A Knowledge Extraction Framework for Biomedical Pathways

Sanda Harabagiu, PhD[1] and **Cosmin Adrian Bejan, PhD**[1]
[1]**Human Language Technology Research Institute,**
The University of Texas at Dallas, Richardson, TX

Abstract

In this paper we present a novel knowledge extraction framework that is based on semantic parsing. The semantic information originates in a variety of resources, but one in particular, namely BioFrameNet, is central to the characterization of complex events and processes that form biomedical pathways. The paper discusses the promising results of semantic parsing and explains how these results can be used for capturing complex medical knowledge.

Introduction

Most previous work in the field of biomedical information extraction has focused on the recognition and classification of named entities mentioned in texts into sets of semantic categories most relevant to researchers in biological sciences and other medical professionals. For example, given a sentence like *"Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of isochamaejasmin"*, named entity recognition systems tailored to the biomedical domain can be used to categorize three different entities of interest: (1) *NF-kappaB* (a protein), (2) *isochamaejasmin* (an organic compound or flavonoid), and (3) *anti-apoptotic effect* (a treatment outcome). While the best of these systems (such as the GENIA tagger[1] developed jointly by the University of Tokyo and the University of Manchester) have made significant progress towards the extraction of entity information from collections of biomedical texts, little work has focused on the extraction of more semantically-complex types of information (such as events, processes, or relations), despite a number of recent advances in the natural language processing (NLP) literature.

More recently, attempts have been made to extract biochemical pathways based on syntactic information[2]. However, no qualitative analysis of the usefulness of the extracted information could be performed, as semantic models were not available, nor were methods for assessing their clinical value. In this paper, we present a knowledge extraction framework that is used for capturing the semantic information relevant to the blood clotting process. This framework is informed by a probabilistic model of blood clotting, reported by Makin and Narayanan[3]. In addition, it has several novelties. First, it uses a modeling framework based on Coordinated Probabilistic Relational Models (CPRMs) which extend graphical models (Bayes nets) with the ability to represent and analyze complex dynamics between events, thus providing an ideal framework for capturing complex biomedical processes represented by biological pathways. Second, the knowledge extraction framework integrates the semantics of CPRMs with the linguistics semantics defined into BioFrameNet. BioFrameNet is a new semantic resource of biological information that captures the frame semantics that is implicit into the text expressions that refer to events, states, and relations from biological pathways. Third, the semantic integration between the CPRM models and the frames defined in BioFrameNet is made possible by an event ontology that offers a rich relational model and a domain-driven context for the semantic extraction context.

To be able to recognize bioframes in scientific articles two resources are needed: (1) BioFrameNet, which encodes the semantic definitions of the bioframes along with exemplar annotations; (2) a semantic parser that uses for its training a variety of other linguistic processing components, including a part of speech (POS) tagger that performs lexical disambiguation, a syntactic parser, which recognizes the syntactic dependencies in biomedical literature, and a term recognizer and disambiguation capable of identifying names of factors, proteins, etc. The syntactic dependencies and the semantic categories contribute to the quality of the semantic parsing based on BioFrameNet. Other important contributors are the relation recognizer and the capability of identifying coreferring entities in text.

Bioframes captured from medical articles, as well as relations between bioframes are linked into an ontology of biological events and states. Information encoded in the ontologies may con-

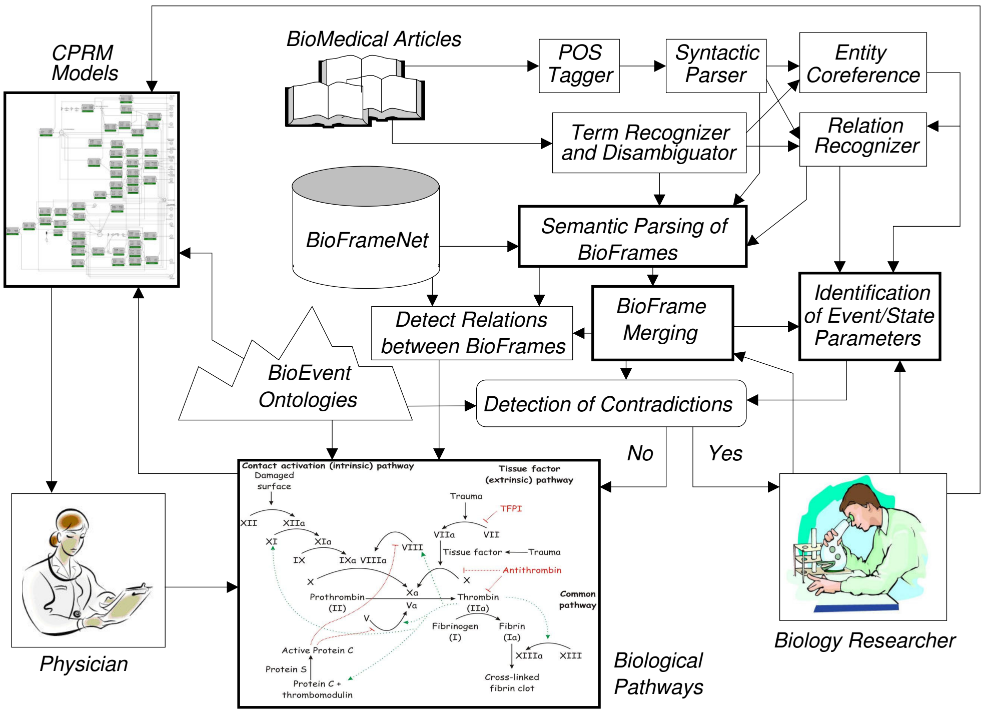

Figure 1: Knowledge extraction framework.

tradict the new knowledge extracted from text. To detect such cases, a mechanism of detecting contradictions is incorporated in the knowledge extraction framework illustrated in Figure 1. When no contradictions between the ontological and textual knowledge exists, the information is encoded into the biological pathway and linked to the CPRM model that allows clinicians to use it and to perform simulations of various pathologies associated with the medical process described in the pathway.

Central to knowledge extraction framework is the ability to perform semantic parsing based on BioFrameNet. In this paper we report on results of experiments conducted to generate semantic parsers for the biomedical pathways relevant to blood clotting.

BioFrameNet - A New Semantic Resource

Technical terms (in this case, biomedical terms) are different from everyday words in part because their meaning is not acquired by children experiencing something and hearing talk about the situation and gradually making a correlation. Rather, technical vocabulary has ascribed meaning, defined by experts and learned consciously, by formal study; this applies not only to entity nouns like *phosphorylase*, but also event nouns like *phosphorylation*, adjectives like *phosphorylated*, etc., all of which evoke the same semantic frame with the same semantic roles (or frame elements), although this fact is not usually represented in other knowledge-bases. To meet the challenges of biomedical texts, FrameNet was transformed into BioFrameNet[4,5]. Table 1 summarizes the changes that will need to be made to produce BioFrameNet.

Table 2 shows the manual annotation of a single sentence in this domain, just to give the reader a feeling for the level of analysis being performed. The sentence contains 6 frame-evoking expressions, so there is a separate annotation set for each frame, on a separate row. Three of the needed frames already exist in FrameNet; the three new frames are marked with asterisks.

This is not the first time the use of frame semantics for biomedical texts has been suggested. Andrew Dolbey has completed dissertation research on the frames and frame elements (FEs) needed to describe intracellular transport[6]. More recently, Uematsu and colleagues[2] compared the biomedical event structures of texts in the GENIA corpus with the corresponding FrameNet frames using a manual alignment; they concluded that the linguistically-oriented semantics of FrameNet "could be favorable to domain portability of a text mining system."

	Current FrameNet	**BioFrameNet**
Source of lexical items	Linguistic data from general corpora; everyday language	Pre-existing ontologies and term banks of technical vocabulary
Meaning model	Semantic frames based on common sense world knowledge; minimal semantic typing on FEs	Frames and FEs linked to precise mathematical model of clotting and to external ontologies
Definitional process	Conventional meaning based on usage	Ascribed meaning = definition by experts
Evaluation task	Various tasks, depending on user	Testing against a specific biomedical model

Table 1: Comparison of current FrameNet and BioFrameNet.

* Protein Activation Frame: The ACTIVATION [$_{\text{Activated Entity}}$ of factor Xa] requires the assemblage of the tenase complex (Ca2+ and factors VIIIa, IX, and X) on the surface of the activated platelets.
Necessity Frame: [$_{\text{Enabled Situations}}$ The activation of factor Xa] requires [$_{\text{Precondition}}$ the assemblage of the tenase complex (Ca2+ and factors VIIIa, IX, and X) on the surface of the activated platelets.]
* Molecular Assembly Frame: The activation of factor Xa requires the assemblage [$_{\text{Resulting Assembly}}$ of the tenase complex (Ca2+ and factors VIIIa, IX, and X)] [$_{\text{Location}}$ on the surface of the activated platelets.]
* Molecular complex Frame: The activation of factor Xa requires the assemblage of the [$_{\text{Function}}$ tenase] complex (Ca2+ and factors VIIIa, IX, and X) [$_{\text{Location}}$ on the surface of the activated platelets].
Part inner outer Frame: The activation of factor Xa requires the assemblage of the tenase complex (Ca2+ and factors VIIIa, IX, and X) on the surface [$_{\text{Whole}}$ of the activated platelets].
* Activation Frame: The activation of factor Xa requires the assemblage of the tenase complex (Ca2+ and factors VIIIa, IX, and X) on the surface of the activated [$_{\text{Activated Entity}}$ platelets].

Table 2: Annotation of a sentence in the coagulation domain.

Semantic Parsing of Pathway Events

The task of bioframe identification resembles the task of word sense disambiguation (WSD) in NLP. We enhanced a semantic parser that received top marks at the 2004 Senseval-3 FrameNet parsing evaluation; it correctly identified relevant FrameNet frames with over 90% F-measure and labeled frame elements with greater than 80% F-Measure. We enhanced this system by considering the same classification models, namely Support Vector Machine (SVM) and Maximum Entropy (ME), but accessed additional lexical and semantic resources from BioFrameNet. Our enhancement is based on (1) the availability of semantic information for the biomedical domain originating in the GENIA annotations and several other resources; and (2) the application of active learning. Figure 2 illustrates the enhanced architecture based on the semantic parser reported by Bejan and Hathaway[7]. The architecture illustrated in Figure 2 has a resource manager that integrates a number of additional resources besides BioFrameNet. We used GENIA[8], the MeSH vocabulary, the Reactome knowledge base, WordNet, and UniProt.

To be able to take advantage of additional examples provided by an expert, we employed selective sampling of the unlabeled biomedical texts that contain the same target word used for bioframe identification. The Active Learning-1 paradigm therefore depends on the confidence estimation of the Maximum Entropy (ME) learner. However, as reported by Kristjansson and colleagues[9], using ME, each of the FEs identified for a bioframe is evaluated independently, so that the propagation of correction after expert user input is minimal. In order to generalize the maximum entropy model used for frame identification to a Conditional Random Field (CRF) which has been shown to perform well on extraction tasks[10,11].

When an expert corrects or accepts an identified bioframe, a novel example is given to the multi-class classifier for bioframe disambiguation. We presented to the user the same example with the frame element annotations resulting from the semantic parser. The semantic parser uses a large set of diverse features, which were reported in [12,13,14,7,15]. These features were grouped into independent sets, enabling a multi-view active learning paradigm, used in the Active Learning-2 module of the architecture illustrated in Figure 2. We repeatedly learned a classifier based on the SVM and ME learning algorithms in each view and queried the expert on examples in which they predicted different FE labels of the bioframe.

The semantic parsing based on BioFrameNet consists of (task 1) the bioframe disambiguation (BFD); and (task 2) the labeling of its FEs. Unlike general frames, bioframes use fairly distinct target words, thus in our experiments, BioFrameNet dis-

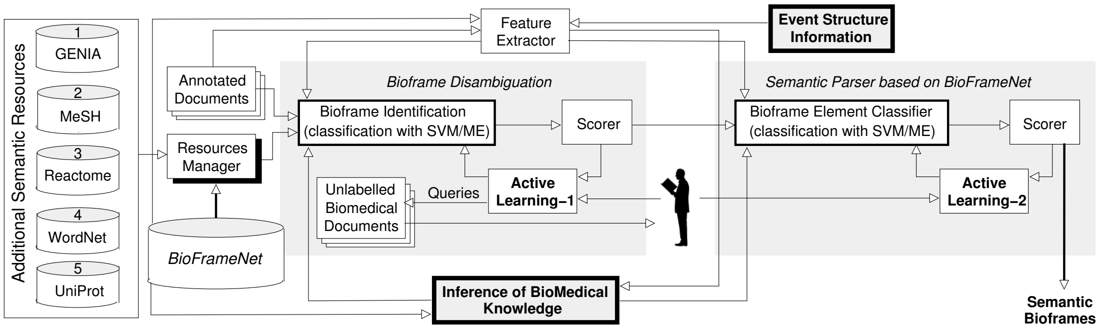
Additional Semantic Resources
1
GENIA
2
MeSH
3
Reactome
4
WordNet
5
UniProt
Annotated Documents
Resources Manager
BioFrameNet
Feature Extractor
Event Structure Information
Biofame Disambiguation
Bioframe Identification (classification with SVM/ME)
Scorer
Active Learning–1
Unlabelled Biomedical Documents
Queries
Semantic Parser based on BioFrameNet
Bioframe Element Classifier (classification with SVM/ME)
Scorer
Active Learning–2
Inference of BioMedical Knowledge
Semantic Bioframes

Figure 2: Enhanced architecture for BioFrameNet based semantic parsing.

ambiguation was performed through a lexical unit look-up. The automatic labeling of bioframe elements was cast as a sequence of two classifiers: (1) a classifier that identifies boundaries of each FE; and (2) a second classifier that labels them.

Frame Element Identification (FEI)

The problem of detecting the FE boundaries is cast as the problem of deciding whether or not a constituent is a valid candidate for a FE. We considered a binary classifier over the entire BioFrameNet data and extract features for each constituent from a syntactic parse tree. Because this experimental setup allows training the binary classifier on a large set of examples, the best feature combination consists of a restrained number of features, listed in Figure 3 selected by the Active-Learning-2.

Frame Element Classification (FEC)

The task of assigning labels to FEs is performed by 58 multi-class classifiers, where each classifier corresponds to a frame in BioFrameNet. Each classifier was trained using the most features listed in Figure 3.

Experimental Results

We conducted a series of experiments that allowed the evaluation of the semantic parser. First, we have been interested to evaluate the quality of the classifier that detects the FE boundaries. Table 3 illustrates the results obtained for this classifier when using a SVM classifier and when using a ME classifier.

Resource	Classifier	Precision	Recall	F1-measure
BioFrameNet	SVM	75.38	88.06	81.23
	ME	73.22	85.72	78.98
FrameNet	SVM	73.65	87.08	79.80
	ME	71.30	83.20	76.79

Table 3: Results obtained for the FEI task.

The superior results used when training BioFrameNet data can be explained by (a) the fact that we have trained a syntactic parser on medical texts and (b) used a Named Entity Recognizer that could recognize a large variety of concepts that are relevant to the biochemical pathways that we studied (related to blood clotting). In addition, the results listed in Table 4 for the BioFrameNet data are produced when using active learning implemented with CPRMs. Active learning has enhanced the precision by 22%. Table 5 illustrates the results of the same task when the general FrameNet data was used on the corpus evaluated in SemEval–2007.

Resource	Classifier	Accuracy
BioFrameNet	SVM	92.34
	ME	89.43
FrameNet	SVM	80.20
	ME	88.93

Table 4: Results obtained for the FEC task.

In our experiments the training and the testing articles were selected from those annotated by the creators of the BioFrameNet.

Conclusions

The results of semantic parsing used for extracting knowledge relevant to biological pathways are very encouraging and they show promise for a novel and complex framework for learning and using through simulations complex knowledge that characterizes such complex processes as blood clotting.

References

1. Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John Mcnaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In *Ad-*

NO	BFD	FEI	FEC	Feature Description
01	v			TW UNIGRAMS: The words, stem words and part of speech (POS) unigrams that are adjacent to target word expressions;
02	v			TW BIGRAMS: The words, stem words and POS bigrams that are adjacent to target word expressions;
03	v			TW WORD: The target word expression;
04			v	TW STEM: The stem word(s) of the target word expression;
05	v			TW POS: The POS of the target word;
06			v	TW CLASS: The lexical class of the target word, e.g. verb, noun, adjective;
07	v		v	NAMED ENTITY FLAGS: Set of binary features indicating whether a constituent contains, is contained or exactly identifies a named entity;
08	v			VERB WSD: If the target word is a verb, extract the head noun of the direct object and the prepositional object included in the verbal phrase;
09	v			NOUN WSD: If the target word is a noun, extract the head word of the verbal phrase that is in a verb–subject or verb–object relation with the noun;
10	v			ADJECTIVE WSD: If the target word is an adjective, extract the head noun that is modified by the adjective;
11			v	PHRASE TYPE: The syntactic category of the constituent;
12		v	v	DIRECTED PATH: Path in the syntactic parse tree between the constituent and the target word preserving the movement direction;
13			v	UNDIRECTED PATH: Same syntactic path as DIRECTED PATH without preserving the movement direction;
14		v		PARTIAL PATH: Path from the constituent to the earlier common ancestor of the target word and the constituent;
15		v	v	POSITION: Test whether the constituent contains the target word, or appears before or after the target word;
16			v	VOICE: Test if the verbal target word has active or passive construction;
17		v	v	HW: The head word of the constituent;
18			v	HW POS: The syntactic head POS of the constituent;
19			v	HW STEM: The stem word of the constituent's head word;
20			v	CW: The content word of the constituent computed as described by Surdeanu and colleagues [13].
21			v	CW POS: The POS corresponding to the content word;
22			v	CW STEM: Stemmed content word;
23			v	GOVERNING CATEGORY: Test whether the noun phrase constituents are dominated by verbal phrases or sentence phrases;
24		v		SYNTACTIC DISTANCE: The length of the syntactic path;
25			v	PP FIRST WORD: If the constituent is a prepositional phrase, return the first word in the phrase;
26			v	HUMAN: Test whether the constituent phrase is either a personal pronoun or a hyponym of first sense of PERSON synset in WordNet;
27			v	CONSTITUENTS NUMBER: The number of candidate FEs;
28			v	CONSTITUENTS LIST: Constituents labels list of the candidate FEs;
29			v	SAME CLAUSE: Test whether the constituent is in the same clause with the target word;
30			v	GF: The grammatical function of a candidate frame element;
31			v	GF LIST: The list of grammatical functions associated to the candidate FEs;
32		v	v	FRAME: The name of the semantic frame that is evoked by the target word;
33			v	NP SISTER: Determine whether the constituent has a noun phrase sister;
34			v	FIRST/LAST WORD: Return the first/last word of the constituent phrase;
35		v	v	FIRST/LAST POS: Return the first/last POS in the constituent;
36			v	LEFT/RIGHT SISTER LABEL: Return the left/right sibling constituent label;
37			v	LEFT/RIGHT SISTER HEAD: Return the left/right sibling head word;
38			v	LEFT/RIGHT SISTER STEM HEAD: Return the left/right sibling stemmed head word;
39			v	LEFT/RIGHT SISTER POS HEAD: Return the left/right sibling head POS;
40			v	TW STEM & HW STEM: Join of TW STEM and HW STEM;
41			v	TW STEM & PHRASE TYPE: Join of TW STEM and PHRASE TYPE;
42			v	VOICE & POSITION: Join of VOICE and POSITION.

Figure 3: Feature set for extracting frame semantic structures.

Task	Best Model	Accuracy		
Frame Disambiguation	SVM	76.71		
FE Label Classification	ME	88.93		
		Precision	Recall	F1-measure
FE Boundary Detection	SVM	73.65	87.08	79.80

Table 5: Task results on the validation set.

vances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746, pages 382–392.

2. Sumire Uematsu, Jin-Dong Kim, and Jun'ich Tsujii. 2009. Bridging the Gap between Domain-Oriented and Linguistically-Oriented Semantics. In *Proceedings of the Workshop on BioNLP. ACL Workshop*.
3. Joseph Makin and Srini Narayanan. 2008. A Hybrid-System Model of Human Blood Clotting. Technical Report, International Computer Science Institute.
4. Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING-ACL-98*.
5. Charles Fillmore, Christopher Johnson, and Miriam Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16.3:235–250.
6. Andrew Dolbey, Michael Ellsworth, and Jan Scheffczyk. 2006. BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. In *KR-MED*.
7. Cosmin Adrian Bejan and Chris Hathaway. 2007. UTD-SRL: A Pipeline Architecture for Extracting Frame Semantic Structures. In *Proceedings of SemEval-2007*, pages 460–463.
8. Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. The GENIA Corpus: An Annotated Research Abstract Corpus in Molecular Biology Domain. pages 82–86.
9. Trausti Kristjansson, Aron Cullota, Paul Viola, and Andrew Mccallum. 2004. Interactive Information Extraction with Constrained Conditional Random Fields.
10. Andrew Mccallum and W. Li. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons.
11. D Pinto, A. Mccallum X. Wei, and Bruce Croft. 2003. Table Extraction Using Conditional Random Fields.
12. Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistic*, 28(3):496–530.
13. Mihai Surdeanu, Sanda M. Harabagiu, John Williams, and John Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of ACL*.
14. Nianwen Xue and Marta Palmer. 2004. Calibrating Features for Semantic Role Labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
15. Sameer Pradhan, Kadri Hacioglu, Valeri Krugler, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2005. Support Vector Learning for Semantic Argument Classification. *Journal of Machine Learning Research*, 60(1):11–39.