote: A survey of current work in biomedical text mining This task has been challenging for several reasons. First, there does not exist a complete dictionary for most types of biological named entities, so simple text- matching algorithms do not suffice. In addition, the same word or phrase can Recognising biological entities in text allows for further extraction of relationships and other information by identifying the key concepts of interest refer to a different thing depending upon context (eg ferritin can be a biological substance or a laboratory test). Conversely, many biological entities have several names (eg PTEN and MMAC1 refer the same gene). Biological entities may also have multi-word names (eg carotid artery), so the problem is additionally complicated by the need to determine name boundaries and resolve overlap of candidate names. Because of the potential utility and complexity of the problem, NER has

attracted the interest of many researchers, and there is a tremendous amount of published research in this topic. With the large amount of genomic information being generated by biomedical researchers, it should not be surprising that in the genomics era, much of the work in biomedical NER has focused on recognising gene and protein names in free text. quote: A survey of current work in biomedical text mining

2.3 Pattern Based Methods

Relation extraction started from the task of extracting protein protein interactions from text by pattern matching. First, a set of part-of-speech rules are applied to split the sentence into simple sentences, e.g. (P1 VB1 P2 VB2 CC P3) is splitted to P1 VB1 P2 and P1 VB2 P3. Next, a set of word patterns is applied to extract relations from these sentences. In addition, citeDiscovering patterns to extract protein–protein interactions from the literature: Part II proposed an idea of minimal description length, as in finding a a pattern set that has the most balance between high presion, short rule length/lower rule complexity by dynamic programming to optimize the rule set. Pattern based methods has achieved quite respectable performances, but it has the limitations that it does not generalize well to incorporate the richness of expressions, such as the anaphora terms like pronouns. In order to achieve that, huge amount of training data is needed.

2.4 Co-occurrence based methods

Finding co-occurring terms within a sentence or abstract has been the foundation of many relation extraction algorithms[3, 6, 9, 17, 18, 23]. Simple co-occurrence measures include probabilistic indicators such as point-wise mutual information, chi-square and log-likelihood ratio. More advanced measures include Concept Space, where thesaurus are mapped to a multi-dimensional Euclidean space[12, 19]. The main advantage of co-occurrence based methods is their simple implementation and low computational complexity. However, they often fail to capture and differentiate the nature of relations. Thus it is more suitable for detecting simple relations like gene-gene relations, but in in the general case of biological events.

2.5 Rule-Based Methods

2.6 Feature Based Methods

2.7 Kernel Based Methods

Kernels provide a similarity measure between two objects in some complex feature space. In contrast to feature based methods, kernel-based methods allow the original representation of the object to be retained and the kernel function will work out the similarity measure. For instance, a sentence maybe represented as a dependency graph, the feature based methods would want to select features such as number of nodes, edges, directions, etc, whereas kernel based methods allows feeding the two entire graph representations into a kernel function and output the distance measure.

2.8 Semi-Supervised Methods

Core

3.1 Data Collection

Our dataset is the Variome Corpus[20], which is openly accessible. [1] Verspoor et al. [20] gave a detailed illustration of the document selection and annotation process. I will summarize the main points here.

3.1.1 Background

A major part of the current biomedical research lies in understanding the relations between human genetic variation and disease phenotypes. The *Human Variome Project*, or *HVP*, is a global initiative to collect all genetic variation information affecting human health[15]. In particular, it acts as a liaison between individuals and organizations to integrate the genetic variants into databases that are open to the general public[20]. The *International Society for Gastrointestinal Hereditary Tumours (InSiGHT)*, is an international organization which aims to benefit patients with hereditary gastrointestinal(GI) tumours by research, education and personal assistance. In 2008, InSiGHT and HVP began a collaboration which propels InSiGHT to refine its process in the integration and interpretation of genetic variants. Consequently, a substantial effort was made to understand the mutation of mismatch repair(MMR) genes, the cause of Lynch Syndrome - one of the main syndromes of GI cancer[16]. A total of 10 full-text articles were selected from PubMed Central®by searching the common Lynch syndrome genes. These documents are mostly about inherited colon cancer. The annotation schema, also known as the Variome Annotation Schema[20], include 11 entity types and 13 relation types. as can be seen in the table 3.1.

In short, the corpus is Inspired by needs of inSIGHT database, but intended for broader applications. Documents relevant to the genetics of Lynch syndrome, which covers inherited

[1] http://www.opennicta.com.au/home/health/variome

Table 3.1 Variome Relation Types

Relation Type	Entity1	Entity2
relatedTo	mutation	disease
relatedTo	disease	gene
relatedTo	disease	body-part
has	gene	mutation
has	mutation	size
has	disease	characteristic
has	cohort-patient	age
has	cohort-patient	characteristic
has	cohort-patient	disease
has	cohort-patient	ethnicity
has	cohort-patient	gender
has	cohort-patient	mutation
has	cohort-patient	size

colon cancer as well as certain other cancers. Selected with PubMed Central To train tools for mining genetic variation and its relationship to disease Here in this information extraction task, we treat the manually annotated data as the *gold data*,

3.2 Dependency Graph and Shortest Path

The dependency graph of a sentence is a directed graph, where nodes represent sentence tokens, and edges indicate their semantic relations. Figure 3.1 shows the dependency graph of the sentence *" The p.Lys618Ala variant was co-existent with pathogenic mutations in two unrelated LS families."* generated by the Stanford Parser. Such a graph preserves the rich semantic structure of a sentence, and has been widely regarded as an informative way of presenting a sentence. A detailed explanation of the relative constituents in the graph can be found in [8]. However, the point is to transfer only human-readable sentences to a computer-understandable data structure. The general idea would be to feed this graph into a learning algorithm and classify relations based on similarity to the sentence graph in the training set. Different approaches exist for this process. Turku Event Extraction System (TEES)[2], for instance, engineers a feature vector which consists of token features(part-of-speech tags and character constituents for each word), sentence features(bag-of-words counts), and graph-based features(dependency path represented as N-grams) and builds a multi-class SVM model with the feature vector[4]. In this project, we decide to use the

[2]https://github.com/jbjorne/TEES

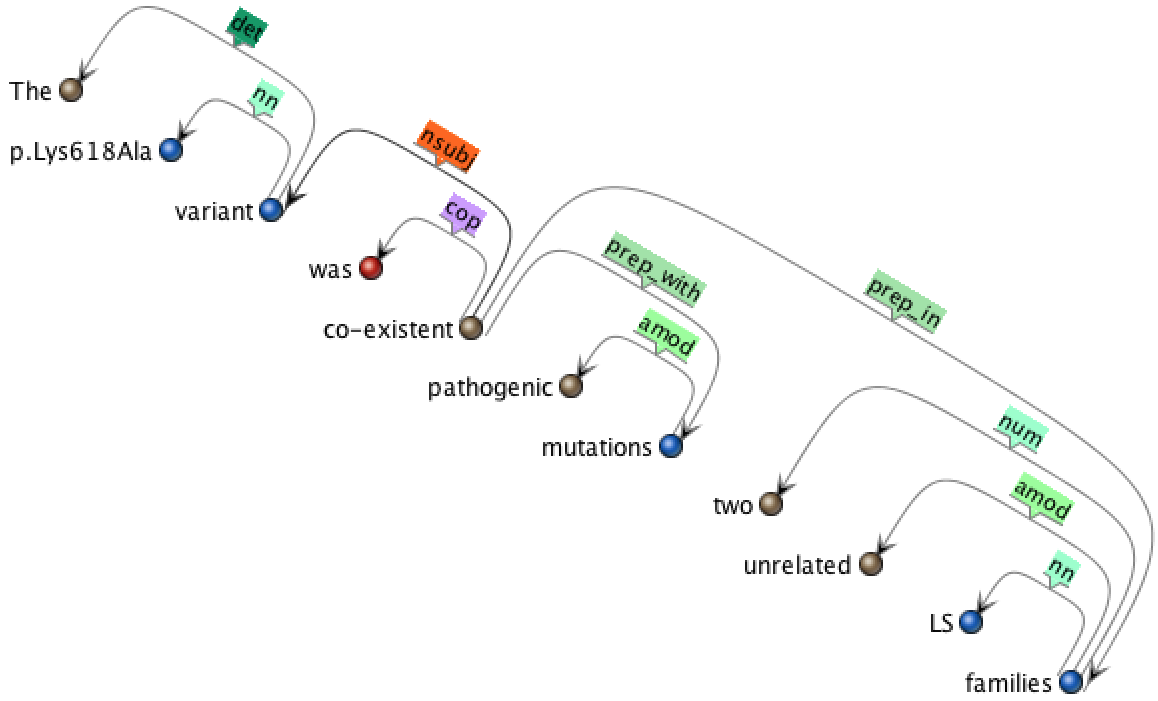

Fig. 3.1 Dependency Graph of " The p.Lys618Ala variant was co-existent with pathogenic mutations in two unrelated LS families."

Shortest Path Hypothesis[5], namely the heuristic that the relation between two entities in a sentence can be distilled from the shortest path between these entities in the undirected version of the sentence dependency graph. This effectively reduces the burden of feature engineering[13], but it also calls for high-quality training data. We believe that with effective parameter tuning and clever graph matching techniques, the shortest path can be a single standalone feature for a relation between two entities.

3.3 Approximate Subgraph Matching Algorithm

Disclaimer: As the system was designed for event extraction, it is more convenient to refer the system as being learning "events", while in essence events are nothing more than complex relations between entities. This is exactly the rationale behind adapting an event extraction system for relations extraction. What was later done in the adaptation process was treating relation as a type of "event". So "relations" and "events" are used (somewhat) interchangeably below.

The Approximate Subgraph Matching framework, proposed by Liu et al. [13] has the following work flow:

3.3.1 Prepossessing

To separate Named Entity Recognition from Relation Extraction problem, the named entity annotations are provided in training, development and test sets. Effectively, this is just asking computer to extract possible relations between these entities without worrying how to recognize them accurately. Of course, named entity recognition is an important step in a real relation extraction task as cite cite, because biological named entity recognition is such a complex problem and sometimes the relation extraction success would depend on the accuracy of NER.

the performance of the subgraph matching method, as an instance-based learning strategy (Alpaydin, 2004), is dependent on having good training examples that express the events in a range of syntactic structures, cite

3.3.2 Rule Learning

For each annotated sentence in the training set, a dependency graph is generated and the nodes representing entities are marked. The entity tokens are then replaced with their entity types, such that the rules learned are about an the generic entity types (e.g. gene mutation) instead of specific entities (e.g. p.Lys618Ala), so that our model has a better ability to generalize as we might not see the specific entities again in the test set. Next, the graph is transformed to its undirected version and the shortest path between entities are found with Dijkstra algorithm. This path is leaned as a rule associated with the event type in the annotation as a rule corresponding to that event type. Provided there is enough training data, a set of rules would be learned for each event type, representing the graph patterns that indicating a specific type of event.

3.3.3 Sentence Matching

The rules generated from the previous step could, in theory, be used to match sentences. For each sentence in the test set, a dependency graph can be generated together with the entity annotations(these are provided, as discussed). The graph can be searched exhaustively looking at the path(s) between each entity tokens, for an exact match with one or more rules within the rule set. This step is know as searching for exact subgraph isomorphism.

The above-mentioned matching approach would invariably lead to low recall, as the richness of language can almost always produce slightly different dependency graph structure representing exactly the same events between the same entities. This leads to the rationale behind approximate subgraph matching, which relaxes the matching process and allows for a penalty-based mismatch.

3.3.4 Rule Optimizing

To avoid learning the idiosyncrasies in the training data, an iterative rule set optimization process is executed. After the initial learning phase, each rule in the rule set is tested first on the training data to see if it get produce accurate enough predictions (above 0.25). The low performing rules are discarded as a consequence.

3.4 System Adaptation

As mentioned in 2.6, the project aims at adapting an existing system used for event-retrieval tasks on relation extraction tasks.

3.4.1 BioNLP Shared Tasks

BioNLP shared Task series is a community-wide text mining challenge specifically for biomedical literatures. The GE task in shared task 2013 aims at retireving events of the following format: The a1 files does (to be added), the a2 files does (to be added),

3.4.2 Variome Annotation Schema

During the duration of this project, most of my attempts to fully adapt the system to relation extraction tasks have failed. In essence, the differentiation in the retrieval process lies in the event retrieval relies on the detection and prediction of a trigger word, where as the relation extraction does not.

My only attempts that worked was transforming the annotation format of the Variome Corpus to that of the BioNLP Shared Task 2013, such that the original system would not break. The transformation is illustrated in this tables. The "has" relationship is transformed to "gene regulation" event, and "relatedTo" relationship is transformed to "Phosphyrilation" event, with the entity annotation slightly too. The entities would be of type "Protein" instead of the original entity types in the Variome Corpus. **The major bottleneck for this adaptation**

Table 3.2 Overall Result

Relation Type	Gold	Answer	Match	Precision	Recall	F1-score
has	1711	1310	402	0.3069	0.2350	0.2661
relatedTo	157	1498	36	0.0240	0.2293	0.0435
TOTAL	1868	2808	438	0.1560	0.2345	0.1874

work is that the original system includes a hard-coded trigger detection and prediction process. As for events detecting trigger words such as "activates" for gene expression is an important step for prediction. However, this process is not included at all in the event prediction process and to cope with the original system I had to add "fake triggers" for these events.
As a result, the first entity of each event/relation annotation is added as the trigger for the notation.

After this first attempt I had a few better ideas to add fake triggers, the best one being adding the entity types (patient-cohort), with parenthesis directly after the entity tokens and do a binding event. Due to the time constraints and these ideas being essentially fool's gold, I did not implement these ideas.

The results of my first attempt is shown in table 3.2.

3.5 Results

The overall result of the system is shown in table 3.2, the main reason the system is not performing well is that it is not predicting the triggers words correctly, because I have chosen to treat entity annotations as triggers.

Conclusion

4.1 Conclusion

We have explored the possibility of adapting an event-extraction system in relation extraction tasks for the Variome Corpus.

4.2 Contributions

The contributions of the research in this thesis are as follows:

- *The first attempt to use the system outside original system developers.* While this may seem like an easy task, a lot of the time for this project has been devoted to resolving dependencies, and debugging scripts to first have an end-to-end system, even for the original shared task data set.
- *The first attempt for literature mining of the Variome Corpus* To our knowledge, this is the first text mining attempt to the Variome Corpus.

4.3 Future Work

4.3.1 Full System Adaptation

Given the nature of the existing software system, I effectively wrote an adapter for the Variome Corpus, transforming the current data set to a format that the system is expecting. This is, of course, far from the best solution. However, system was developed solely for the purpose of event extraction tasks, and has numerous constraints and data format expectations as hard-coded strings and methods, to the extent that all of my efforts to change the code have failed. It is natural and understandable for academic software like this to be one-and-done for things like the shared task, yet I think it is in the interest of future system users and

developers to have a well-documented, adequately extensible system in place, with room for tweaking algorithms and file formats with parameters. Ideally, the shortest path kernel will expose interfaces for users to adjust what is happening under the hood with a few parameters. For instance, the existence of trigger should be optional and can be set with a parameter applied to the both rule learning process and event extraction process. Even in the shared task 2013, the existence of triggers for events like relations or coherence is optional. Next, annotation format should be parametrized, the system should establish a relationship between user-set entity and relation format and the entities and relations. Moreover, the named entity recognition could be a dispatch-able unit of the system too with options of a user-given entities or the output of a named-entity recognizer. A perfect scenarios would be that the user can input a configuration schema indicating annotation formats, entity types, entity output(whether from the annotation file or from a named entity recognizer), event/relations types and textual data. That way the system can be used for different kinds of tasks, possibly even beyond the domain of biomedical text mining and provide valuable feedback for the plausibility of the approximate subgraph matching paradigm.

4.3.2 Relation Type Fine-Tuning

Currently, the system only has two relations types, a “has” relation and a “relatedTo” relation. However, a patient-has-disease relation is semantically quite different from a gene-has-mutation relation, despite both of them being chunked to a single relation type called “has”. One might be tempted to have fine-grained relation types and incorporate the entity types into the relation, such as a separate “patient-cohort-has-disease” and a separate “gene-has-mutation” instead of a “has” relation for both. Nevertheless, this would cause the training set to become extremely sparse and the system learning behavior could change drastically. The correlation between granularity of event types and system performance has yet to be explored.

4.3.3 Parser Effect

This project uses the exact same preprocessing pipeline as citeExtracting Biomedical Events and Modifications Using Subgraph Matching with Noisy Training Data, namely the JULIE Sentence Boundary Detector, or JSBD (Tomanek et al., 2007), We then parse using a version of clearnlp1 (Choi and McCal- lum, 2013), a successor to ClearParser (Choi and Palmer, 2011), However, it concluded that changing the parser from the one originally in [13] has limited recall to an effect that can not be offset by increasing amount of training data, mainly

because the longer dependency graph produced by the clearnlp parser is harder to generalize. The parser effect regarding relation extraction task for the Variome Corpus is yet to be explained. The parser, too should be a loosely coupled component of the system such that the effect of different dependency parsers can be explored easily.

4.3.4 Parameter Tuning

The parameters such as subgraph weights, thresholds, rule set optimization aggressiveness for the training set and test them on the test set.

4.3.5 Kernel Combining

A significant limitation of the ASM based approach is the lower recall compared with other systems. Successful general literature mining at the semantic level might require a combination of many approaches. The shortest path kernel should be a loosely coupled component of the system and replaceable or by other kernels or a combination of kernels.

References

[1] Ananiadou, S., Kell, D. B., and Tsujii, J.-i. (2006). Text mining and its potential applications in systems biology. *Trends in biotechnology*, 24(12):571–579.

[2] Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., and Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–i48.

[3] Becker, K. G., Hosack, D. A., Dennis, G., Lempicki, R. A., Bright, T. J., Cheadle, C., and Engel, J. (2003). Pubmatrix: a tool for multiplex literature mining. *BMC bioinformatics*, 4(1):61.

[4] Björne, J. and Salakoski, T. (2011). Generalizing biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 183–191. Association for Computational Linguistics.

[5] Bunescu, R. C. and Mooney, R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 724–731. Association for Computational Linguistics.

[6] Chaussabel, D. and Sher, A. (2002). Mining microarray expression data by literature profiling. *Genome Biol*, 3(10):1–16.

[7] Cohen, A. M. and Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71.

[8] De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. *URL http://nlp. stanford. edu/software/dependencies manual. pdf*.

[9] Jenssen, T.-K., Lægreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature genetics*, 28(1):21–28.

[10] Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

[11] Krallinger, M. and Valencia, A. (2005). Text-mining and information-retrieval services for molecular biology. *Genome biology*, 6(7):224.

[12] Leroy, G. and Chen, H. (2005). Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. *Journal of the American Society for Information Science and Technology*, 56(5):457–468.

[13] Liu, H., Hunter, L., Kešelj, V., and Verspoor, K. (2013). Approximate subgraph matching-based literature mining for biomedical events and relations. *PloS one*, 8(4):e60954.

[14] MEDLINE (2015). Key medline® indicators.

[15] Ring, H. Z., Kwok, P.-Y., and Cotton, R. (2006). Human variome project: an international collaboration to catalogue human genetic variation. *Pharmacogenomics*, 7(7):969–972.

[16] Silva, F. C. C. d., Valentin, M. D., Ferreira, F. d. O., Carraro, D. M., and Rossi, B. M. (2009). Mismatch repair genes in lynch syndrome: a review. *Sao Paulo Medical Journal*, 127(1):46–51.

[17] Stapley, B. J. and Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. In *Pac Symp Biocomput*, volume 5, pages 529–540.

[18] Tanabe, L., Scherf, U., Smith, L., Lee, J., Hunter, L., and Weinstein, J. (1999). Medminer: an internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, 27(6):1210–4.

[19] van der Eijk, C. C., van Mulligen, E. M., Kors, J. A., Mons, B., and van den Berg, J. (2004). Constructing an associative concept space for literature-based discovery. *Journal of the American Society for Information Science and Technology*, 55(5):436–444.

[20] Verspoor, K., Yepes, A. J., Cavedon, L., McIntosh, T., Herten-Crabb, A., Thomas, Z., and Plazzer, J.-P. (2013). Annotating the biomedical literature for the human variome. *Database*, 2013:bat019.

[21] Wikipedia (2014). Biocurator — wikipedia, the free encyclopedia. [Online; accessed 31-May-2015].

[22] Witten, I. H. (2005). Text mining. *Practical handbook of Internet computing*, pages 14–1.

[23] Wren, J. D., Bekeredjian, R., Stewart, J. A., Shohet, R. V., and Garner, H. R. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, 20(3):389–398.