TOWARDS PATHWAY CURATION THROUGH LITERATURE MINING – A CASE STUDY USING PHARMGKB

RAVIKUMAR K.E., KAVISHWAR B. WAGHOLIKAR, HONGFANG LIU

Department of Health Sciences Research, College of Medicine, Mayo clinic, Rochester, MN, 55905
Email: {KomandurElayavilli.Ravikumar, Wagholikar.Kavishwar, Liu.Hongfang}@mayo.edu

The creation of biological pathway knowledge bases is largely driven by manual effort to curate based on evidences from the scientific literature. It is highly challenging for the curators to keep up with the literature. Text mining applications have been developed in the last decade to assist human curators to speed up the curation pace where majority of them aim to identify the most relevant papers for curation with little attempt to directly extract the pathway information from text. In this paper, we describe a rule-based literature mining system to extract pathway information from text. We evaluated the system using curated pharmacokinetic (PK) and pharmacodynamic (PD) pathways in PharmGKB. The system achieved an F-measure of 63.11% and 34.99% for entity extraction and event extraction respectively against all PubMed abstracts cited in PharmGKB. It may be possible to improve the system performance by incorporating using statistical machine learning approaches. This study also helped us gain insights into the barriers towards automated event extraction from text for pathway curation.

1 Introduction

Genome-wide high throughput studies have led to an increased emphasis on understanding the biological interactions at the systems level rather than the individual molecular interactions. Biological pathway knowledge bases provide systems level interaction information, and are constructed by manual curation of the scientific literature. Due to extensive manual effort required, there is a significant delay in capturing the information in knowledge bases after the publication of scientific literature. Baumgartner et al 2007 (1) suggests that manual curation of biological databases is beyond human life span without significant assistance from text mining. Increase in the volumes of biomedical literature has witnessed simultaneous improvements in the ability to apply natural language processing (NLP) methods to full text articles and entire PubMed collection (2-4).

Despite a decade of research in biomedical text mining the effort to semi-automate the curation workflow of various biological databases and pathway databases in particular is still evasive (5). Some of the earlier systems targeted the acquisition of protein networks (binary relations) from literature are simply based on co-occurrence such as iHOP (6), Chillibot (7), or grammar-based rules such as Pathway Studio (8) and GeneWays (9). While extraction of such networks is useful, the networks cannot be easily mapped to pathways, which model information flow in biological cascades.

While most of the systems mentioned above extract binary relations there has been significant improvement in the state of the art by progressing the extraction from simple binary interactions to complex events, which form building blocks of a pathway. In the recent past the efforts to achieve automated biomedical text mining have been catalyzed by a series of BioCreative (10, 11) and BioNLP shared tasks (5, 12, 13). These competitions saw the emergence of systems (2, 3, 14, 15) that extract complex events where simple events are part of other events using both machine

learning and rule-based approaches. PathText (16) proposed an integrated approach to ease the manual effort involved in pathway curation task but still requires lot of manual effort. The most recent BioNLP shared task 2013 (5) organized a task dedicated to pathway curation. Only two systems, TEES (3) and NacTeM (17) participated in this task, which reported an F-measure of 52.84% and 51.10% respectively on the task. Schmidt et al 2012 (18) also explored text mining assisted pathway curation in a limited context of a specific pathway involving kinases.

While the recent studies indicate a step forward in the direction of pathway curation, they do not completely address all the issues necessary for pathway curation. We are not aware of any study that evaluates a text mining system for extracting biological pathways that uses a manually curated pathway database as the gold standard.

In this study we describe an event extraction that uses pattern templates (covering nearly 450 verbs describing biological events) to extract arguments and assign semantic roles for events described within a single sentence. In addition the system uses linguistic rules to connect information across sentences, which is a major distinguishing feature of the system from rest of the systems described above. Finally we investigate an important problem of great significance, the role our text mining system can play in assisting pathway curation through extraction of events and identify the challenges to our text mining system in extracting the event annotations in PharmGKB (19) pathway database.

2 Methods

Figure 1 shows the overall system architecture and the individual components of our text mining system.

2.1 Pre-processing and Named entity recognition

The pipeline starts with tokenization and sentence detection for a given document. The sentences are then assigned part of speech using Brill Tagger (20) trained on GENIA corpus (21). POS tagging is augmented by post-processing error correction rules. This is followed by shallow parsing using fnTBL chunker (22) trained on GENIA corpus (21). The shallow parsing is supplemented with detection of additional syntactic constructions related to noun phrases, which include co-ordination, appositives and verb groups.

The next component is named entity recognition (NER) component consisting of manually developed rules as outlined by Narayanaswamy et al 2003 (23) and dictionaries of words and morphological features like prefixes, suffixes and infixes for biomedical entities. The NER component classifies entities into 8 major categories namely Protein/Gene, protein sites, chemicals, drugs, organism, bodypart (include organ, tissue, cells and sub-cellular location), disease, quantitative parameter (e.g. conductance, voltage, binding constant, dissociation constant, IC50) and values (e.g. 20 nM, 30 pS, 10 ms). Based on the NER results we corrected the errors in POS tagging and shallow parsing module by having a feedback loop in order to improve the performance of event extraction.

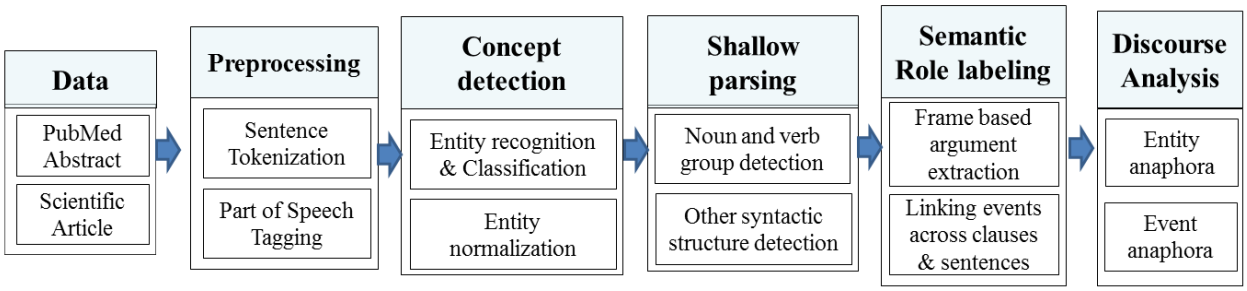


Figure1 – System Architecture

2.2 Event extraction

The event extraction module consists of two major sub components 1) detection of events within a clause or sentence based on pattern templates and 2) connecting events across sentences through discourse analysis.

2.2.1 Argument extraction based on verb frames

The system consists of rules for different classes of verbs or its nominal forms that extract and assign thematic roles to its arguments based on verb category and the semantic type of the arguments. The patterns for each verb were developed using a corpus of 300 abstracts related to electrophysiology sub-domain describing events about ion channel physiology. Currently there are 9 major classes and 50 sub-classes of verbs. The patterns consider the verbal forms such as activate, inhibit, transport and nominal forms such as activation and phosphorylation. These verb/nominal forms are marked as potential triggers and there are 450 such triggers identified across all categories. Table 1 lists the major category of event classes and the corresponding verbs for defining frames for argument extraction. Some example patterns are included below with example sentences can be found in Figure 2.

Pattern 1: <Agent> (PRP NP)* REGULATE_VERB <Theme> (PRP NP)*

This template matches a clause with a verb and extends the clause on either side of the verb as long as each of the base noun phrases that it crosses is headed only by a preposition (shown in Figure 2A). Regulatory verbs (both positive, negative and neutral) such as "increased", "stimulated", "blocked", and "prevented", "regulated" have the above argument structure and are matched by this pattern.

Pattern2: < Nominal form NP> of <THEME> by <AGENT>

This pattern matches the sentence and extracts arguments (shown in Figure 2B). A similar pattern handles passive forms of the verb as shown in Figure 2C.

Pattern 3: <AGENT>, [Nominal form NP] of <THEME>

Pattern 3 handles nominal forms within appositive expressions like in "**Gd3+**, *an inhibitor of the* ***flow -induced Ca2+ increase***, prevented the hyperpolarization" and extracts the arguments ("*Gd3+*" as agent and "*flow -induced Ca2+ increase"* as theme) for the trigger "inhibitor".

2.2.2 Connecting events across clausal boundaries

We explored a few linguistic motivated approaches to connect or transfer arguments across clausal boundaries. Our strategy involve three steps: 1) fill empty semantic slots by transferring the arguments across events, 2) merge relevant frames and write parser to connect discourses, 3) resolve anaphoric expressions to find the right antecedent for both entities and events. Figure 2 shows the examples for frame based argument extraction output using BRAT annotation tool.

Table 1. Verb categories

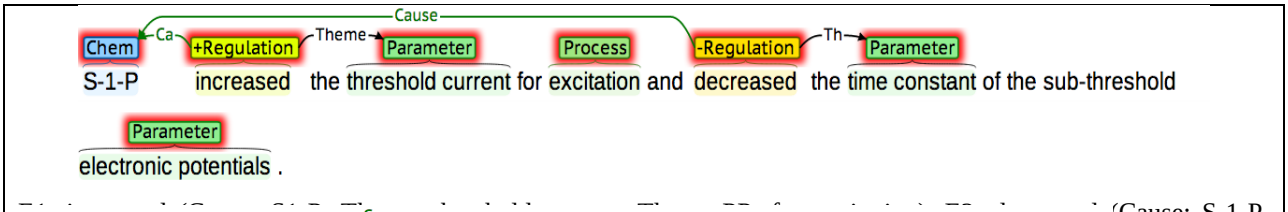
Category	Example verbs
Conversion	Phosphorylation, methylation, de-phosphorylation etc. and other PTMs
Localization	Transport, trans-located, movement
Gene expression	Expression, transcription, translation
Degradation	Degradation
Binding	Bind, binding, complex formation
Dissociation	Dissociate, bond break
Regulation	
Positive	Activation, induce, trigger
Negative	Inhibition, inactivation
Neutral	Modulate, regulate

Filling empty slots by transferring arguments across events - Quite often, syntactic arguments of verbs or its nominalized form, either the subject/object will be empty. Such situations demand mechanisms to fill the empty arguments by linking the current frame with another. Consider the example sentence shown in Figure 2D. While "ZD7288" and "CsCl" and "latencies" are extracted as the cause and theme respectively for the verb "increased", the "properties of both open states" is extracted as the theme of the verb "decreased". Our rule to allow transfer of arguments (either Cause or Theme) if the verbs are in co-ordination and belong to the same category ("Regulation" in this case) enable easy identification of "ZD7288 and CsCl" as the agent for the verb "decreased".

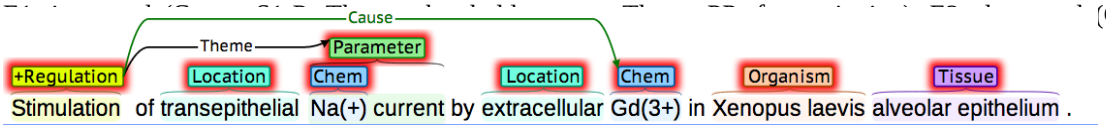
The above co-ordination rule can handle even more complex co-ordination structures beyond clausal boundaries as shown in example in Figure 2E. Here the co-ordination between the verbs "did not affect" and "did increase" is identified, which triggers the argument transfer rule to help identify "8Br-cAMP" as the cause of the verb "increase".

Linking sequential events by merging frames - We also link sequential events as conveyed in the text by merging the frames and connecting discourses. Consider the sentence shown in Figure 2F. From that sentence our verb frame based extraction module extracts the following outputs: EVENT1: effects (NO; I (Na)), and *EVENT3: inhibition (UNK, both cGMP and cAMP pathways) for the events "effects" and "inhibition" respectively. If we carefully notice on either side of the verb "blocked" we have the nominal form of verb followed by a prepositional phrase. In such cases we connect both the events as EVENT2: (Event3, Event1).*

We also have rules to extract lexical chains by handling discourse connectives such as "thereby", via whereas etc., which are often used to connect two events in the text.

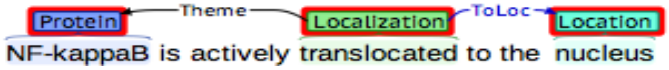


(Cause: S-1-P,



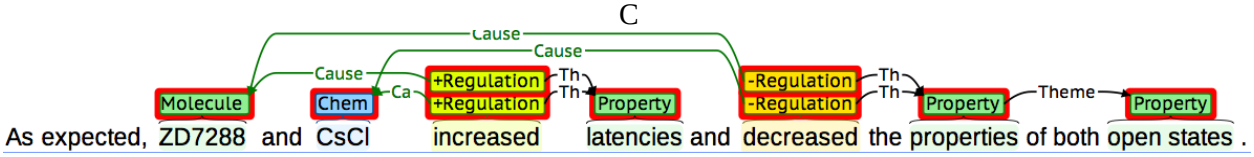
E1: Stimulation (Cause: extracellular Gd(3+), Theme: transepithelial Na(+) current).

B



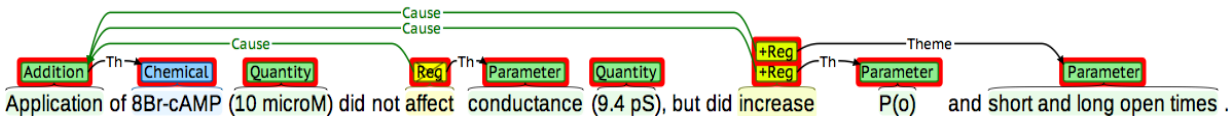
E1: translocated (Theme: NF-kappaB; FromLoc:UNK; ToLoc: nucleus)

C



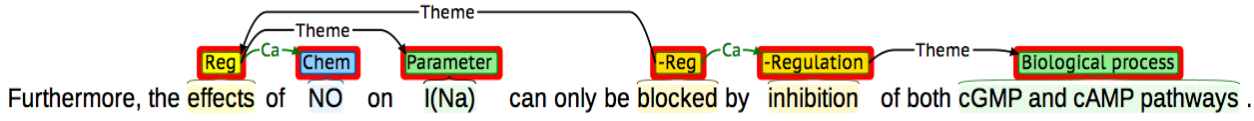
E1: increased (Cause: ZD7288 and Cscl; Theme: Latencies), E2: decreased (Cause: ZD7288 and Cscl; Theme: properties; Theme PP: both open states)

D



E1: application (Theme: 8Br-cAMP), E2: NOT_Effect (Cause:E1; Theme: conductance); E3: increase (Cause:E1, Theme:E1)

E



E1: effect (Cause: NO; Theme: I(Na)), E2: inhibition (Theme: cGMP and cAMP pathways); E3: blocked (Cause:E2, Theme:E1)

F

Figure 2. Examples of verb frame based argument extraction. A) Active forms, B) Nominal forms, C) Passive forms, D) Filling empty arguments through verb co-ordination, E) Filling empty arguments beyond clausal boundaries, and F) Linking events through merging verb frames. +Regulation/+Reg: positive regulation; -Regulation/-Reg: Negative regulation

Anaphora resolution - We also have a simple anaphora resolution module to resolve both anaphoric entities and events. Our approach to anaphora resolution for entities is linguistic rules described in Kennedy and Boguraev 1996 (24). For demonstrative NPs such as “this kinase”, “these transcription factors” we consider features such as semantic type of the NPs, the distance

between the antecedent and the candidate anaphora and number (singular or plural form of NP) while deciding the right antecedent. For the anaphoric phrase “both sites” in the following snippet “Dephosphorylated hsp 90 is phosphorylated at *both sites* by casein kinase II …”, the candidate antecedents that our method would consider are those phrases which refer to two objects of the type dictated by the head word “site” (protein sites). We look for antecedent phrases which are of the semantic type “protein sites”. In this case, the rule correctly identified the anaphor “serine 231 and serine 263” which appeared in a preceding sentence, “For the alpha protein, these sites correspond to *serine 231 and serine 263*.” Anaphora resolution play critical role in recovering the actual arguments as shown in the following example.

Besides resolving anaphors at the entity level we also have rules to resolve event anaphora. Our strategy to resolve event anaphora is based on the identity of the verbs if they have the same root form post- lemmatization. For example, consider the following sentence, “*This modulation* may **contribute** to the **migratory effect** of **MIP1-alpha on microglia**”. The system extracted two outputs for the trigger “contribute” and “effect” as given below: **Event1**: contribute (this modulation, the migratory effect of MIP1-alpha on microglia); **Event2**: effect (MIP1-alpha, microglia). In the first event (Event1) the phrase “*this modulation*” is resolved to as referring to the modulation event, described in the prior sentence, “Thus, microglia in hippocampi from epileptic patients expresses *high-conductance Ca2+-dependent K+ channels* that are **modulated** by the *chemokine MIP1-alpha*”. For the event "modulated” the system extracted the following output: “**Event3**: ***modulated*** **(the chemokine MIP1-alpha, high-conductance Ca2+-dependent K+ channels)**. After anaphora resolution the system finally gets the consolidated output as **Event1: (Event3, Event2)**.

3 Experiments

3.1 Data set

We evaluated the performance of the system by extracting events from PubMed abstracts cited as literature evidence in PharmGKB, and comparing the system output with the manual annotations in the PharmGKB (19). PharmGKB pathway is a rich resource, which catalogs both the pharmacodynamics and pharmacokinetics pathways involving the interplay between the drugs, metabolites and genes through manual curation along with the citation to primary literature evidence namely the PubMed (25). PharmGKB pathway resource’s latest version (As on July 1st 2013) contains 99 pathways with citations to primary literature. Besides these it also contains other pathways assembled from other resources such as Reactome(26). In addition to events we evaluated the system for identifying all the participating molecules (genes/chemicals) involved in the pathways. We reported the performance as precision, recall and F-measure. For each event in a pathway we compared the individual fields (see Table3) namely From, To, and ControlledBy against the manual annotations. True positives were required to match all the four fields. For the manual evaluation we considered additional criteria during evaluation. By ignoring the gene

normalization we considered the extraction to be correct if the biology intuition tells that the identified gene mentioned in the text is synonymous to the one in the PharmGKB.

3.2 Post-processing the system output to compare against PharmGKB annotation

For the current study we retrieved all the PubMed IDs (1,036) cited as literature evidence in the 99 PharmGKB pathways and retrieved them from PubMed through Entrez batch search (27). We formatted our system's output to generate the annotations in the same format as that of PharmGKB event annotation. In order to further align our gene mentions with the PharmGKB annotation, we normalized the textual mentions of gene/protein to Gene symbols using GeNO (28). We remapped the entity annotations produced by our system with that of GeNO by comparing the output span indices of the two systems. Even if there were overlap in the indices we aligned both the annotations and assigned the Gene symbols identified by GeNO to the corresponding entity mentioned in the text. We mapped the Agent/Cause of the verb extracted by our system to the "Controlled By" field in PharmGKB while the Theme identified by our system is mapped to "From" field. If the theme of the verb did not undergo any transformation in its molecular state through post-translational modifications, metabolism etc. then the same theme is assigned to the "To" field as well. For example consider the sentence (PMID: 11287982)

Table 2. Sample PharmGKB annotation

From	To	Controlled By	Evidence
BCR-ABL	BCR-ABL	imatinib	11287972;12755554;13679030;16122278
imatinib	CGP	CYP1A2;CYP2C19;CYP2C9;CYP2D6;CYP3A4;CYP3A5	15828850;16122278

"**Imatinib** is a potent and selective ***inhibitor*** of the **protein tyrosine kinase Bcr-Abl**, **platelet-derived growth factor receptors** (PDGFRalpha and PDGFRbeta) and **KIT**". Imatinib, the agent of the verb "inhibitor" in the above sentence is mapped to the "ControlledBy" field and one of the theme "**Bcr-Abl"** is mapped to the "From" field. Since the verb inhibitor do not involve any transformation of the theme it is also assigned to the "To" field.

3.3 Evaluation

We performed two evaluations 1) automated evaluation on all the event descriptions in PharmGKB pathways 2) manual evaluation of event extraction for four selected pathways. The four pathways are Platelet aggregation inhibitor pathway, Warfarin pathway, Metformin pathway, and Aromatase inhibitor pathway. We assessed the utility of our system output in pathway curation. Besides events, we also evaluated the ability of the system to identify all the participating molecules (genes/chemicals) in the pathways. We used the standard metrics namely precision, recall and F-measure for evaluation. For each event in a pathway we compared the individual fields namely From, To, and ControlledBy against the manually curated one and if all the four fields are found to be correct we count them to be a true positive event. Otherwise we count them as both precision and recall error. We did not report the partial recall for the fields correctly identified by the system.

4 Results and discussion

4.1 Evaluation on complete PharmGKB data set

PharmGKB pathway annotation contains 894 events involving 1040 molecules (839 genes and 201 drugs) annotated from 99 PharmGKB pathways. We evaluated the ability of our system in identifying the molecules participating in events annotated in PharmGKB pathways as shown in Table 3. Out of the two classes of entities the performance of Gene named entity was extremely lower (F-measure: 56.96) as it involve normalizing the gene mentions in the text to Entrez gene symbol as per the requirements of PharmGKB annotations. However for identifying drugs and chemicals the F-measure was fairly high (82.68%) as it doesn't involve entity normalization.

Table 3. Evaluation of system's performance on entity identification on complete PharmGKB

Entity Type	Total Entity (Gold)	Total Extracted (Total correct)	Precision (%)	Recall (%)	F-measure (%)
Gene	839	632 (419)	66.30	49.94	56.96
Drug/Chemical	201	261 (191)	73.18	95.02	82.68

Table 4 lists the performance of our event extraction system on the 1036 abstracts cited as literature evidence in PharmGKB pathways. The 99 pathways in PharmGKB contain 894 events. Our system identified 952 events from the 1036 abstracts out of which only 323 were found to be correct leading to precision of 33.93%, recall of 36.13% and F-measure of 34.99%. However we observed that extra-sentential processing modules contributed to only 4.5% improvement to the final output. The likely reason may be that PharmGKB annotation of pathway events mostly involves only simple entities such as genes and proteins but not complex events such as biological processes.

Table 4. Evaluation of system's performance on event extraction from PharmGKB

Total Events (Gold)	Total Extracted (Total correct)	Precision (%)	Recall (%)	F-measure (%)
894	952 (323)	33.93	36.13	34.99

4.2 Manual evaluation of four hand-selected pathways

While we expected the recall to be lower we were surprised to observe lower precision, a feature atypical of rule-based systems. In order to better understand the reason behind the low precision we manually evaluated the performance on abstracts related to four hand-selected pathways, which has citations to 34 abstracts as literature evidence. The manual inspection of the system output on these 34 abstracts aimed to identify the reason behind the low recall and precision. We observed the following discrepancies between the extracted output and the gold standard annotation in the PharmGKB:

1) Certain annotations in PharmGKB are not actually present in either the abstract or in the full text article. For example in the Platelet Aggregation inhibitor pathway we have the following annotation in PharmGKB as given in Table 5 below.

We did not find any mention of the individual G-protein in the ControlledBy column either in the cited abstracts or in the full text articles. However, there is a general mention about the involvement of G-proteins from the G-12&13 families, which our system extracted correctly. Out of the total 24 annotations for this pathway in PharmGKB, there were 7 annotations, which do not have direct evidence in the literature considering both the abstract and full text article. Instead they were derived through biological inference. None of these annotations were identified by our system. While from a biologist perspective the annotation in the pathway database is correct we believe that the current state of the art of literature mining has not matured enough to extract such annotations. Inferencing by using the background knowledge from knowledge bases such as PRO, UniProt etc. alone can help resolve such uncertainties.

Table 5. PharmGKB annotation from platelet aggregation pathway

From	To	Controlled By	Evidence
ADCY3	ADCY3	GNA11,GNA12,GNA13,GNA15,GNAI1,GNAI2,GNAI3,GNAQ,GNB3,GNAS	15187029, 11997386

2) Another notable reason for lower recall is that the information in pathway database is synthesized from multiple abstracts while our system extracts information only from a single article.

3) Another observation clearly explains the reasons for the lower precision of the system. Our system extracted a few annotations with no corresponding entries in PharmGKB. On manual inspection we found that while those annotations are not wrong they do not confirm to the event definition of the PharmGKB database. For example from an abstract (PMID: 15187029) the system extracted two relations namely, *regulate (P2Y(12), PP1)* and inhibit *(P2Y(12), adenylate cyclase)* from the sentence "Furthermore, the Src family kinase inhibitor **PP1** selectively potentiates the contribution to the calcium response by **P2Y(12)**, although *inhibition* of **adenylate cyclase** by **P2Y(12)** is unaffected." which are not annotated in PharmGKB. While both the relations extracted are correct from the biologist perspective it is not relevant in the context of PharmGKB annotation. The errors in gene normalization (both recall and precision) also contributed to the errors in event extraction as well. Table 6 lists the performance of our system on the selected 4 pathways through manual evaluation with and without ignoring the gene normalization.

Table 6. Evaluation of system's performance on event extraction on handpicked PharmGKB dataset

Event Type	Total Events (Gold)	Total Extracted (Total correct)	Precision (%)	Recall (%)	F-measure (%)
Event ignoring normalized entities	58	69 (39)	56.52	67.24	61.41
Events with normalized entities	58	41 (25)	60.97	43.13	50.50

The first row in Table 6 corresponds to the evaluation where we considered the event annotations to be considered as correct even if the genes were not normalized to the correct Entrez gene symbols. We used the biological inference to judge if the extracted gene matches the gene definition annotated in PharmGKB. However we wish to clarify if the event is not represented in

the annotation we considered the text extraction to be a false positive as our underlying focus in this study is to evaluate the utility of literature mining in pathway curation. The second row in Table 6 considers the extraction to be correct only if the genes are normalized to the correct gene symbols. We observed an appreciable drop in the recall (>20%) and very little increase in precision (~3%) when we consider gene normalized events, which illustrates that it is an important limitation in the performance of standardizing event extraction. Another limitation that we would like to point out is that our system being a rule-based one may require substantial manual effort to tune it to scale and improve its performance further.

5 Conclusions and future directions

Despite these limitations we believe that in this study we have made sincere efforts to explore and understand the limitations of a literature mining system in the context of extracting event descriptions which will be useful in finding literature evidences for actual pathway curation in a limited context of PharmGKB database. Our results are substantially lower than the recently reported studies (2, 3). However it is not fair to compare the performance of the system evaluated in this study with that of other systems as there is significant difference in the evaluation schema itself. Most of the previous studies evaluate the event annotation capability against the annotations at the textual level either abstracts (4, 15, 29) or full-text articles (30) aimed at benchmarking the text mining effort. However in this study, we explored the comparison of text-based extraction against events annotated in an independently curated pathway knowledge base. The performance of our system is comparable to the other state of the art system against text-based annotations (2, 3). This study further allowed us to identify the gaps between the current state of the art in literature mining and the demands of text mining assisted pathway curation. However we believe that our current system will be useful for finding the evidence needed for curation of the pathways. We plan to explore the following steps to improve text mining assisted pathway curation:

1) Improve the state of the art in gene normalization, which we hope to improve since we are working on this task in parallel for the BioCreative 4 Track3 (31);

2) Explore hybrid approaches by combining the rule-based system with machine learning approach to reduce the amount of manual effort required to tune the systems to new data sets;

3) Understand the pathway curation workflow and design annotation schema and corpora for pathway curation. The current available corpora limit the annotation to single abstracts or articles. Quite often we need to synthesize information across articles. But we realize that it is not possible without the understanding the pathway curation workflow;

4) Assess the needs of pathway curators to set more realistic and achievable text mining goals. We realize that working closely with the database curators and building an intuitive interface to facilitate pathway curation will not only help us understand the curation workflow but also help improve the state of the art in literature mining significantly.

6. Acknowledgments

The authors acknowledge that the study was supported by two grants: National Science Foundation ABI:0845523 and National Library of Medicine R01LM009959 grants. The authors also acknowledge the support received from Centre for Individualized Medicine, Mayo Clinic.

References

1. Baumgartner WA, Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics. 2007;**23**(13):i41-i8.
2. Björne J, Ginter F, Pyysalo S, Tsujii Ji, Salakoski T. Complex event extraction at PubMed scale. Bioinformatics. 2010;**26**(12):i382-i90.
3. Björne J, Heimonen J, Ginter F, Airola A, Pahikkala T, Salakoski T. Extracting complex biological events with rich graph-based feature sets. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task; 2009: Association for Computational Linguistics; 2009. p. 10-8.
4. Liu H, Komandur, R., Verspoor, K. From graphs to events: A subgraph matching approach for information extraction from biomedical text. In proceedings, BioNLP-ST'11 Workshop; 2011; 2011.
5. BioNLP Shared Task 2013. [cited; Available from: http://2013.bionlp-st.org/
6. Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics. 2005;**21**(suppl 2):ii252-ii8.
7. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC bioinformatics. 2004;**5**(1):147.
8. Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio—the analysis and navigation of molecular networks. Bioinformatics. 2003;**19**(16):2155-7.
9. Rzhetsky A, Iossifov I, Koike T, et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. Journal of biomedical informatics. 2004;**37**(1):43-53.
10. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC bioinformatics. 2005;**6**(Suppl 1):S1.
11. Lu Z, Kao H, Wei C, et al. The Gene Normalization Task in BioCreative III. BMC Bioinformatics. 2011; **12**(Suppl 8):S2.
12. Kim J-D, Pyysalo S, Ohta T, Bossy R, Nguyen N, Tsujii Ji. Overview of bionlp shared task 2011. Proceedings of the BioNLP Shared Task 2011 Workshop; Association for Computational Linguistics; 2011. p. 1-6.
13. Kim JD, Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J. Overview of BioNLP'09 shared task on event extraction. Proceedings of the Workshop on BioNLP: Shared Task; 2009; 2009. p. 1-9.
14. Bandy J, Milward D, McQuay S. Mining protein–protein interactions from published literature using Linguamatics I2E. Protein Networks and Pathway Analysis: Springer; 2009. p. 3-13.

15. Van Landeghem S, Ginter F, Van de Peer Y, Salakoski T. EVEX: a PubMed-scale resource for homology-based generalization of text mining predictions. Proceedings of BioNLP 2011 Workshop; 2011: Association for Computational Linguistics; 2011. p. 28-37.
16. Kemper B, Matsuzaki T, Matsuoka Y, et al. PathText: a text mining integrator for biological pathway visualizations. Bioinformatics. 2010;**26**(12):i374.
17. Miwa M, Thompson P, McNaught J, Kell DB, Ananiadou S. Extracting semantically enriched events from biomedical literature. BMC bioinformatics. 2012;**13**(1):108.
18. Schmidt CJ, Sun L, Arighi CN, et al. Pathway curation: Application of text-mining tools eGIFT and RLIMS-P. Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on; 2012: IEEE; 2012. p. 523-8.
19. Hewett M, Oliver DE, Rubin DL, et al. PharmGKB: the pharmacogenetics knowledge base. Nucleic acids research. 2002;**30**(1):163-5.
20. Brill E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational linguistics. 1995;**21**(4):543-65.
21. Kim JD, Ohta, T., Tateisi, Y., Tsujii, J. GENIA corpus : a semantically annotated corpus for bio-textmining. Bioinformatics. 2003;**19**(Suppl1):i180-i2.
22. Florian R, Ngai G. Fast transformation-based learning toolkit. Johns Hopkins University, http://nlp.cs.jhu.edu/rflorian/fntbl/documentation.html; 2001.
23. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K. A biological named entity recognizer. Pac Symp Biocomput; 2003; p. 427-438.
24. Kennedy C, Boguraev B. Anaphora for everyone: pronominal anaphora resoluation without a parser. Proceedings of the 16th conference on Computational linguistics-Volume 1; 1996: Association for Computational Linguistics; 1996. p. 113-8.
25. PubMed database. [cited; Available from: http://www.ncbi.nlm.nih.gov/pubmed/
26. Joshi-Tope G, Gillespie M, Vastrik I, et al. Reactome: a knowledgebase of biological pathways. Nucleic acids research. 2005;**33**(suppl 1):D428-D32.
27. Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: Molecular biology database and retrieval system. Methods in enzymology. 1996;**266**:141-62.
28. Wermter J, Tomanek K, Hahn U. High-performance gene name normalization with GeNo. Bioinformatics. 2009;**25**(6):815-21.
29. Liu H, Hunter L, Kešelj V, Verspoor K. Approximate Subgraph Matching-Based Literature Mining for Biomedical Events and Relations. PloS one. 2013;**8**(4):e60954.
30. Garten Y, Altman R. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC bioinformatics. 2009;**10**(Suppl 2):S6.
31. Biocreative IV 2013 [cited; Available from: http://www.biocreative.org/tasks/biocreative-iv/track-3-CTD/