MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Extracting Phenotypic Information from the Literature via Natural Language Processing

Lifeng Chen, Carol Friedman

Department of BioMedical Informatics, Columbia University, NY, USA

Abstract

In recent years, the amount of biomedical knowledge has been increasing exponentially. Several Natural Language Processing (NLP) systems have been developed to help researchers extract, encode and organize new information automatically from textual literature or narrative reports. Some of these systems focus on extracting biological entities or molecular interactions while others retrieve and encode clinical information. To exploit gene functions in the post-genome era, it is necessary to extract phenotypic information automatically from the literature as well. However, few NLP projects have focused on this. We present the development of a system called BioMedLEE that extracts a broad variety of phenotypic information from the biomedical literature. The system was developed by adapting MedLEE, an existing clinical information extraction NLP engine. A feasibility evaluation study of BioMedLEE was performed using 300 randomly chosen journal titles. Results showed that experts achieved an average precision rate of 65.4%, (95%CI: [58.0%, 72.8%]) and a recall rate of 73.0%, (95%CI: [66.2%, 80.0%]). BioMedLEE had 64.0% precision and 77.1% recall respectively, according to expert agreements.

Keywords

Natural language processing, text mining, data mining, phenotypic information extraction.

Introduction

The fundamental aim of modern genetics is to connect phenotype with genotype. In the past decade, genomic sequencing, microarray analysis and electronic publishing have resulted in the explosion of biomedical data and knowledge. The completion of genome projects in human and several other organisms has increased the amount of sequence information drastically. However, the role of many sequences is unknown. For example, *human inherited diseases*, as a very small part of phenotypes, undergo vigorous investigation. Currently although LocusLink [1] and OMIM [2] record approximately 1,500 of the roughly 4,000 Mendelian inherited diseases with known genes, about 700 of the single-gene diseases do not have known associated genes. On the other hand, *complex conditions* such as asthma, diabetes and Alzheimer's disease are more common and impact millions of people. They are thought to result from the combination of a set of "susceptible genes" and "risk factors" from the environment. Nevertheless, the connections between the genes and complex disorders are by nature much harder to establish.

Online literature has become an important resource for investigating genotype-phenotype relationship. For example, OMIM, an online catalog of human genes and associated genetic disorders, is maintained by use of human curators manually reading and extracting information from the online literature. Other databases, such as FlyBase [3] and the Mouse Genome Informatics (MGI) [4] are also curating phenotypic information by manually extracting information from the published literature.

Nevertheless, utilization of the online literature is problematic. One issue is that current information from published literature is generally stored manually into the knowledge base by human curators. Thus, the maintenance of the knowledge base requires a tremendous amount of human effort. The large volume of information makes it impractical for manual identification and entry of relevant information into a knowledge base. Secondly, most databases are managed separately, and each represents a single, well-defined area. This makes search by researchers over different databases very time consuming. Thirdly, records in the databases occur mainly in the form of free text, which although convenient for human beings, is difficult for computerized system to reliably access.

Natural language processing (NLP) has the potential to solve this problem by extracting and structuring text-based biomedical information, making these data available for current use and future analysis. Several NLP engines have been developed in the clinical and biological domains. These systems are able to extract clinical [5-7] or biological [8;9] [10] terms and interactions from medical reports or biological literature, respectively. Many of these systems have shown effectiveness acquiring medical knowledge or genomic terms and interactions from corresponding textual sources. Important as it is, however, there is no NLP system to date dedicated to recognizing and extracting phenotypic information.

There are multiple ways to address this problem. A brand new system could be built. Alternatively, a current system may be adapted to achieve the same goal. Clinical NLP systems, for example, extract a variety of findings, such as diseases, symptoms and body locations. Since those types of information are essentially types of phenotypes, the success of extracting different types of clinical information from medical reports suggests the possibility of adapting a clinical NLP engine to

enable extraction of phenotypic information from online literature. In this work, we present BioMedLEE, a system based on adaptation of MedLEE, a clinical NLP engine to extract phenotypic information from the online literature. The feasibility of the adaptation was evaluated. Results are shown and discussed.

Background

In the past few years, NLP systems have been developed and implemented in many biomedical domains. In the clinical domain, many of the systems (MPLUS [5], LSP [6], MEDLEE [7], MENELAS [11], and RECIT [12]) focus on processing narrative medical reports such as discharge summaries and radiology reports. The NLP systems extract clinical terms and relations from the patient records. In the biological domain, NLP knowledge extraction mainly focuses on two categories: 1) biological entities such as genes and proteins [8;9] and 2) biological relations between those entities, e.g. protein-protein interaction [13-15] or protein-drug interaction [16]. Most of these NLP engines have shown promise; they report precision and recall ranging from 60-80% in the biological domain (For review, see [17]). Similar results were seen in the medical domain, with specificity and sensitivity (or precision and recall) rates of 60-90% [5;7;11].

There have been several projects that investigated gene-phenotype relationships from the online literature. Perez-Iratxeta has described a method using data mining to identify association of genes to inherited diseases [18]. They established a scoring system based on co-occurrence of curated MeSH headings of MEDLINE articles, GO terms [19] and protein functions recorded in RefSeq database. The main idea is that the more frequent the co-occurrence, the stronger the association the terms have. Using this method they were able to identify a set of potential disease-related genes. Their evaluation with 100 known disease-associated genes showed that there were 25% and 50% probabilities that the disease-associated genes would be among the 8 and 30 best-scoring genes respectively, suggesting there is a relationship between the score of a gene and its likelihood of being associated with a particular disease in an article.

Adamic described a similar statistical method that identifies sets of genes associated with given diseases from the literature [20]. They initially obtained gene symbols and aliases from HUGO (Human Genome Organization) [21], OMIM and LocusLink. Then they performed automated searches of MEDLINE abstracts to produce a "gene list" using a straightforward pattern matching. Particular diseases, such as leukemia, were selected for investigation. A scoring system was then used to calculate scores for "gene-disease" pairs according to the frequencies of their co-occurrences. They were able to identify most of breast cancer genes in one of the human edited breast cancer gene databases (http://tyrosine.biomedcomp.com) using this method.

However, these efforts mainly focused on diseases, which constitute a small part of phenotypes. Our system BioMedLEE differs from the above since we are focus on extracting a broad variety of phenotypic information from the online literature, and not just diseases. Additionally Perez-Iratxeta's work depends on curated resources, such as MeSH headings and annotated GO terms. BioMedLEE uses automated techniques and thus, does not depend on manually curated data. Adamic's system, while obtaining a comprehensive gene list from several databases, was limited to only a few diseases in implementation. Currently we are only recognizing phenotypic information. In a separate project, we refining a gene name recognition and identification module, and plan on integrating the gene name recognition module with BioMedLEE to facilitate interpretation of gene-phenotype relationship.

Overview of MedLEE

Detailed descriptions of MedLEE have been previously published [7;22]; in this paper, we present a brief summary. MedLEE consists of a series of processing components that utilize different knowledge components. This design enables extension to similar domains by augmenting or changing the knowledge sources while leaving the processing components intact. For example, a system GENIES, which extracts biomolecular interactions and other relations, was adapted from MedLEE by replacing the lexicon associated with the clinical domain with one relevant for the biological domain and by utilizing a new set of grammar patterns [13].

The relevant components for this paper are the preprocessor, the parser, and the partial parsing components. The preprocessor determines sentence boundaries, and also identifies and categorize phrases in the sentences. Identification of phrases is accomplished by using a knowledge source in the form of a lexicon, which contains entries for phrases associated with phenotypic information, their corresponding semantic or syntactic categories, and target forms. The next component, the parser, recognizes the structure of the sentences and generates the initial target output form. The parser uses grammar rules that combine semantic and syntactic patterns to recognize relevant clinical findings and modifier relations, and to generate target forms. If a parse cannot be obtained by strictly following the grammar rules, partial parsing is used. The representation of the output is in the form of frames, where the first element of a frame corresponds to the type of information, the second to the value. The remaining elements in the frame are modifier frames, which have a similar "type-value-modifier" frame structure. For example, the following frame was generated for the phenotypic information *developmental defects in the inner ear:*

[problem,defect,[bodyfunc,development],[bodyloc,ear,[region, inner]

In this example, the primary finding was a problem **defect**, which was qualified by the function **development**, and which occurred in the body location **ear**, in the region **inner**. The frame is subsequently transformed into XML form, which is a straightforward process.

Methods

Collecting a Corpus

We used the resources associated with the mouse model organism and automatically obtained all the abstracts from the MGI website as of February 2003. A set of genes was associated

with each of the abstracts from manual curation, which was specified in the website. In our current work, we focused on extracting phenotypic information from the set of abstracts and assumed that extracted phenotypes had some relationship with the genes associated with the abstract.

Development of the BioMedLEE system

Adapting MedLEE to enable extraction of phenotypic information requires less work than building a brand new system. However, two tasks have to be accomplished: 1) grammar rules associated with semantic and syntactic patterns in the new domain have to be edited to adjust the context change from medical narrative reports to biomedical literature; and 2) the lexicon has to be modified to enable recognition of more phenotypic terms, such as cellular functions and model organism anatomy.

To identify issues in adapting MedLEE to this new domain, we first parsed 50 randomly chosen abstracts using the original version of MedLEE, and then made incremental changes and refinements, consisting of changes to the lexicon and grammar. Terms associated with particular semantic categories were removed from the lexicon because they were not relevant to phenotypic information. For example, diagnostic procedures (e.g. *chest x-ray, biopsy)*, laboratory procedures (*chem.-7, apgar score*), healthcare devices (*catheter, surgical clips*) and medications (*Tylenol, vitamin A*) were removed. New terms were added to the lexicon that were not frequently observed in the clinical record. These terms were obtained from several resources. One resource was the Unified Medical Language System (UMLS) [23]. Approximately 19,000 UMLS terms associated with certain semantic categories were automatically imported into the lexicon. For example, terms with classes corresponding to cellular body functions (*antibody formation, cell adhesion, blastogenesis*), cellular dysfunction (*chromosome deletion, neuron degeneration, polyploidy*), and cellular components (*cell nucleus, cytoskeleton*) were added. Seventy anatomic terms were added according to terms found in the MGI website that were associated with anatomy (*tail, whiskers, hindlimb*). The third resource was the Mammalian Ontology [24], which specified morphological behavior (*circling, bobbing, compulsive biting)*, and other phenotypes (*curly whiskers, polydactyly)*. Three hundred and forty-four terms were added from this resource. The fourth resource was biological expertise. Two hundred and fifty terms that were found to be relevant but missing were added, such as *corticogenesis, dysmorphology, cyclopia*).

Changes to the grammar involved removing certain patterns. For example, medication patterns, diagnostic procedure patterns, demographic patterns, and recommendation for follow up examination patterns were removed from the grammar. Only one new grammar rule was added, to accommodate absence of body parts, and no changes to the other rules were made.

Since the revised version of MedLEE was designed to handle a different domain, it was named BioMedLEE. The text was reparsed using BioMedLEE and problems were noted and their causes analyzed. Only simple corrections were made in subsequent rounds of refinement because our aim was to perform a feasibility evaluation early on to assess performance; if the results were promising, we aimed to spend more effort in refining the system. For example, we removed about 30 lexical entries because they had different senses in the biological environment than in the clinical environment. For example, *growth* with a sense *tumor* is rare in the biological domain. When the new revisions were completed, the same set of abstracts was parsed using them. Several rounds of refinements were made prior to the feasibility study.

Evaluation

To perform the feasibility study, a test set of 300 titles was randomly chosen from the set of MGI abstracts. Nine experts were used as subjects to obtain a gold standard. Each title was assigned in such a way that it was examined by three different experts. Each expert was given instructions on how to read the titles and highlight phenotypic information. In this study, we were primarily focusing on diseases, functions and behaviors above the molecular level, i.e. phenotypes on cellular, tissue, organ, body system and organism level (e.g. *skin carcinogenesis, epithelial migration*). Examples were given to help the experts understand the task. A set of 5 titles was provided for self-testing before they read the titles from the test set. Then each expert read 100 titles and highlighted terms that were judged to contain phenotypic information. All the results were gathered and examined by LC (first author of the paper) to build a gold standard. If there was a disagreement between experts, a majority vote was used. Some words, such as "*of*", "*in*", "*a*" and "*the*" were not considered when comparing terms. In some cases, one expert highlighted a subset of terms chosen by another, that was different only because of modifiers, e.g. "*embryonic death*" and "*early embryonic death*". This was considered to be an agreement and the smallest relevant phrase was chosen for the gold standard. In the example above, "*embryonic death*" was considered as the phenotype in the sentence. The same set of titles was then processed by BioMedLEE and phenotypic information extracted. The outputs were compared with the gold standard using the same criteria that were used to judge the experts' opinions. We then calculated the recall and precision measurements for both the experts and BioMedLEE. Recall was calculated as the number of correct phenotypic relations extracted by BioMedLEE or experts divided by the number obtained by the gold standard. Precision was computed as the number of correct phenotypic relations extracted by BioMedLEE or individual experts divided by all relations that were extracted. To avoid bias, when evaluating individual experts, the particular expert was excluded from the reference standard. For the cases where the expert only agreed with one of the two other experts that were to form the reference standard, a dice was thrown to decide if the phenotypic relation should be considered a reference standard.

Results

Of the 300 titles, a total of 481 phenotypic relations were marked up by the experts, among which the experts agreed completely on 170 (35.3%) while they were unable to achieve an agreement on 167 (34.7%). For the remaining 144 (29.9%) phenotypic relations, 2 of the 3 experts agreed and a majority vote was used.

Therefore 314 phenotypic relations, for which complete agreement or a majority vote was achieved, formed the gold standard in our evaluation study. Of the 314 phenotypic relations, BioMedLEE correctly recognized 242 of them, giving a recall of 77.1%. BioMedLEE also extracted 136 more phenotypes that were not highlighted by the experts. The precision rate was thus 242/(242+136), or 64.0%. At the same time, an average precision rate of 65.4%, (95%CI: [58.0%, 72.8%]) and a recall rate of 73.0%, (95%CI: [66.2%, 80.0%]) were achieved for individual experts.

Discussion

Our feasibility evaluation showed promising results. BioMedLEE was able to extract phenotypic information from textual titles with comparable precision and recall to individual experts. Also many of BioMedLEE's problems appeared easy to fix. We found that most false positives were caused by 6 general sense terms (*function, deficiency, development, disruption, tissue* and *cell*) that were extracted by BioMedLEE but not experts when these terms occur without any modification. It seemed that BioMedLEE was penalized because these terms appeared in the outputs frequently. We applied a simple filter that required these terms be retained only when they have certain types of modifiers. When this simple postprocessor was applied, half (68) of the false positives were eliminated and the precision rate was improved to 78.1% without affecting the recall rate.

Expert evaluation was costly. It took about 40-60 minutes for an expert to highlight 100 titles while BioMedLEE required only 35 seconds to process the same set of titles using a Sun Blade 2000 workstation with 2 900 MHz 64-bit UltraSPARC III CPUs and 2G RAM.The disagreement among the experts seems very high (~40%). This was probably due to the vague nature of the question we addressed: what constitutes "phenotypic information"? According to the Merriam-Webster dictionary, phenotype is "the visible properties of an organism that are produced by the interaction of the genotype and the environment". Thus, phenotype is the outward, physical manifestation of an organism. This can include the physical parts, the sum of the atoms, molecules, macromolecules, cells, structures, metabolism, energy utilization, tissues, organs, reflexes and behaviors or anything that is part of the observable structure, function or behavior of a living organism. Although instructions were given to the experts to help them highlight the desired phenotypic information in this study, that were the "higher levels" phenotypes than molecular interactions, frequently the experts highlighted molecular interactions such as "*transcription*" or "*alternative splicing*". In many cases, one expert would highlight the information while another decided not to, resulting in the high disagreement rate. Instructions given to the experts may have played a role in the high disagreement rate. The instructions explicitly stated what information should be highlighted but did not explicitly list what should not be. In the future, more negative examples will be provided, and we expect the disagreement rate will decrease.

Since MedLEE is an NLP engine for the clinical domain, when we adapted it to extract biological information, some problems occurred due to incorrect interpretation of terms. For example, "*interaction*" was interpreted as "*drug interaction*", which is incorrect in most biological environments. Also there were excessive terms in the lexicon, which caused false positives. For example, "*autosomal homologue*", which is a molecular level term, was captured by BioMedLEE. This occurred because supplemental information associated with body functions were automatically imported into the lexicon from the UMLS, which did not differentiate between molecular and observable levels of phenotypes. This can be fixed using knowledge engineering. Another more difficult problem was due to ambiguity. For example, the gene symbol "*fat/fat*" was interpreted as the phenotype "*obesity*". Ambiguity represents 7% (10/136) of all errors in our examples. However, this rate can be higher (up to 10/68 = 14%) if the general terms causing false positives are fixed. Additionally, some types of phenotypic information were not recognized by BioMedLEE because the corresponding terms were not included in the lexicon. The "incompleteness" of the lexicon, however, is time consuming to deal with. Since we derived our lexicon from certain sources, e.g., the UMLS metathesaurus, the completeness and correctness of our lexicon depends on the corresponding resources. The lack of a comprehensive terminology for phenotypes of all organisms will necessitate that relevant terms will have to be manually added, which is a substantial amount of work. Thus, the lack of availability of a comprehensive resource may be a limiting factor in our research. Many researchers have addressed this problem. For example, some groups have used uncurated terminology, i.e., phrases dynamically extracted from the literature in combination with a curated terminology to relieve this problem [16;25]. Majoros has described a model using a hidden Markov heuristic to identify key concepts in biomedical literature, to help improve speed and accuracy in ontology construction [26].

Although promising, our results showed, BioMedLEE will have to undergo further revision. Future work will involve: (1) extending the lexicon; (2) integrating the module that extracts gene or protein (GENIES) information with BioMedLEE so we can capture and organize gene-phenotype relationships; (3) work on resolving ambiguous words and acronyms. This is an important but difficult task since ambiguity appears to be big problem even in a single domain for a single species. It certainly gets more severe when considering multiple species and domains [27; 28].

Conclusion

Genotype-phenotype relations are important in modern genetics. NLP can facilitate extraction of this information automatically from the biomedical literature. Phenotypic information is very broad, and resembles clinical information in many ways. We developed an NLP system, BioMedLEE, which extracts phenotypic information from the biomedical literature and performed a feasibility study. Our results showed that individual experts disagreed on phenotypic information contained in the titles 40% of the time. BioMedLEE provided comparable precision (64.0%) and recall (77.1%) to experts. Future effort will be focusing on refinement, improvement and evaluation of BioMedLEE to help extract valuable phenotypic information from biomedical literature.

Acknowledgments

We thank the experts who helped us in evaluation of the BioMedLEE. We thank Dr. Judith Blake for her help with the MGI resources. This work was supported in part by grant EIA-031 from the National Science Foundation, and LM06274 and LM7659 from the National Library of Medicine.

References

[1] Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 2001; 29: 137-140.

[2] Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* 2000; 15: 57-61.

[3] Drysdale R. Phenotypic data in FlyBase. Brief. *Bioinform.* 2001; 2: 68-80.

[4] Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT. MGD: the Mouse Genome Database. *Nucleic Acids Res.* 2003; 31: 193-195.

[5] Christensen L, Haug P, Fiszman M. MPLUS: a probabilistic medical language understanding system. *Proceedings of ACL Workship in Natural Language Processing* , 29-44. 2003.

[6] Sager N, Lyman M, Nhan NT, TIck LJ. Medical language processing: applications to patient data representation and automatic encoding. *Meth Inform Med* 1995; 34: 140-146.

[7] Friedman C, Alderson PO, Austin J, Cimino JJ, Johnson SB. A general natural language text processor for clinical radiology. *JAMIA* 1994; 1: 161-174.

[8] Fukuda K, Tsunoda T, Tamura A, Takagi T. İnformation extraction: identifying protein names from biological papers. 707-718. 1998. Hawaii. *Proceedings of the Pacific Symposium on Biocomputing '98.*

[9] Jenssen T, Vinterbo SA. A set-covering approach to specific search for literature about human genes. Overhage M. 384-388. 2000. *Proc AMIA Symp* 2000.

[10]Krauthammer M, Rzhetsky A, Morozov P, Friedman C. Using BLAST for identifying gene and protein names in journal articles. *Gene* 2000.

[11]Zweigenbaum P, Bachimont B, Bouaud J, Charlet J, Boisvieux JF. A multi-lingual architecture for building a normalized conceptual representation from medical language. Gardner, RM. 357-361. 1995. Phil, Hanley & Belfus. *Proceedings of the 19th Annual SCAMC.*

[12]Rassinoux AM, Wagner JC, Lovis C, Baud RH, Scherrer JR. Analysis of medical texts based on a sound medical model. Gardner, RM. 27-31. 1995. Phil, Hanley & Belfus. *Proceedings of the 19th Annual SCAMC.*

[13]Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* 2001; 17 Suppl 1: S74-S82.

[14]Ono T, Hishigaki H, Tanigami A, Takagi T. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics.* 2001; 17: 155-161.

[15]Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M. Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.* 2000; 541-552.

[16]Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.* 2000; 517-528.

[17]Hirschman L, Park JC, Tsujii J, Wu CH. Accomplishments and challenges in literature data mining for biology. *Bioinformatics.* 2002; 18: 1553-1561.

[18]Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* 2002; 31: 316-319.

[19]Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et.al. Gene ontology: tool for the unification of biology. *Nat Genet* 2000; 25: 25-9.

[20]Adamic A, Wikinson D, Huberman A, Adar E. A literature based method for identifying gene-disease connections. *IEEE Computer Society Bioinformatics Conference* , 109-117. 2002.

[21]Wain HM, Lush M, Ducluzeau F, Povey S. Genew: the human gene nomenclature database. *Nucleic Acids Res.* 2002; 30: 169-171.

[22]Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. *Meth Inf in Med* 1998; 37: 334-344.

[23]Lindberg D, Humphreys B, McCray AT. The Unified Medical Language System. *Meth Inform Med* 1993; 32: 281-291.

[24]Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT. MGD: the Mouse Genome Database. *Nucleic Acids Res.* 2003; 31: 193-195.

[25]Yoshida M, Fukuda K, Takagi T. PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics.* 2000; 16: 169-175.

[26]Majoros WH, Subramanian GM, Yandell MD. Identification of key concepts in biomedical literature using a modified Markov heuristic. *Bioinformatics.* 2003; 19: 402-407.

[27]Hirschman L, Morgan AA, Yeh AS. Rutabaga by any other name: extracting biological names. *J. Biomed. Inform.* 2002; 35: 247-259.

[28]Tuason O, Chen L, Liu H, Blake J, Friedman C. Acquisition of lexical Knowledge using Biological Nomenclatures. *Pac.Symp.Biocomput.* 2004.

Address for correspondence

Lifeng Chen, MS,
Department of BioMedical Informatics, Columbia University
Vanderbilt Clinic Building, 5th Floor,
622 West 168th Street
New York, NY 10032
mailto:lifeng.chen@dbmi.columbia.edu