KR-MED 2006 "Biomedical Ontology in Action"
November 8, 2006, Baltimore, Maryland, USA

BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies

Andrew Dolbey[1,2], Michael Ellsworth[3], and Jan Scheffczyk[3]
[1]University of Colorado, Center for Computational Pharmacology, Aurora, Colorado
[2]University of California Berkeley, Linguistics Dept., Berkeley, California
andy.dolbey@gmail.com
[3]International Computer Science Institute, Berkeley, California
{infinity, jan}@ICSI.Berkeley.EDU

Biomedical domain ontologies could be better put to use for automatic semantic linguistic processing if we could map them to lexical resources that model the linguistic phenomena encountered in this domain, e.g., complex noun phrase structures that reference specific biological entity names and processes. In this paper, we introduce BioFrameNet – a domain-specific FrameNet extension. BioFrameNet uses Frame semantics to express the meaning of natural language, is augmented with domain-specific semantic relations, and links to biomedical ontologies like the Gene Ontology – all of which are expressed in the Description Logic (DL) variant of OWL. Thus, BioFrameNet annotations of natural-language text precisely map to biomedical ontologies, which in turn facilitates inference using DL reasoners.

INTRODUCTION

Many currently available Natural Language Processing (NLP) tools limit language processing to levels of linguistic detail that involve form, e.g. Part of Speech tagging and syntactic parsing (Stanford Parser[1]). In this endeavor, they are quite successful. What is missing is, however, an automated analysis of meaning. With the vast amount of knowledge expressed via textual resources publicly available, we see an increasing demand to include automated meaning analysis in our NLP toolkits. We intend to develop tools that provide users with fast access to what is being discussed in a large set of documents of potential interest. This will include tasks like entity recognition, question answering, thread discovery, and summarization.

At the same time, there has been a rapid emergence of a great number of ontological resources including the Gene Ontology and Entrez Gene Database. This is particularly true in the domains of molecular biology and biomedicine. This emergence offers opportunities to achieve new levels of success in Natural Language Understanding (NLU), the task of automatically determining and extracting meaning from texts. But for this to happen, the *interface between form and meaning* must also be modeled.

We propose to model this interface by combining Frame semantics [1] with links to domain-specific biomedical ontologies, all of which we express in the Description Logic (DL) variant of OWL in order to facilitate inference by means of DL reasoners like Racer [2] or FaCT++ [3]. The primary goal of *BioFrameNet* (BioFN), a resource currently being developed, is to model the mapping of form and meaning in the linguistic structures that occur in biomedical texts.

BioFN is the dissertation project of the first author. It extends and refines FrameNet (FN) [4] – a lexicon for English, which is based on Frame semantics [1]. A semantic Frame (hereafter simply Frame) represents a set of concepts associated with an event or a state, ranging from simple (Bringing, Placing) to complex (Revenge, Criminal_process). For each Frame, a set of roles (or arguments), called Frame Elements (FEs), is defined, about 10 per Frame. We say that a word can evoke a Frame, and its syntactic dependents can fill the FE slots. Semantic types (STs) constrain the types of FE fillers. Semantic relations between Frames are captured in Frame relations, each with corresponding FE-to-FE mappings. Syntactic-semantic mapping in FN and BioFN is captured by means of defining sets of valence patterns, where triples of FE, grammatical function, and phrase types observed in natural language text are enumerated for each Lexical Unit (LU) = word sense. FN currently contains more than 780 Frames, covering roughly 10,000 LUs; these are supported by more than 135,000 FrameNet-annotated example sentences.[2]

[1] See http://www-nlp.stanford.edu/software/lex-parser.shtml.

[2] For further information on FrameNet, see http://Framenet.icsi.berkeley.edu.

This paper proceeds as follows: First, we briefly discuss related work. Second, we introduce BioFN. We then propose mappings to biomedical ontologies and show our technique for creating these mappings, which will use OWL DL. This is followed by a description of how biomedical natural-language text can be annotated using BioFN and how these annotations can be put to work for reasoning by expressing them in OWL DL. Finally, we discuss lessons learned and show how others can benefit from our approach.

RELATED WORK

The HunterLab[3] transport ontology has also been developed to model transport processes [5], and shares certain properties with BioFN. However, by using the explicit semantics provided in (Bio)FrameNet, we get, for free, a more inclusive formal analysis of the semantics of a transport event. Therefore, we would not need to produce and specify separate axioms with systems such as PAL. We model this semantics directly with BioFN.

BioFN uses our OWL DL translation of FrameNet [6] and augments it with domain-specific semantic relations between FEs and links to GO, the Entrez Gene database, and the protein transport knowledge representation created by the HunterLab [4]. Thereby, BioFN leverages on our experiences with linking FrameNet to the Standard Upper Merged Ontology (SUMO) [7], which, so far, are not domain specific.

PASBio [8] is a project that aims to produce definitions of Predicate Argument Structure (PAS) frames, similar in spirit to PropBank [9], but focusing on the domain of molecular biology. Although the PAS frames have much in common with BioFN valence patterns, it does not offer a direct linking of the predicates or their arguments to domain or general merged ontologies. The work of Korhonen et. al. [10] reports on the automatic induction of lexical verb classes for the domain of biomedicine, where the classes link together syntactic and semantic properties of groups of verbs, much like the work of Levin [11] and Kipper [12]. Providing syntax-semantic linking at the level of lexical class helps compensate for missing individual lexical entries, but runs the risk of error for individual predicates that share most of the semantics of the class, but nevertheless show divergent linking behavior [13].

"Kicktionary"[5] is a multi-lingual application of the FrameNet methodology to the domain of soccer. The kicktionary structure can be brought into accordance with ontological principles [14] and thus be mapped to soccer ontologies, e.g. [15]. BioFN can be extended to a multi-lingual lexicon based on the principles shown in [14]. Additional domain-specific semantic relations between FEs distinguish BioFN from the kicktionary.

BIOFRAMENET

BioFN is a lexical resource modeled after FrameNet (FN) proper [4]. Indeed, it is an extension of FrameNet, one that builds on – i.e., includes and links to – the general FN frames. The primary data of the project is a collection of text data items (discussed later in the paper) annotated by biologists associated with the HunterLab of the University of Colorado Health Sciences Center.[6] The text data has a primary focus on the domain concept of intracellular transport. The annotations were carried out with a reported consistency score of over 90%. For purposes of this work, the annotations provide reliable indications of the locations of the spans of text that correspond to FE values.

The primary additions to FN proper consist of semantic frames relevant to the domain of molecular biology. As is the case elsewhere in FrameNet, these frames are linked with other frames in a set of clearly defined ways. For each Frame, there is a definition of Frame elements – the "arguments" or "slots" that the Frame licenses. Each Frame is also associated with a list of predicators, the lexical units that evoke the Frame.

For example, BioFN includes the domain-specific Frame "Transport_intracellular", which describes the biological process of intracellular transport of molecular entities. The Frame elements for this Frame are Cargo (the transported entity), Carrier (the transporting entity), Origin (the start point of transport), and Destination (the end point of transport). The following predicators, with part of speech appended to the name, are among the more frequently occurring lexical units that evoke this Frame:

translocate.v, translocation.n, transport.v, transport.n, shift.v, shuttle.v, export.v

[3] Center for Computational Pharmacology, of University of Colorado Health Sciences Center, directed by Dr. Lawrence Hunter.

[4] See http://compbio.uchsc.edu/grifs/transport/schema.shtml.

[5] See http://www.kicktionary.de.

[6] See http://compbio.uchsc.edu.

In many cases, new Frames added are related to other Frames that already exist in FN proper. For example, the Transport_intracellular Frame is included as a subtype of the Brining Frame, a Frame concerning the movement of a Theme and an Agent and/or Carrier.[7] It should be noted that the focus of the texts in the HunterLab corpus data will place a limit on the number and coverage of biomedical Frames included in the initial version of BioFN.

An important question that arises when incorporating new Frames in FN is whether or not a new Frame is warranted. This ties in to a general lumping vs. splitting decision the FN team often faces [4]. When the Frame under consideration is for domain specific semantics, there are special pros and cons to splitting with a new Frame. One disadvantage is an increase in the complexity of the network of Frames. We believe this is outweighed by the advantage of being able to specify richer information and constraints specific to the particular domain. Thus it will be possible to elaborate and constrain the general semantics of bringing with meaning, entailments, and domain knowledge particular to the event of intracellular transport. This shows up most clearly in the linking of Frames and FEs to domain specific ontologies. Maintaining close relations with more general Frames allows access to the more general semantics as well, thus simplifying the task of connecting the Bio-specific Frame to related Frames, since many of the connections will already be modeled in the general vocabulary.

MAPPING BIOFRAMENET TO DOMAIN ONTOLOGIES

The domain ontologies we used for BioFN's mappings are GO, Entrez Gene, and a small transport knowledge representation schema of the HunterLab (HL) [5]. These were chosen for three reasons. First, and foremost, the community consensus is that GO and Entrez Gene are reliable, trusted, and actively updated. Second, all three are free and publicly available. And third, the HunterLab transport schema is currently under active development, and itself makes use of the other two domain resources.

There are two levels of mappings that must be formalized. On one level, the Transport_intracellular Frame and its Frame Elements are described. This frame is mapped to a node in the GO biological_process tree, "protein transport". The FEs "Origin" and "Destination" are mapped to nodes in the cellular_component tree. The FEs "Cargo" and "Carrier" are disjunctively mapped to either an Entrez Gene element, or otherwise to the HL items "molecule or molecular complex" or "molecular part". This is shown in Fig. 1. On another level, we also need to map SemanticType (ST) filler constraints to the same (or related) ontologies[8].

We have developed an approach that automatically translates a crucial portion of FrameNet (and its specializations) and annotations into OWL DL [6]. Fig. 1 shows the OWL DL translation of the Transport_intracellular Frame.

Frames, STs, and FEs are represented as OWL classes, where an FE class represents the type of the FE fillers. Frame and FE relations are modeled as existential restrictions on these classes; inheritance is represented via OWL subclassing. This way the generated ontology stays OWL DL – a crucial precondition for automated reasoning. The connection between a Frame and an FE filler is represented by the "hasFE" relation. We do so because in OWL relations are not first-class objects.[9] For example, the FE filler for Origin_relation is in fact a relation but we represent it as an OWL class in order to connect spans of text to it and to have the possibility of specifying relations to other FEs (like the Origin FE, which fills Origin_relation).

BioFN also uses the FrameNet STs, which are linked to the Standard Upper Merged Ontology (SUMO) [7]. Thereby, BioFN immediately benefits from SUMO's rich axiomatization.

We augment the OWL translation of BioFN with links to the Gene Ontology (GO), the Entrez Gene Ontology (EG), the HunterLab transport ontology (HL), and Smith's Relation Ontology (RO) [16]. These links are represented via subclass relationships and appear as bold arrows in Fig. 1.[10]

For example, the Frame class Transport_intracellular is a subclass of GO:Biological_process. Our way of modeling supports the use of OWL's expressive class language, e.g., to create anonymous union classes. For example, the class Cargo is a subclass of the

[7] See http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=118&frame=Bringing&.

[8] Mappings of the ST filler constraints are not shown in Fig. 1.

[9] OWL does not support relations between relations other than inheritance.

[10] The subclass relationships were added by hand in the OWL representation, they are not expressible in FrameNet itself.

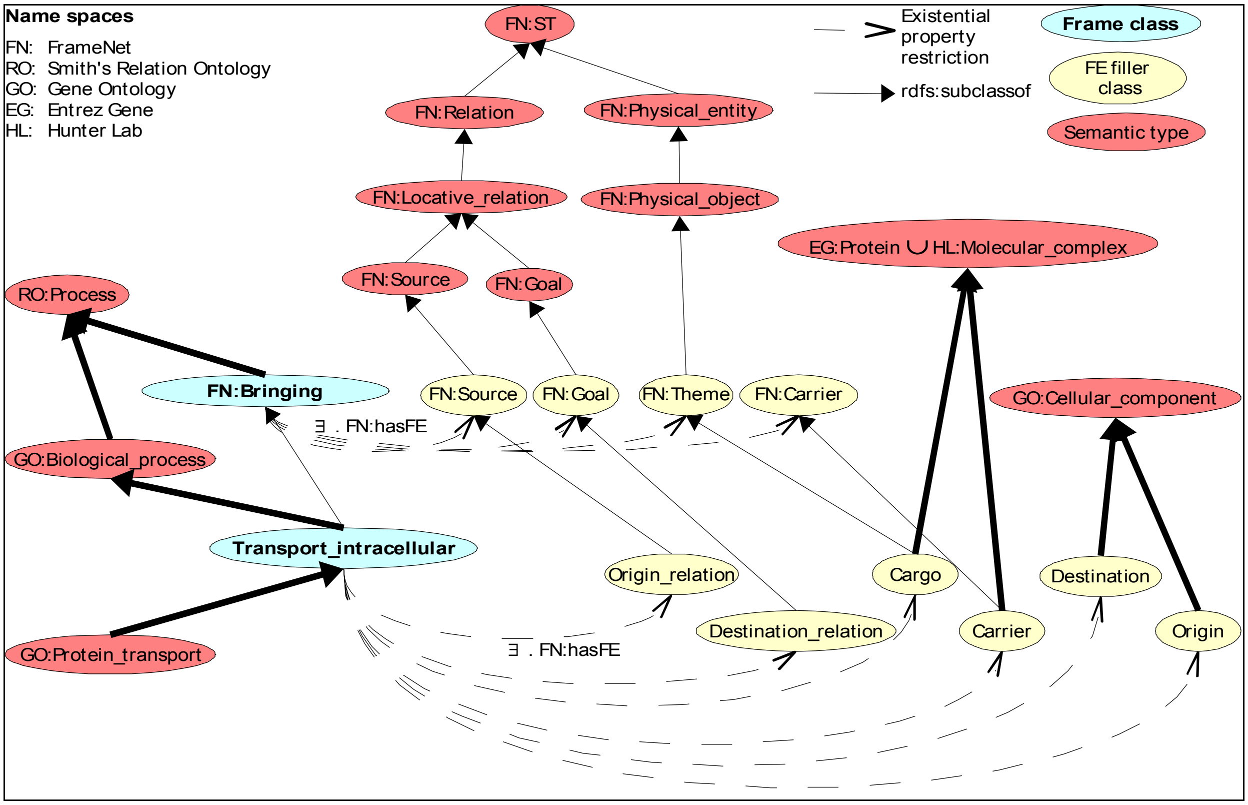

Figure 1 – OWL translation of Transport_intracellular Frame.

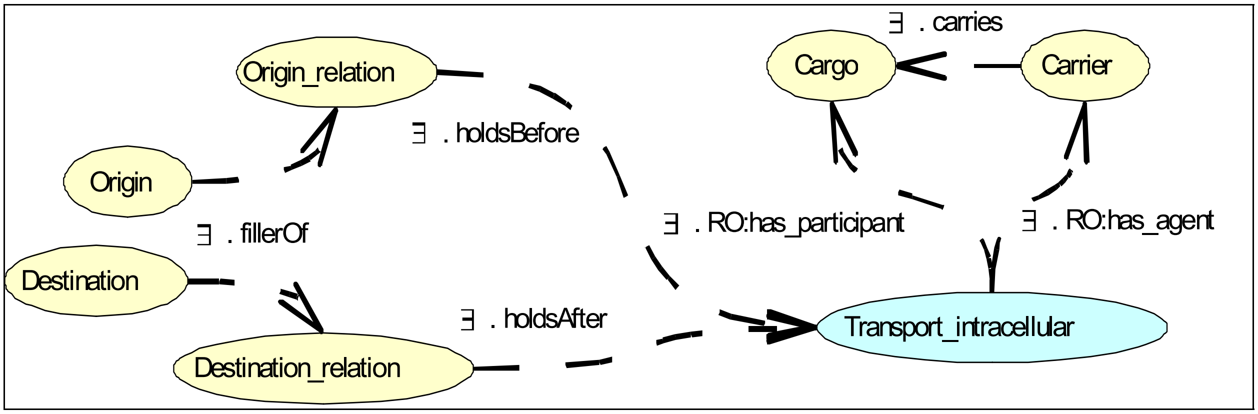

Figure 2 – Extra FE Relations.

union of EG:Protein and HL:Molecular_complex.

In order to aid reasoning we specify further semantic relations between FE filler classes of the same Frame (see Fig. 2).

Wherever possible we use relations and constraints defined in Smith's Relation Ontology in order to

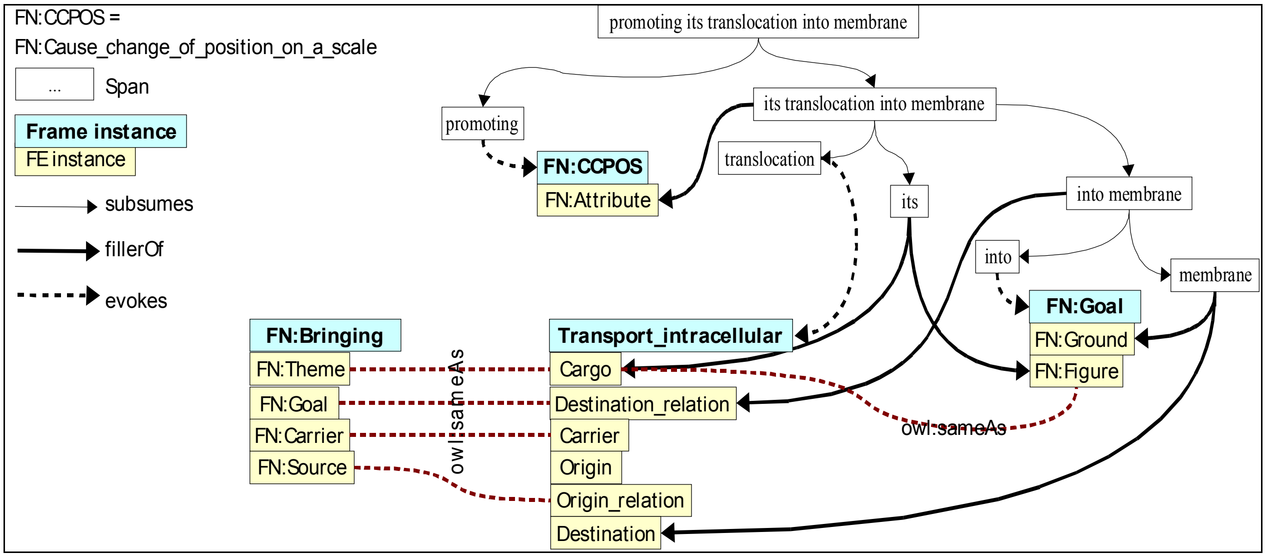

Figure 3 –Annotation of example GRIF.

leverage from their formal definitions. For other relations we are working on a formal definition much similar to those proposed in [7]. For example, we say that each Transport_intracellular process must have a participant of type Cargo and an agent of type Carrier, which carries the cargo. Again, these relations are expressed as existential class restrictions.

TEXT DATA EXAMPLE

A particular kind of text available in the domain of molecular biology is that of GRIF, "Gene References in Function"[11]. GRIFs provide relatively short descriptions of the function(s) of particular genes. This kind of text serves as a useful initial target of analysis due to their close links to particular genes in publicly available and widely used databases of genes and gene products. In this paper, as an illustrative example we will show our BioFN analysis of a portion of a particular GRIF (the analyzed portion is underlined):

> SCD1 deficiency specifically increases
> CTP:choline cytidylyltransferase activity
> by <u>promoting its translocation into membrane</u>
> and enhances phosphatidylcholine
> biosynthesis in liver

This GRIF makes an assertion about the transport of one entity, "CTP:choline", into the cellular component "membrane". There are other assertions that can be inferred in this GRIF, both about the nature of the transport process itself and about other processes that are also involved. Due to space limits, we will not include a BioFN analysis of the language that evokes these other inferred phenomena, including "deficiency", "activity", "enhances", and "biosynthesis", though the analysis of these items has been done in a similar fashion.

SEMANTIC REPRESENTATION OF BIOMEDICAL NATURAL LANGUAGE TEXT

From the full text annotation of a GRIF, we automatically generate an Annotation Ontology that uses the BioFN Ontology as a template. An Annotation Ontology populates the BioFN Ontology with instances of Frames and FEs as well as the actual text data and satisfies the existential constraints (which express Frame and FE relations).

Fig. 3 shows a part of the Annotation Ontology for our example GRIF. Text spans are represented as instances that fill FE instances or evoke Frame instances.[12]

Spans can syntactically include other spans, which we express by the *subsumes* relation. Whenever a span fills more than one FE we generate an owl:sameAs relation between the FE fillers, based on this syntactic evidence. Since we need to satisfy all the constraints from the BioFN ontology, we generate for each existential restriction on some relation R with target class C a new instance of C

[11] See http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html.

[12] For simplicity we let instances share the names of their respective classes and omit classes. Also, we omit hasFE relations that point from a Frame to each of its FEs.

and connect this instance by the relation R. Also, for FE mappings (including inheritance) we generate owl:sameAs relations between the generated FE instances, which aid reasoning [6]. Thus we generate a new instance of the FrameNet:Bringing Frame because the Transport_intracellular Frame inherits from FrameNet:Bringing. We also express that the connected FE instances are the same. Therefore, the span "its" in the example GRIF actually evokes three FEs, all of which have an identical filler: Cargo (in Transport_intracellular), FrameNet:Figure (in FrameNet:Goal), and FrameNet:Theme (in FrameNet:Bringing).

Generation of BioFN-specific semantic relations between FEs and Frames is straightforward. Fig. 4 shows the additional semantic relations generated for the Transport_intracellular Frame instance.

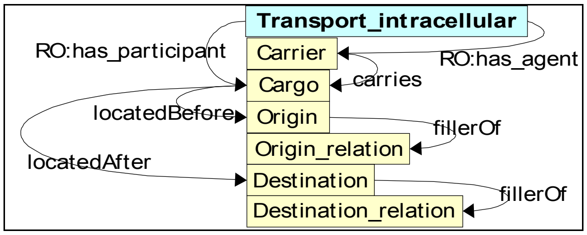

Figure 4 –Transport_intracellular relations.

In Fig. 5, we represent an instance of the Dimension Frame bound (via the Cargo FE) to an instance of the Transport_intracellular Frame.

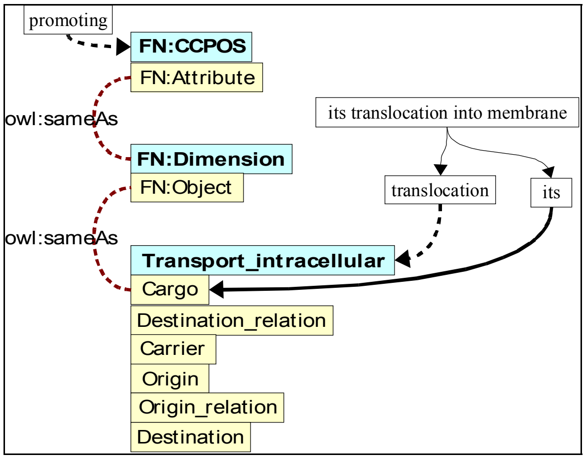

Figure 5 – Dimension Frame : an instance of metonomy.

This interpretation arises through a metonymic relation between events and quantities which is beyond the scope of the current paper; the interpretation with Transport_intracellular filling the Attribute role of Cause_change_of_position_on_a_scale ought to be discarded since Attributes and Events are disjoint.

LESSONS LEARNED

Changing FrameNet

Even during our preliminary investigation of annotation for BioFN, we have discovered new LUs (e.g. *promote.v* and *enhance.v*) for the Cause_change_of_position_on_a_scale Frame. This is despite FrameNet having studied this concept in some detail, showing definitively that domain-specific annotation will be necessary to capture the vocabulary of the biological domain.

This elaboration of FN is similar in spirit to other current efforts to link FN with other similar resources like VerbNet, PropBank, and Cyc [17]. These resources will be used for comparison and evaluation, when appropriate, as BioFN work proceeds.

Changing biomedical ontologies

The lack of reference to GO in many entries of the HunterLab ontology will make integrated processing very difficult. The ultimate usefulness of BioFN will rely on a merged ontology and knowledge-base, with seamless references to FrameNet, SUMO, the HunterLab ontology, GO, and Entrez-Gene. The cross-reference between the ontologies required by BioFN will reveal errors and unnecessary points of difference between these ontologies, thus enabling their improvement.

The impact of our approach for reasoning

We have already demonstrated elsewhere [6] that our OWL DL model of FrameNet is usable for the kind of reasoning needed for question answering, using queries in Racer. With some loss of power, the method could be made more efficient by implementation as a graph-traversal or querying of an SQL version of the ontology.

However, since the approach was not integrated with a large-scale ontology, it has so far been hampered by variations in the linguistic form of objects not captured in FrameNet or even in WordNet. Since BioFN will be integrated with the appropriate ontologies from its inception, the same approach should be much more powerful using the BioFN resource (together with its associated ontologies) than it is with FrameNet resources alone. In addition, applications built with BioFN or FrameNet will make use of other NLP tools such as stemmers and lemmatizers for handling variation in linguistic form. We predict having similar success with BioFN in carrying out Question Answering and a variety of other NLU tasks.

How can others benefit from our approach?

Current biological ontologies have very few relations and events, and considerably less experience with modeling language than FrameNet. The work demonstrated here shows that FrameNet-style ontological descriptions of language can be integrated with information from biological ontologies using the expressive power of Description Logic.

How can our technique be applied to other problems/domains?

Since FrameNet provides a general-domain (if limited) ontology, it seems promising to apply our methodology to other domains that have associated ontologies and a need for textual processing. One area in which some work has already proceeded is event tracking in the terrorism domain [18].

CONCLUSIONS, FUTURE WORK

In this paper we introduced BioFN – a domain-specific FrameNet extension. BioFN bridges form and meaning of natural-language biomedical texts by (1) new domain-specific Frames, (2) links to established biomedical ontologies like GO and Entrez Gene, and (3) domain-specific semantic relations between FEs. We model BioFN as an OWL DL ontology, which we populate with BioFN annotations of biomedical texts. Thus, natural-language biomedical texts become available for DL-based reasoning.

Since the BioFN project is dissertation work currently in progress, we are not yet able to provide full numbers and statistics for coverage of the data under consideration and counts and definitions of all the new Frames that need to be created. This is indeed one of the primary goals of the dissertation: a complete analysis of the collection of GRIFs in the HunterLab corpus. An analysis of coverage of WMD-related[13] text by the FN project shows that analyzing texts in a particular domain does yield significantly greater coverage of new texts in the same genre.[14]

In the future, we will enhance BioFN with more biomedical Frames and richer semantic relations. Also, we aim at an (OWL DL + SWRL) axiomatization of domain-specific relations much in the fashion of [16]. We will conduct experiments in automatic parsing using the Shalmaneser Frame parser [19]. GO and Entrez Gene classes provide narrow semantic types, which can significantly aid automatic Frame recognition and role (i.e., FE) labeling.

Finally, we envision operationalizing the generation of ontology instances of metonymy by unpacking types of metonymy in the ontology itself. Currently, to the best of our knowledge, no ontology includes the explicit indications of metonymy that this would require, but ongoing work [7] is moving in this direction.

We are confident that the technique we use for BioFN scales well to other domains. Domain-specific lexical resources that are linked to domain-specific ontologies – under the roof of an upper lexical resource (like FrameNet), an upper ontology (like SUMO), and modeled using a common formal language (like OWL DL) – seem to be a reasonable approach to natural-language understanding. Thus, in the long run, we see FrameNet as a backbone of several domain-specific FrameNets that in turn are linked to domain-specific ontologies.

Acknowledgments

This work is supported in part by grant #5R01-LM008111-02 from the National Library of Medicine.

References

1. C. J. Fillmore. Frame semantics and the nature of language. Annals of the New York Academy of Sciences, (280):20-32, 1976.
2. M. Wessel and R. Möller. A high performance semantic web query answering engine. In Proc. International Workshop on Description Logics, 2005.

[13] WMD = Weapons of Mass Destruction.

[14] See updated FN FAQ for further discussion of this point, at http://framenet.icsi.berkeley.edu/index.php?option=content&task=category&sectionid=11&id=86&Itemid=49.

3. I. Horrocks. The FaCT system. In H. de Swart, editor, Automated Reasoning with Analytic Tableaux and Related Methods: International Conference Tableaux'98, number 1397 in Lecture Notes in Artificial Intelligence, pages 307-312. Springer-Verlag, May 1998.
4. J. Ruppenhofer, M. Ellsworth, M. R. Petruck, and C. R. Johnson. FrameNet: Theory and Practice. ICSI Berkeley, 2005. http://framenet.icsi.berkeley.edu.
5. Z. Lu. Mining protein transport data from GeneRIFs. University of Colorado, Center for Computational Pharmacology presentation, 2006.
6. J. Scheffczyk, C. F. Baker, and S. Narayanan. Ontology-based reasoning about lexical resources. In Proc. of OntoLex 2006: Interfacing Ontologies and Lexical Resources for Semantic Web Technologies, pages 1-8, Genoa, Italy, 2006.
7. J. Scheffczyk, A. Pease, and M. Ellsworth. Linking FrameNet to the suggested upper merged ontology. In Proc. of FOIS 2006, Baltimore, MD, 2006. to appear.
8. T. Wattarujeekrit, P. K. Shah, and N. Collier. PASBio: predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics, 5:155, 2004.
9. P. Kingsbury and M. Palmer. From Treebank to PropBank. In Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC2002), 1989-1993. Las Palmas, Spain, 2002.
10. A. Korhonen, Y. Krymolowski, and N. Collier. Automatic Classification of Verbs in Biomedical Texts. To appear in Proceedings of ACL-COLING 2006. Sydney, Australia, 2006.
11. B. Levin. English Verb Classes and Alternations. Chicago University Press, Chicago. 1993.
12. K. Kipper, H. T. Dang, M. Palmer. Class-Based Construction of a Verb Lexicon. AAAI/IAAI 2000: 691-696. North Falmouth, MA, 2000.
13. C. F. Baker and J. Ruppenhofer. FrameNet's Frames vs. Levin's Verb Classes. In J. Larson and M. Paster (Eds.), Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society. 27-38. 2002.
14. T. Schmidt. Interfacing lexical and ontological information in a multilingual soccer FrameNet. In Proc. of OntoLex 2006: Interfacing Ontologies and Lexical Resources for Semantic Web Technologies, pages 75-81, Genoa, Italy, 2006.
15. P. Buitelaar et al. Generating and visualizing a soccer knowledge base. In Proc. of the EACL'06 Demo Session, Trento, Italy, 2006.
16. B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector, and C. Rosse. Relations in biomedical ontologies. Genome Biology, 6(5), 2005.
17. A. Meyers, A. C. Fang, L. Ferro, et. al.. Annotation Compatibility Working Group Report, Coling/ACL, Sydney, Australia, 2006.
18. S. Sinha and S. Narayanan. Model based answer selection. In Textual Inference in Question Answering Workshop, AAAI 2005, Pittsburgh, PA, 2005
19. K. Erk and S. Padó. Shalmaneser – a flexible toolbox for semantic role assignment. In Proc. of LREC 2006, Genoa, Italy, 2006, to appear.