Bridging the Gap between Domain-Oriented and Linguistically-Oriented Semantics

Sumire Uematsu Jin-Dong Kim Jun'ich Tsujii
Department of Computer Science
Graduate School of Information Science and Technology
University of Tokyo
7-3-1 Hongo Bunkyo-ku Tokyo 113-0033 Japan
{uematsu,jdkim,tsujii}@is.s.u-tokyo.ac.jp

Abstract

This paper compares domain-oriented and linguistically-oriented semantics, based on the GENIA event corpus and FrameNet. While the domain-oriented semantic structures are direct targets of Text Mining (TM), their extraction from text is not straghtforward due to the diversity of linguistic expressions. The extraction of linguistically-oriented semactics is more straghtforward, and has been studied independentely of specific domains. In order to find a use of the domain-independent research achievements for TM, we aim at linking classes of the two types of semantics. The classes were connected by analyzing linguistically-oriented semantics of the expressions that mention one biological class. With the obtained relationship between the classes, we discuss a link between TM and linguistically-oriented semantics.

1 Introduction

This paper compares the linguistically-oriented and domain-oriented semantics of the GENIA event corpus, and suggests a factor for utilizing NLP techniques for Text Mining (TM) in the bio-medical domain.

The increasing number of scientific articles in the bio-medical domain has contributed in drawing considerable attention to NLP-based TM. An important step in NLP-based TM is obtaining the domain-oriented semantics of sentences, as shown at the bottom of figure 1. The BioInfer (Pyysalo et al., 2007) and the GENIA event corpus (Kim et al., 2008) provide annotations of such semantic structures on collections of bio-medical articles. Domain-oriented semantic structures are valuable assets because their representation suits information needs in the domain; however, the extraction of such structures is difficult due to the large gap between the text and these structures.

On the other hand, the extraction of linguistically-oriented semantics from text has long been studied in computational linguistics, and has recently been formalized as Semantic Role Labeling (Gildea and Jurafsky, 2002), and semantic structure extraction (Baker et al., 2007)(Surdeanu et al., 2008). Semantic structures in such tasks are exemplified in the middle of figure 1. The linguistically-oriented semantic structures are easier to extract, although the information is not practical to the domain.

We aim at relating linguistically-oriented frames of semantics with domain-oriented classes, thus making a step forward in utilizing the computational linguistic resources for the bio-medical TM. Of all the differences in the two type of semantics, we focused on the fact that the former frames are more sensitive to the perspective imposed by the sentence writer. In the right hand-side example of figure 1, the linguistically-oriented structure treats *PBMC*, a cell entity, as an agent; however the bio-medical structure reflects the scientific view that there are no agents, objects acting with intention, in bio-molecular phenomena.

As a preliminary investigation, we selected four representative classes of bio-molecular phenomena; Localization, Binding, Cell_adhesion, and Gene_expression, and investigated domain-oriented annotations for the classes in the GENIA

Proceedings of the Workshop on BioNLP, pages 162–170,
Boulder, Colorado, June 2009.

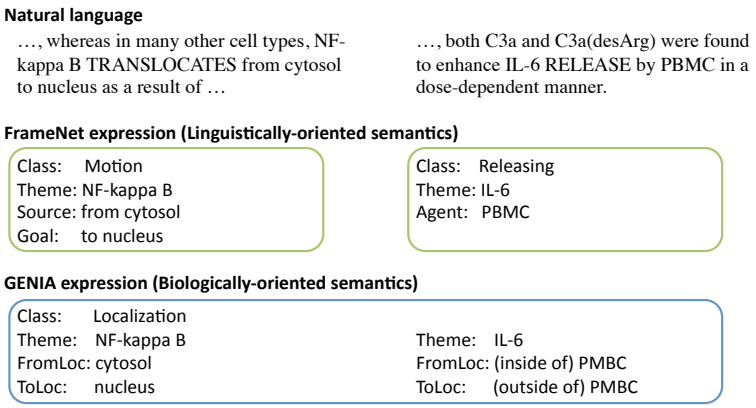

Figure 1: A comparison of the linguistically-oriented and biologically-oriented structure of semantics

event corpus. Expressions mentioning the four classes were examined and manually classified into linguistically-oriented frames, represented by those defined in FrameNet (Baker et al., 1998). FN frames associated to a bio-molecular event class constitute a list of possible perspectives in mentioning phenomena of the class.

The rest of this paper is structured in the following way: Section 2 reviews the existing work on semantic structures and expression varieties in the bio-medical domain, and provides a comparison to our work. In section 3, we describe the GENIA event corpus, and the FrameNet frames used as linguistically-oriented classes in our investigation. Sections 4 and 5 explain the methods and results of the corpus investigation; in particular the sections investigate how the linguistic frames were associated to the domain-oriented classes of semantics. Finally, we provide discussion and conclusion in section 6 and 7.

2 Related Work

Existing work on semantics approached domain-oriented semantic structures from linguistically-oriented semantics. In contrast, our approach uses domain-oriented semantics to find the linguistic semantics that represent them. We believe that the two different approaches could complement each other.

The PASbio(Wattarujeekrit et al., 2004) proposes Predicate Argument Structures (PASs), a type of linguistically-oriented semantic structures, for domain-specific lexical items, based on PASs defined in PropBank(Wattarujeekrit et al., 2004) and NomBank(Meyers et al., 2004). The PASs are defined per lexical item, and is therefore distinct from a biologically-oriented representation of events. (Cohen et al., 2008) investigated syntactic alternations of verbs and their nominalized forms which occurred in the PennBioIE corpus(Kulick et al., 2004), whilst keeping PASs of the PASBio in their minds.

The BioFrameNet(Dolbey et al., 2006) is an attempt to extend the FrameNet with specific frames to the bio-medical domain, and to apply the frames to corpus annotation. Our attempts were similar, in that both were: 1) utilizing the FN frames or their extensions to classify mentions of biological events, and 2) relating the frames and the FEs (roles of participants) with classes in domain ontologies; e.g. the Gene Ontology(Ashburner et al., 2000).

As far as the authors know, it is the first attempt to explicitly address the problem of linking linguistically-oriented and domain-oriented frames of semantics. However, it has been indirectly studied through works on TM or Relation Extraction using linguistically-oriented semantic structures as features, such as in the case with (Harabagiu et al., 2005).

3 Corpora

We used domain-oriented annotations of the GENIA event corpus and linguistically-oriented frames defined in FrameNet (FN), to link domain-oriented and linguistically-oriented frames of semantics. We briefly describe these resources next.

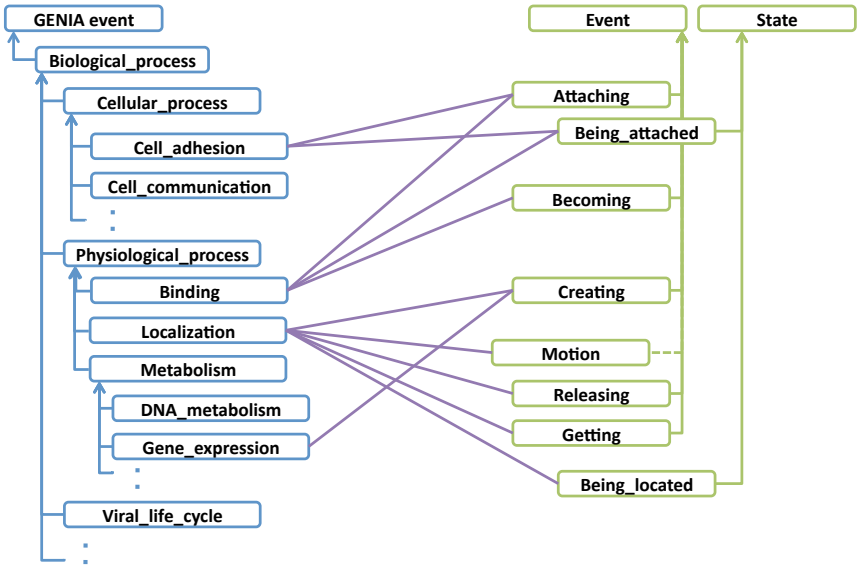

Figure 2: The resulting relationship between linguistically-oriented and biologically-oriented frames.

The GENIA event corpus consists of 1,000 Medline abstracts; that is, 9,372 sentences annotated with domain-oriented semantic structures. The annotation was completed for all mentions of biological events, and resulted in 6,114 identified events. Examples of annotated event structures are shown at the bottom of figure 1. Each structure has attributes *type* and *theme*s, which respectively show the biological class of the mentioned event and phrases expressing the event participants. The event classes are defined based on the terms in the Gene Ontology. For example, the Localization class in the GENIA event corpus is defined as an equivalent of the GO term Localization (GO0051179). The event classifications used in the corpus are depicted in the left hand-side of figure 2. Arrows in the figure depict the inheritance relations defined in the GENIA event ontology. For instance, the Localization class is defined as a type of Physiological_process. Each of the annotated structures has additional attributes that point phrases that the annotator of the structure used as a clue. Among the attributes, the *clueType* attribute shows a clue phrase to the event class. In our investigation, the attribute was treated as a predicate, or an equivalent of the lexical unit in the FN.

FN is a network of frames that are are linguistically-oriented classifications of semantics. A FN frame is defined as "a script-like conceptual structure that describes a particular type of situation, object, or event and the participants and propositions involved in it," and is associated with words, or lexical units, evoking the frame. For instance, the verbs *move*, *go* and *fly* are lexical units of the Motion frame, and they share the same semantic structure. Each FN frame has annotation examples forming an attestation of semantic overlap between the lexical units. Additionally, FN defines several types of frame-frame relations; e.g. inheritance, precedence, subframe, etc. The right hand-side of figure 2 shows some FN frames and inheritance relationships between them. The FN provides linguistically-oriented classifications of event mentions based on surface expressions, and also shows abstract relations between the frames.

4 Additional Annotation

Our aim is to link linguistically-oriented and domain-oriented frames of the bio-medical text's semantics. A major problem in this task was that there were no annotated corpora with both types of semantic structures. Therefore, we decided to concentrate on the mentions of a few classes of biological phenomena, and to annotate samples of the mentions with linguistically-oriented structures conforming to

Freq.	Keyword	Frame
693	binding	Attaching
247	bind	Attaching
125	interaction	Attaching, Being_attached
120	complex	–
99	bound	Attaching, Being_attached
91	interact	Attaching, Being_attached
61	form	Becoming
52	crosslink	Attaching
46	formation	Becoming

Table 1: The most frequent keywords of the Binding class, mentioned 2,006 times in total.

Freq.	Keyword	Frame
131	translocation	Motion
81	secretion	Releasing
75	release	Releasing
32	secrete	Releasing
25	mobilization	Motion
23	localization	Being_located
20	uptake	Getting
18	translocate	Motion
15	expression	Creating
9	present	Being_located

Table 2: The most frequent keywords of the Localization class, mentioned 582 times in total.

the FrameNet annotations.

The following provides the annotation procedures. First, we collected linguistic expressions that mention each of the selected GENIA event classes from the GENIA event corpus. We then sampled and annotated them with their linguistically-oriented semantics which conformed to the FrameNet.

4.1 Target Classes and Keywords

We concentrated mainly on the mentions of four GENIA classes; Localization, Binding, Cell_adhesion, and Gene_expression. Gene_expression, Binding, and Localization are three of the most frequent four classes in the GENIA event corpus.[1] Binding and Localization are the two most primitive molecular events. The Cell_adhesion class was included as a comparison for the Binding class.

Counting keywords for mentioning events was close to automatic. We extracted phrases pointed by a *clueType* attribute from each event structure. We then tokenized the phrases, performed a simple stemming on the tokens, and counted the resulting words. The stemming process simply replaced each inflected word to its stem by consulting a small list of inflected words with their stems. Manual work was only used in making the small list.

4.2 FN Annotation

A major challenge encountered in annotating a sampled expression with a semantic structure conforming to FN, was in the assignment of a FN frame to the mention. Our decision was based on the following four points: 1) keywords used in the mention, 2) description of FN frames, 3) syntactic positions of the event participants, and 4) frame-frame relations.

The first indicates that a FN frame became a candidate frame for the mention, if the keyword in the mention is a lexical unit of the FN frame. FN frames and their lexical units could be easily checked by consulting the FN dictionary. If there were no entries for the keyword in the dictionary, synonyms or words in the keyword's definition were used. For example, the verb *translocate* has no entries in the FN dictionary, and the frames for verbs such as *move* were used instead.

For the second point, we discarded FN frames that are either evoked by a completely different sense of the keyword, or too specific of a non-biological situations.

Before we assigned a FN frame to each mention, we manually examined the syntactic positions of all event participants present in the sampled GENIA mentions. Combinations of the syntactic position and event participants observed for a keyword were compared with sample annotations of the candidate FN frames.

We checked frame-frame relations between the candidate frames, because they can be regarded as evidence that shows that the conception of the frames is related. For our aim, it was sufficient to choose a set of frames that best describes the different perspectives for mentioning one type of molecular phenomena. Even when some keywords seemed to be dissimilar in the three points mentioned above,

[1] Except correlation and regulation classes which express relational information rather than events.

Freq.	Keyword	Frame
98	adhesion	Being_attached
19	adherence	Being_attached
16	interaction	Being_attached, Attaching
15	binding	Attaching
8	adherent	Being_attached

Table 3: The most frequent keywords of the Cell_adhesion class, mentioned 193 times in total.

Freq.	Keyword	Frame
1513	expression	Creating
357	express	Creating
239	production	Creating
71	overexpression	Creating
69	produce	Creating
62	synthesis	Creating

Table 4: The most frequent keywords of the Gene_expression class, mentioned 2,769 times in total.

a single frame could be assigned to them if it was quite clear that they shared a similar perspective. The frame-frame relations provided in the FN were treated as clues to the similarity.

Keywords frequently used in each event class are listed in tables 1, 2, 3, and 4, with the final assignment of FN frames to each keyword.

5 Analysis

After the linguistic annotation was performed, we compared the GENIA event structure and the frame structure of each sampled expression, and obtained relations of the GENIA class-FN frame and GENIA slot-FN participant. The resulting relationships between FN frames and the four GENIA classes demonstrate a gap between linguistically-oriented and domain-oriented classification of events, as shown in figure 2.

The relations can be explained by decomposing it into two cases: 1) 1-to-n mappings, and 2) n-to-1 mappings. The n-to-n mapping from GENIA to FN can then be regarded as a mix of the two cases. In the following sections, the two cases are described in detail. Further, we show conversion examples of a FN structure to a GENIA event structure, which were supported by the obtained GENIA participant-FN participant relations.

5.1 1-to-N Mapping: Different Perspectives on the Same Phenomena

A 1-to-n mapping from GENIA to FN can be explained as the case where the same molecular phenomena are expressed from different perspectives.

5.1.1 Binding Expressed in Multiple frames

The Binding class in GENIA is defined as "the selective, often stoichiometric interaction of a molecule with one or more specific sites on another molecule." We associated the class with three frames, and two frames of the three, Attaching and Becoming frames, represent different perspectives for mentioning the class. The Being_attached frame shares the same conception as Attaching, but expresses states instead of events. See table 1 for keywords of the class, and the frames assigned to the words.

Attaching: In the perspective represented by this frame, a binding phenomenon was recognized as a event in which protein molecules were simply attached to one another.

> [The 3'-CAGGTG E-box$_{\text{Item}}$] could BIND [USF proteins$_{\text{Goal}}$], $\cdots$
> (PubMed ID 10037751, Event IDs E11, E12, E13)

Becoming: In the perspective represented by this frame, a product of a binding event was treated, on the surface, as a different entity from the original parts.

> When activated, [glucocorticoid receptors$_{\text{Entity}}$] FORM [a dimer$_{\text{Final_category}}$] $\cdots$
> (PubMed ID 10191934, Event ID E5)

This type of expression was possible because a product of a binding often obtains a different functionality, and can be treated as a different type of entity. Note that this frame was not associated with the Cell_adhesion class described in section 5.2.

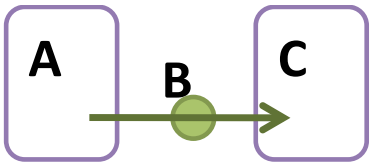

Figure 3: A schematic figure of translocation.

Being_attached: Annotators recognized a protein binding event from the sentence below, which basically mentions a state of the NF-kB.

> In T cells and T cell lines, [NF-kB$_{\text{Item}}$] is BOUND [to a cytoplasmic proteic inhibitor, the IkB$_{\text{Goal}}$].
> (PubMed ID 1958222, Event ID E2, E102)

Although this type of expression shares a similar point of view with the Attaching frame, we classified these expressions into the Being_attached frame in order to demonstrate cases in which a prerequisite Binding event was inferred from a state.

5.1.2 Translocation Expressed in Multiple Frames

The Localization class in the GENIA corpus is defined as a class for "any process by which a cell, a substance, or a cellular entity, such as a protein complex or organelle, is *transported to, and/or maintained in* a specific location." Sampled expressions of the class separated into mentions of a process, by which an entity was *transported to* a specific location, and those of the process in which an entity was *maintained in* a specific location. We concentrate on the former in this section, and describe the latter in section 5.1.3.

We associated the frames: Motion, Releasing and Getting with what we call translocation events, or Localization events in which an entity was *transported to* a specific location. Figure 3 provides a schematic representation of a translocation event. Each of the three frames had a different perspective in expressing the translocations. See table 2 for keywords of the frames.

Motion: This group consists of expressions centered on the translocated entities of the translocation - namely, B in the figure 3.

> [NK cell NFAT$_{\text{Theme}}$] $\cdots$ MIGRATES [to the nucleus$_{\text{Goal}}$] upon stimulation,$\cdots$
> (PubMed ID 7650486, Event ID E33)
> Activation of T lymphocytes $\cdots$ results in TRANSLOCATION [of the transcription factors NF-kappa B, AP-1, NFAT, and STAT$_{\text{Theme}}$] [from the cytoplasm$_{\text{Source}}$] [into the nucleus$_{\text{Goal}}$].
> (PubMed ID 9834092, Event ID E67)

These expressions are similar to those of the Motion frame in the FN.

> [Her foot$_{\text{Theme}}$] MOVED [from the brake$_{\text{Source}}$] [to the accelerator$_{\text{Goal}}$] and the car glided forward.

Releasing: This group consists of expressions centered on a starting point of the translocation - namely, A in the figure 3.

> In [unstimulated cells which$_{\text{Agent}}$] do not SECRETE [IL-2$_{\text{Theme}}$], only Sp1 binds to this region, $\cdots$
> (PubMed ID 7673240, Event ID E13)
> Activation of NF-kappaB is thought to be required for [cytokine$_{\text{Theme}}$] RELEASE [from LPS-responsive cells$_{\text{Agent}}$], $\cdots$
> (PubMed ID 1007564, Event ID E14)

The verbal keywords occurred as a transitive in most cases, and had subjects and objects that expressed starting points and entities in the translocations. This is a typical syntactic pattern of the Releasing frame, if we regarded an Agent in the FN as a starting point of the movement of a Theme.

> [The police$_{\text{Agent}}$] RELEASED [the suspect$_{\text{Theme}}$].

Getting: This group consists of expressions centered on a goal point of the translocation - namely, C in figure 3. We assumed that this group has an opposite point of view from the Releasing frame. The noun *uptake* was found to be a keyword in this group.

> The integral membrane $\cdots$ appears to play a physiological role in binding and UPTAKE [of Ox LDL$_{\text{Theme}}$] [by monocyte-macrophages$_{\text{Recipient}}$], $\cdots$
> (PubMed ID 9285527, Event ID E10)

To summarize, we observed three groups of expressions that mention translocation events, and each group represented different perspectives to mention the events. Each of the groups and the associated frame seemed similar, in that they shared similar keywords and possible syntactic positions to express the event participant.

5.1.3 Localization excluding Translocation Expressed in Multiple Frames

Localization events excluding translocations were expressed in the Being_located and Creating frames.

Being_located: This group consists of expressions that simply mention an entity in a specific location.

> $\cdots$ [recombinant NFAT1$_{\text{Theme}}$] LOCALIZES [in the cytoplasm of transiently transfected T cells$_{\text{Location}}$] $\cdots$
> (PubMed ID 8668213, Event ID E23)

Creating: A noun *expression* was observed to be used by instances mentioning the presence of proteins.

> horbol esters are required to induce [AIM/CD69$_{\text{Created_entity}}$] Cell-surface EXPRESSION as well as $\cdots$
> (PubMed ID 1545132, Event ID E12)

*Expression*s in these cases indicate an abbreviation for *gene expression*, which is a event of Gene_expression class. This type of overlap between the Localization and Gene_expression is explained in section 5.2.2

5.2 N-to-1 Mapping: Same Conception for Different Molecular Phenomenon

In contrast to the cases described in section 5.1, the same conception could be applied to different biological phenomena.

5.2.1 Shared Conception for Binding and Cell_adhesion

Molecular events classified into Binding and Cell_adhesion shared the conception that two entities were attached to each other. However, types of the entities involved are different. They are: the protein molecule in Binding, and cell in Cell_adhesion.

> CD36 is a cell surface glycoprotein $\cdots$, which INTERACTS with thrombospondin, $\cdots$, and erythrocytes parasitized with Plasmodium falciparum.

In the sentence above, an event involving *a cell surface glycoprotein* and *thrombospondin* was recognized as a Binding, whereas an event involving *a cell surface glycoprotein* and *erythrocytes* was classified as a Cell_adhesion event.

5.2.2 Shared Expressions of Localization and Gene_expression

Both Localization and Gene_expression classes are connected with the Creating frame. Some Localization events have a dependency on the Gene_expression event. Protein molecules are made in events classified into the Gene_expression class.

> [Th1 cells$_{\text{Creator}}$] PRODUCE [IL-2 and IFN-gamma$_{\text{Created_entity}}$], $\cdots$
> (PubMed ID 10226884, Event ID E11, E12)

The molecules are then translocated somewhere. Consequently, localized protein molecules might indicate a Gene_expression event, and a phrase "protein expression" was occasionally recognized as mentioning a Localization.

> horbol esters are required to induce [AIM/CD69$_{\text{Created_entity}}$] cell-surface EXPRESSION as well as $\cdots$
> (PubMed ID 1545132, Event ID E12)

5.3 Conversion of FN Structures to GENIA Events

During the investigation, we compared participant slots of GENIA and FN structures, in addition to the structures themselves. Figures 4 and 5 depict conversion examples from a FN structure and its participants to a GENIA structure, with the domain-oriented type of each participant entity. The conversions were supported by samples, and need quantitative evaluation.

6 Discussion

By annotating sentences of the GENIA event corpus with semantic structures conforming to FrameNet, we explicitly compared linguistically-oriented and

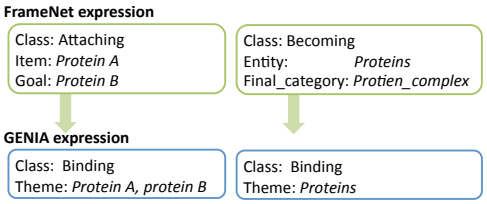

Figure 4: FN-to-GENIA conversions for Binding

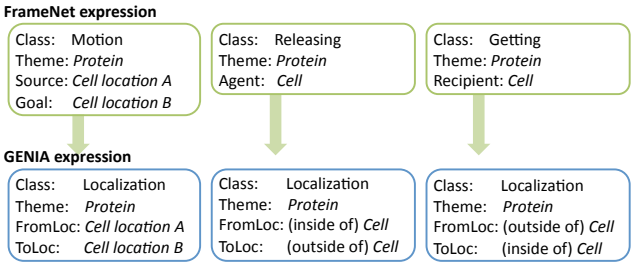

Figure 5: FN-to-GENIA conversions for Localization.

domain-oriented semantics of the bio-molecular articles. Our preliminary result illustrates the gap between the two type of semantics, and a relationship between them. We discuss development of a Text Mining (TM) system, in association with the extraction of linguistically-oriented semantics, which has been studied independently of TM.

First, our result would show that TM involves at least two qualitatively different tasks. One task is related to our results; that is, recognizing equivalent events which are expressed from different perspectives, and hence expressed by using different linguistic frames, and at the same time distinguishing event mentions which share the same linguistic frame but belong to different domain classes. Our investigation indicates that this task is mainly dependent on domain knowledge and how a phenomenon can be conceptualized. Another task of TM is the extraction of linguistically-oriented semantics, which basically maps various syntactic realizations to the shared structures. In order to develop a TM system, we need to solve the two difficult tasks.

Second, TM could benefit from linguistically-oriented frames by using them as an intermediating layer between text and domain-oriented information. The domain-oriented semantic structures, which is a target of TM, are inevitably dependent on the domain. On the other hand, the extraction of linguistically-oriented semantics from text is less dependent. Therefore, using the linguistically-oriented structure could be favorable to domain portability of a TM system.

Our aim was explicitly linking linguistically-oriented and domain-oriented semantics of the bio-molecular articles, and the preliminary result show the possibility of the extraction of linguistically-oriented semantics contributing to TM. Further investigation of the relationship would be a important step forward for TM in the bio-molecular domain.

Our investigation was preliminary. For example, conversions from FN structures to GENIA event structures, depicted in figures 4 and 5, were based on manual investigation. Further, they were attested by limited samples in the corpus. For our results to contribute to a TM system, evaluation of the conversions and automatic extraction of such conversions must be considered.

7 Conclusion

This paper presents a relationship of domain-oriented and linguistically-oriented frames of semantics, obtained by an investigation of the GENIA event corpus. In the investigation, we annotated sample sentences from the GENIA event corpus with linguistically-oriented semantic structures as those of FrameNet, and compared them with domain-oriented semantic annotations that the corpus originally possesses. The resulting relations between the domain-oriented and linguistically-oriented frames suggest that mentions of a biological phenomenon could be realized in a number of linguistically-oriented frames, and that the linguistically-oriented frames represent possible perspectives in mentioning the phenomenon. The resulting relations would illustrate a challenge in developing a Text Mining system, and would indicate importance of linguistically-oriented frames as an intermediating layer between text and domain-oriented information. Our future plan includes evaluation of our conversions from a linguistically-oriented to a domain-oriented structure, and automatic extraction of such conversions.

References

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics*, pages 86–90, Morristown, NJ, USA. Association for Computational Linguistics.

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, Prague, Czech Republic, June. Association for Computational Linguistics.

K. Bretonnel Cohen, Martha Palmer, and Lawrence Hunter. 2008. Nominalization and alternations in biomedical language. *PLoS ONE*, 3(9):e3158, 09.

Andrew Dolbey, Michael Ellsworth, and Jan Scheffczyk. 2006. Bioframenet: A domain-specific framenet extension with links to biomedical ontologies. In *Proceedings of the Second International Workshop on Formal Biomedical Knowledge Representation: "Biomedical Ontology in Action" (KR-MED 2006)*, volume 222 of *CEUR Workshop Proceedings*. CEUR-WS.org, Nov.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Sanda M. Harabagiu, Cosmin Adrian Bejan, and Paul Morarescu. 2005. Shallow semantics for relation extraction. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 1061–1066.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1):10.

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, Lyle Ungar, Scott Winters, and Pete White. 2004. Integrated annotation for biomedical information extraction. In Lynette Hirschman and James Pustejovsky, editors, *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, pages 61–68, Boston, Massachusetts, USA, May 6. Association for Computational Linguistics.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The nombank project: An interim report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjorne, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1):50.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177, Manchester, England, August. Coling 2008 Organizing Committee.

Tuangthong Wattarujeekrit, Parantu Shah, and Nigel Collier. 2004. Pasbio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5(1):155.