# A System for Building FrameNet-like Corpus for the Biomedical Domain

**He Tan**
School of Engineering
Jönköping University, Sweden
he.tan@jth.hj.se

## Abstract

Semantic Role Labeling (SRL) plays an important role in different text mining tasks. The development of SRL systems for the biomedical area is frustrated by the lack of large-scale domain specific corpora that are annotated with semantic roles. In our previous work, we proposed a method for building FramenNet-like corpus for the area using domain knowledge provided by ontologies. In this paper, we present a framework for supporting the method and the system which we developed based on the framework. In the system we have developed the algorithms for selecting appropriate concepts to be translated into semantic frames, for capturing the information that describes frames from ontology terms, and for collecting example sentence using ontological knowledge.

## 1 Introduction

Semantic Role Labeling (SRL) is a process that, for each predicate in a sentence, indicates what semantic relations hold among the predicate and its associated sentence constituents. The associated constituents are identified and their semantic role labels are assigned, as in: [$_{Transporter}$CBG] **delivers** [$_{Entity}$cortisol] [$_{Destination}$to target cells]. SRL could play an important role in text mining tasks such as information extraction, question answering and text summarization. With the advent of large resources like FrameNet (Fillmore et al., 2001) and PropBank (Palmer et al., 2005), SRL has become a well-defined task with a substantial body of work and comparative evaluation. Much of this work has focused on the arguments of verbs, and has been trained and evaluated on newswire text.

Recently, work has turned to bring SRL to the biomedical area (Wattarujeekrit et al., 2004; Tsai et al., 2006; Dolbey et al., 2006; Bethard et al., 2008). Biomedical text considerably differs from the PropBank and FrameNet data, both in the style of the written text and the predicates involved. Predicates in the data are typically verbs, biomedical text often prefers nominalizations, gerunds and relational nouns (Cohen et al., 2008; Kilicoglu et al., 2010). Predicates like *endocytosis* and *translocate*, though common in biomedical text, are absent from both the FrameNet and PropBank data (Wattarujeekrit et al., 2004; Bethard et al., 2008; Tan, 2010). Predicates like *block*, *generate* and *transform*, have been used in biomedical documents with different semantic senses and require different number of semantic roles compared to FrameNet (Tan, 2010) and PropBank data (Wattarujeekrit et al., 2004).

The projects, such as PASBio (Wattarujeekrit et al., 2004), BioProp (Tsai et al., 2006) and BioFrameNet (Dolbey et al., 2006), have made efforts on building resources for training SRL systems in the biomedical domain. PASBio annotated the semantic roles for 31 predicates (distributed 29 verbs) in style of PropBank. It used a model for a hypothetical signal transduction pathway of an idealized cell, to motivate verb choices. BioProp, also a PropBank-like corpus, annotated the semantic roles of 30 frequent biomedical verbs found in the GENIA corpus. BioFrameNet built a FrameNet-like corpus having 32 verbs and nouns annotated with the semantic roles. It considers a collection of GeneRIF (Gene References in Function) texts that are annotated by the protein transport classes in the Hunter Lab knowledge base. Up until recently, these corpora are relatively small.

One of obstacles to building FrameNet-like resources is to manually construct large, coherent and consistent frame sets for the domain. In (Tan et al., 2012) we argue that we can build large-scale FrameNet-like resources using domain knowledge from ontologies. A large number of ontologies



*Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014*, pages 46–53,
Gothenburg, Sweden, April 26-30 2014. 

have been developed in biomedical area, such as OBO ontologies (Smith et al., 2007). Many of them represent the knowledge of domain-specific events (any activities, processes and states). Although most of the ontologies are controlled vocabularies and do not explicitly describe the attributes of events, this information is implicitly contained in ontology terms. Together with the knowledge explicitly represented in the data models of ontologies the information can guide us in constructing large, coherent and consistent frame sets and also ease the task of collecting example sentences. In next section we describe the background knowledge and then present how the ontological knowledge can be used to build frame-semantic descriptions. Section 3 describes a general framework that supports this ontology-driven construction of frame-semantic descriptions and the current system we have developed based on the framework. Related work is given in section 4. Then we conclude the paper with a conclusion and the discussion of future work.

## 2 Ontology and Frame Semantics

Ontology is a formal representation of knowledge of a domain of interest. An ontology includes concepts that represent classes of entities within a domain, and defines different types of relations among concepts, as well as the rules for combining these concepts and relations. Most currently widely used ontologies in the biomedical domain are controlled vocabularies. The data models essentially contain lists of concepts, and organize them in an *is-a* and *part-of* hierarchy.

In practice, a concept contains one or more terms that are chosen for naming the concept. A preferred term is assigned as the name of the concept, and others could become synonyms. Terms are carefully chosen to clearly and precisely capture the intended meaning of the entities the concept refer to. The terms are noun or noun phrases. As showed in the results of the survey of naming conventions in OBO ontologies (Schober et al., 2009), multi-word terms are constructed in a consistent manner. They are created by re-using strings that appear in the terms already defined in this or in other ontologies. Although attributes of the entities belonging to concepts are not explicitly described in the data model, they remain implicit in the terms (Stevens et al., 2000). The constituents of the terms might contain the information.

Table 1: Protein Transport Concepts

| |
|---|
| GO:0009306 protein **secretion** |
| GO:0017038 protein **import** |
| GO:0071693 protein **transport** *within* extracellular region |
| GO:0072322 protein **transport** *across* periplasmic space |
| GO:0072323 chaperone-mediated protein **transport** *across* periplasmic space |
| GO:0042000 **translocation** *of* peptides or proteins *into* host |
| GO:0051844 **translocation** *of* peptides or proteins *into* symbiont |
| GO:0051808 **translocation** *of* peptides or proteins *into* other organism *involved in* symbiotic interaction |

The Gene Ontology (GO) (The Gene Ontology Consortium, 2000) is the most widely used controlled vocabulary in the area. It provides the terms for declaring molecular functions, biological processes and cellular components of gene and gene products. Table 1 lists the names of 8 subclasses of `GO:0015031 protein transport` in the *is-a* hierarchy. The head of a phrase determines the semantic category of object or situation which the phrase refer to. Therefore, the head words of the terms, `translocation`, `import`, `secretion` and `transport`, refer to a "protein transport" category, since the concepts represent different kinds of "protein transport". Other constituents of the terms express the attributes or properties of the event. For example, `translocation of peptides or proteins into other organism involved in symbiotic interaction (GO:0051808)`, express the entity (`peptides or proteins`), the destination (`into other organism`) and the condition (`involved in symbiotic interaction`) of a protein transport event. These information are not represented in the model of the ontology.

Frame Semantics (Fillmore, 1985) is the study of how the words evoke or activate frame knowledge, and how the frames thus activated can be used to understand the text that contains the words. Frame semantics assumes that in order to understand the meanings of the words in a language, we must first have knowledge of the background and motivation for their existence in the language and for their use in discourse. The knowledge is defined in the conceptual structures (frames). In the FrameNet, the lexicographic application of the theory, a semantic frame describes an event, a situation or an object, together with the frame ele-

ments (FE) that represent the aspects and components of the frame. Lexical units (LU) that belong to the frame, are the words that evoke the frame. Each frame is associated with example sentences within which LUs and FEs are marked. The FrameNet builds frames by collecting and analysing the attestations of words with semantic overlap from the British National Corpus (BNC).

We propose that the domain knowledge contained in ontologies can instruct us in building a FrameNet-like corpus, without having an existing large scale domain corpus like BNC. The construction starts with creating large coherent and consistent frame sets and then collecting associated example sentences. The information implicitly contained in ontology terms together with the knowledge represented in the models of ontologies provide the background knowledge that is required to building the frame-semantic descriptions. After the frames are created, associated example sentences can be collected using knowledge based search engines for biomedical text, and then be annotated.

For example, a frame Protein Transport can be characterized based on the concept `GO:0015031 protein transport`. In the frame, by studying the terms of the subclasses and descendants of the concept (such as those in table 1), the aspects and components of the frame (such as entity, destination and condition), and the domain-specific words evoking the frame (like translocation, import, secretion and transport) are captured. Furthermore, we can identify a *inheritance* relation between this frame and the frame `Transport` built based on the concept `GO:0006810 transport`, since there is the *is-a* relation between `GO:0006810 transport` and `GO:0015031 protein transport` in the GO. Now a complete frame-semantic description for Protein Transport, including FEs, LUs, and relations to other frames, is obtained after all the related concepts and relations are studied.

## 3 The System

In this section we present a framework that supports this ontology-driven construction of FrameNet-like corpus and describe the current system we have developed based on the framework.

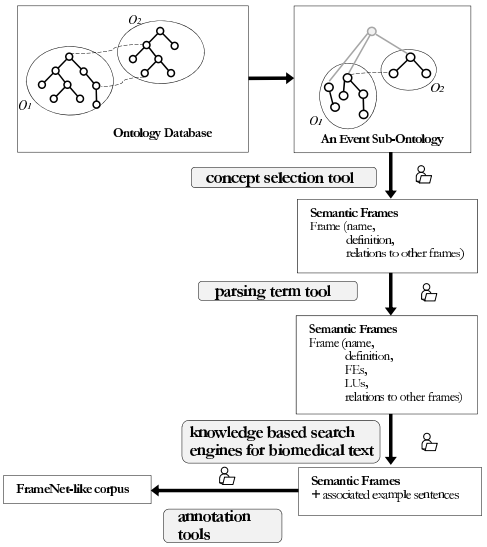


Figure 1: A Framework of Ontology-driven Building Corpus

### 3.1 Framework

In Fig 1 we propose the framework for supporting the ontology-driven construction of domain corpus with frame-semantics annotations. Before starting the building process, a sub-ontology of biomedical events is extracted from an ontology or an ontology database in which relations between ontology terms are identified. Firstly, concepts representing biomedical events, are gathered. A concept represents a biomedical event if it is a concept that is classified to a type of event in top-domain ontology (like the semantic type `T038 Biologic Function` in UMLS semantic network (Bodenreider, 2004)), or is a subclass or descendant of a concept that has been identified as a representation of biomedical event, or can be defined as a concept describing event based on its definition. After the concepts are identified, an event sub-ontology, including the concepts and the relations between them in original ontologies, is extracted. A root is assigned to the sub-ontology if the concepts are from more than one sub-trees in original ontologies.

The concept selection tool suggests the appropriate concepts that will be translated into frames. The algorithm may consider the characteristics that indicate the generalness of a concept as the selection criteria, such as the location of the concept in the hierarchy and the number of subclasses and descendants the concept has. Further, the concept could be manually identified by domain experts. After a concept is selected, the frame de-

scribing the event represented by the concept, is created. Relations between frames are decided according to the relations between the corresponding concepts. The name and definition of a frame is edited by domain experts based on the definition of the concept.

The frame description is accomplished by studying the sub-tree under the concept. After collecting the terms in the sub-tree, the parsing term tool analyses the compositional structure of the term, which elucidate the meaning of a term. The tool may derive the compositional structure of a term based a syntax parse of the term. LUs and FEs then are suggested based on the compositional structures. A final frame-semantic description is decided with interactions to domain experts.

The associated example sentences of a frame could be collected using semantic search engines for biomedical text, like GoPubMed (Doms and Schroeder, 2005). Such search engines annotate documents with concepts of domain ontologies by mapping phrases in text of documents to concepts. Based on this underlying domain knowledge search engines are able to maximize precision and recall of the documents that the user is interested to collect specific information. Therefore, example sentences can be collected from the documents annotated by the concepts in the sub-tree used to characterize the associated frame. In the end annotating example sentences with LUs and FEs of the associated frame is completed by domain experts under the assistance of annotation tools.

### 3.2 The System

We have developed a system based on the framework for building FrameNet-like corpus using domain ontologies.

#### An Event Sub-Ontology

In the current system we experimented with the GO biological process ontology (data-version: 2012-10-27). In UMLS semantic network the root node of the ontology `biological process (GO:0008150)` is classified into the semantic type `T038 Biologic Function` . The ontology contains 24,181 concepts and 65,988 terms. The terms include the names of the concepts and their *exact* synonyms. Other synonyms (*broad*, *narrow* and *related* synonyms) are not included, since only terms intending to precisely capture the meaning of a concept are considered. For example, `fat body metabolism`, a *broad* synonym of `GO:0015032 storage protein import into fat body`, describes a much broader activity than that belongs to the concept.

#### Method for Concept Selection

Different types of frames are used to describe different situations. Frames can range from highly abstract to very specific. To assist the user in selecting appropriate concepts to be translated into frames, the system provides the structure information of the ontology, and the definitions of the concepts and their locations in the ontology.

The event ontology $O$ can be represented as a directed graph $G$. Graph elements are considered to calculate the structure information of $O$ and the location of the concepts in $G$ including,

- the $root$, the node having no outgoing *is-a* arcs. The graph $G$ has one root.
- a $leaf$ node , a node having no ingoing *is-a* arcs in the graph.
- $sibling$ nodes, nodes connected to the same node through *is-a* arcs.
- $descendant$ nodes of a node $n_i$, nodes in the sub-tree rooted at $n_i$.
- a path $p_{ij}$, any sequence of directed edges from the node $n_i$ to the node $n_j$.
- a generation $g_i$, the set of all sibling nodes connected to the node $n_i$.
- $depth$, the cardinality of a path
- $breadth$, the cardinality of a generation.

As the structure information of $O$ we calculate the number of nodes in $G$, the average and maximal shortest paths from the root to leaves, the average and maximal breadth of the generations having different distances from the root. To show the location of a concept in G, we calculate the shortest path from the concept to the root, and the number of its descendants and siblings.

The user selects appropriate concepts based on the above information, and may also using their own domain knowledge. For example, a frame could be constructed based on the concept `GO:0006810 transport`. The structural information as showed in table 2 suggests that the concept is richly described in the ontology and it covers a large set of related events. Further, the user (a domain expert) himself/herself could be aware that transport events have been studied in the area over

| | #node | depth of shortest path to root (SPR) | #sibling | avg. depth of SPR from leaves | max. depth of SPR from leaves | avg. breadth | max. breadth. | #leaf |
|---|---|---|---|---|---|---|---|---|
| biological_process | 24181 | - | - | 6.5 | 14 | 3.7 | 413 | 12871 |
| transport | 1210 | 2 | 5 | 5.9 | 14 | 3.5 | 41 | 754 |
| protein transport | 182 | 3 | 41 | 5.7 | 9 | 4.2 | 40 | 132 |

Table 2: The structural information of GO biological process ontology (data-version: 2012-10-27) and the sub-trees under the concept `GO:0006810 transport` and `GO:0015031 protein transport`.

the last 30 years. Most cellular processes are accompanied by transport events. For understanding biomedical texts, transport events are among the most important things to know about.

**Method for Parsing Terms**

After a concept is selected, the terms in the subtree rooted at the concept are collected to be analysed for building frame description. In the current system the analysis is separated into three steps.

Terms are noun phrases (NP). The first step is to tokenize phrase string into an ordered sequence of an atomic (non-decomposable) token. The phrase string is split on white-space characters and non-alphanumeric characters. White-space character are discarded, but non-alphanumeric characters are preserved and treated as special word tokens. For example, "`alpha-glucoside transport (GO:0000017)`" is tokenized into {alpha, -, glucoside, transport}

The second step is to identify the head word of NP. We assume that the head of a phrase is composed of only one token. A naive Bayes classifier classifies a token as the head of a phrase, if the highest value for the posterior probability of being the head word given the token is obtained among all the tokens in the phrase. The posterior probability of being the head word $w$ given token $t$ is estimated using Bayes rule (Mitchell, 1997):

$$P(w|t) = \frac{P(w)P(t|w)}{P(t)}$$

As $P(t)$ is independent of $w$ being the head, it can be ignored. This gives: $P(w|t) = P(w)P(t|w)$.

A token is either the head word or not the head work of a phrase, so $P(w)$ is a constant. $P(t|w)$ is estimated by the feature probabilities of token $t$. Assuming that the features $x_i$ are all conditionally independent of one another, we have

$$P(t|w) = \prod_{i=1}^{n} P(x_i = a_{ik}|w)$$

$P(x_i = a_{ik}|w)$ is estimated using the maximum likelihood estimation method. Let $n(x_i = a_{ik}, t)$ be the number of occurrences of token $t$ where attribute $x_i$ is $a_{ik}$ and $t$ is a head word, and $n(w)$ be the number of occurrences of the token $t$ where $t$ is a head word. Then $P(x_i = a_{ik}|w)$ is estimated by

$$P(x_i = a_{ik}|w) = \frac{n(x_i = a_{ik}, w) + \lambda}{n(w) + \lambda|V|}$$

where $\lambda$ is the earlier defined Laplace smoothing parameter, and $|V|$ is the number of distinct values of the attribute $x_i$.

Attributes of a token $t$ in a phrase $p$ include,

- token string,
- the part-of-speech (POS) of $t$ in $p$, (the POS of $t$ in $p$ is assigned using MedPost POS Tagger (Smith et al., 2004)),
- the POS of the tokens before and after $t$ in $p$,
- the length of $p$ (the number of tokens in $p$),
- the position of $t$ in $p$.

We have evaluated the method on identifying the heads of terms in GO biological process ontology. The length of terms in the ontology ranges from 1 to 39. For each length, 10% of terms are randomly selected as training data if it is applicable. The result of 10-fold cross validation showed that 93.9% of the heads are correctly identified on average.

A term, a NP, has a noun as its head. The system collects other forms (such as verb, objective, etc.) having the same meaning as the head by looking up the SPECIALIST Lexicon (Bodenreider, 2004), a general English lexicon including many biomedical term. Words in different forms are all suggested as predicates for frame.

The last step is to capture the information hidden in modifiers in phrases. Modifiers describe the head word of a phrase and makes its meaning more specific. They modify phrases by adding information about "where", "when", or "how" something

Table 3: Major Modifier Types in Ontology Terms

| Pre-modifiers | head | Post-modifiers |
|---|---|---|
| attributive adjective | noun | prepositional phrase |
| ed-participial adjective | | ed-clause |
| ing-participial adjective | | ing-clause |
| noun | | to-clause |
| | | appositive noun phrase |

is done. The information gives the suggestions on what FEs to be defined for a frame. In a NP, the head word is preceded by a determiner or one or more pre-modifiers, and may also be followed by one or more post-modifiers. The major structural types of pre-modifiers and post-modifiers are given in table 3. We observed that determiners and relative clauses rarely appear in ontology terms.

The number of FEs is limited in a frame. The information about the major attributes of event appears frequently in the terms. For example, in the sub-tree under `GO:0006810 transport`, 92.6% terms contain the entity undergoing the "transport" event, and 19.3% terms describe the destination (see Table 4). Therefore, although there maybe a large number of terms in a sub-tree, a very small number of the terms can be used to capture the major attributes of the event.

To facilitate the user in identifying the FEs, the system collects a smallest set of terms covering all the attributes of the event that have been described in the sub-tree. The attributes of the event reside in different modifier types appearing in the terms. Further, prepositional phrase modifiers starting with different prepositions may describe different properties. The algorithm for collecting the term set is given as follows,

$T$ ={the set of terms in the sub-tree} ;
$M$={the set of modifier types $m$} ;
$P, L = \emptyset$ ;
**repeat**
  $l$ = the longest $t \in T$ ;
  **foreach** $m$ *in* $l$ **do**
    **if (** $m$ *is a prepositional phrase* **and** $m$ *starts with a preposition* $p \notin P$**) or** $m \notin P$ **then**
      add $l$ to $L$ ;
      **foreach** $m$,$p$ *in* $l$ **do**
        **if** $m, p \notin P$ **then**
          add $m, p$ to $P$ ;
      **end**
      break ;
  **end**
  remove $l$ from $T$ ;
**until** $T = \emptyset$ **or** $length(l) = 1$;
**return** $L$

**Method for Collecting Example Sentences**

The example sentences are retrieved from the PubMed/MEDLINE database by using the GoPubMed. The sentences to be annotated, are always the most relevant and from the latest publications. For a LU, we acquired sentences by using the GO terms with the head from which the LU is derived. The query starts from using the most general GO terms. In cases when only specific GO terms are available and the number of query results is too small, the query term is generalized by removing modifiers from terms. For example, the lexical units, `release.n` and `release.v`, are derived and only derived from `renin release into blood stream` (a synonym of `GO:0002001 renin secretion into blood stream`). No query result returns for the GO term (AND operator is used to combine the tokens in the term in the query). The general term "protein release" is used as the query term instead.

**Annotation Tool**

The current system contains a tool that supports manual annotation following the FrameNet's guidelines described in (Ruppenhofer et al., 2005).

**File Format**

The corpus is stored in XML files using the same format as the FrameNet. The correspondences between frames and ontology concepts are stored in a RDF file. Such relations could benefit integrations of different lexical resources and/or knowledge bases in the future. A correspondence is encoded as follows:

```
<correspondence id="1">
 <concept rdf:about=
  "http://www.geneontology.org/go#GO:0006810"/>
 <frame rdf:about=
  "http://hj.se/ontobiofn/frames#0000001"/>
 <comment/>
</correspondence>
```

It provides the features: concept (the URI of some concept of an ontology); frame (the URI of the frame translated from the concept); comment (the comment on this correspondence given by the user); and an id assigned to this correspondence.

### 3.3 Evaluation of the System

We have successfully built a FrameNet-like corpus using the method of ontology-driven construction (Tan et al., 2012). The construction is done manually by 2 master students with biology background. The corpus covers transport events in the domain. The GO is used as the source ontology for domain knowledge. The corpus contains 2 frames.

| | **TE** | **TO** | **TDS** | **TC** | TL | TP | TT | TDR | TA | TPL |
|---|---|---|---|---|---|---|---|---|---|---|
| Protein Transport (581 terms) | 99.5% (578) | 8.6% (50) | 37.4% (159) | 16.4% (95) | 7.1% (41) | 4.6% (27) | 1.0% (6) | 0.3% (2) | 0.2% (1) | 0% (0) |
| Transport (2235 terms) | 92.6% (2070) | 12.2% (272) | 19.3% (432) | 9.9% (221) | 5.7% (127) | 7.3% (164) | 1.9% (43) | 1.5% (34) | 1.8% (40) | 0.36% (8) |

Table 4: The percentage of the GO terms that indicate the FEs (the number of GO terms). FEs are Transport_Entity (TE), Transport_Origin (TO), Transport_Destination (TDS), Transport_Condition (TC), Transport_Location (TL), Transport_Path (TP), Transport_Transporter (TT),Transport_Direction (TDR), Transport_Attribute (TA), Transport_Place (TPL).

Table 5: Time for Building the Corpus

| | **using system** | **manual** |
|---|---|---|
| construct frames | 2 days | 2 weeks |
| gather and annotate example sentences | 2.5 weeks | 3 weeks |

The Transport frame follows the definition of the GO concept, `GO:0006810 transport` (Tan et al., 2012). It has a sub-frame Protein Transport, which characterizes transport events of proteins (Tan et al., 2011). It follows the definition of `GO:0015031 protein transport`. To accomplish the description of the two frames, 2235 terms and 581 terms, respectively, were collected and analysed from the GO. Based on the background knowledge implicitly described in the terms, 10 FEs are identified for the frame Transport (inherited by the frame Protein Transport), and 129 LUs are collected. Maximally for each LU 10 annotated sentences are gathered. Totally, 955 example sentences were retrieved from PubMed and annotated.

We evaluate the effectiveness and efficiency of the system. 2 different master students are asked to build a FrameNet-like corpus covering transport and protein transport events using the method. The GO is also provided as the source ontology. The 2 students have biology background and have the knowledge of the FrameNet and ontology. Both students correctly complete the task using the system in the evaluation. They build the 2 frames Transport and Protein Transport, and construct the same frame descriptions using the domain knowledge from the GO. They are also required to maximally collect and annotate 10 sentences for each LU. The set of the example sentences are not exactly the same set of sentences chosen in the previous corpus. Table 5 shows the time they use on average and the time spent in the manual construction.

## 4 Related Work

Interfacing ontologies and lexical resources has been initiated in several work (Guarino, 1998; Gangemi et al., 2003; Niles and Pease, 2003). The work in (Gangemi et al., 2003; Niles and Pease, 2003) has attempted to reorganize WordNet's top-level synset taxonomy using ontology concepts. More recently, the FrameNet project links FrameNet's semantic types to ontology concepts, to constrain the filler types of frame elements for specific domains (Scheffczyk et al., 2006). It is the first step of their work aiming at improving FrameNet capability for deductive reasoning with natural language. The authors suggest that the alignment between lexicons and ontologies could restructure the lexicon on the basis of ontological-driven principles, and enables ontologies to be used automatically by natural language processing (NLP) applications.

## 5 Conclusion

In this paper we present our method for building FrameNet-like corpus for biomedical area starting with use of ontological domain knowledge. Ontological knowledge can lead to well-defined semantics exposed on the corpus, which can be very valuable in NLP and text mining applications. We have developed a framework of supporting the method and implemented a system based on the framework. In the current system we developed the algorithms for selecting appropriate concepts to be translated into semantic frames, for capturing the information that describes aspects and components of frames from ontology terms, and for collecting example sentence using ontology concepts.

In the future we will continue to extend the corpus using ontological knowledge. The event ontology to be used as domain knowledge will include terms from different ontologies. We will evaluate our system when it deals with different ontologies and their terms. Another direction of the future work is to investigate how the ontological knowl-

edge bundled with the corpus are used by NLP and text mining applications.

## References


Steven Bethard, Zhiyong Lu, James H Martin, and Lawrence Hunter. 2008. Semantic role labeling for protein transport predicates. *BMC Bioinformatics*, 9:277.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, pages D267–D270.

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

Andrew Dolbey, Michael Ellsworth, and Jan Scheffczyk. 2006. Bioframenet: A domain-specific framenet extension with links to biomedical ontologies. In *Proceedings of KR-MED*, pages 87–94.

Adress Doms and Michael Schroeder. 2005. Gopubmed: exploring pubmed with the gene ontology. *Nucleic Acids Research*, 33:W783–786.

Charles J. Fillmore, Charles Wooters, and Collin F. Baker. 2001. Building a large lexical databank which provides deep semantics. In *Proceedings of the PACLIC*.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2003. Sweetening wordnet with dolce. *AI Magazine*, 3(24):13–24.

Nicola Guarino. 1998. Some ontological principles for designing upper level lexical resources. In *Proceedings of First International Conference on Language Resources and Evaluation*, pages 527–534.

Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, Sean Marimpietri, and Thomas C. Rindflesch. 2010. Arguments of nominals in semantic interpretation of biomedical text. In *Proceedings of the 2010 Workshop on BioNLP*.

Tom Mitchell. 1997. *Machine Learning*. McGraw-Hill.

Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31:71–105.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2005. FrameNet II: Extended theory and practice. Technical report, ICSI.

Jan Scheffczyk, Adam Pease, and Michael Ellsworth. 2006. Linking framenet to the sumo ontology. In *International Conference on Formal Ontology in Information Systems*.

Daniel Schober, Barry Smith, Suzanna Lewis, Waclaw Kusnierczyk, Jane Lomax, Chris Mungall, Chris Taylor, Philippe Rocca-Serra, and Susanna-Assunta Sansone. 2009. Survey-based naming conventions for use in obo foundry ontology development. *BMC Bioinformatics*, 10(1):125.

L. Smith, T. Rindflesch, and W. J. Wilbur. 2004. MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*, 20(14):2320–2321.

Barry Smith, Michael Ashburner, and et al. 2007. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255.

Robert Stevens, Carole A. Goble, and Sean Bechhofer. 2000. Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, 1:398–414.

He Tan, Rajaram Kaliyaperumal, and Nirupama Benis. 2011. Building frame-based corpus on the basis of ontological domain knowledge. In *Proceedings of the 2011 Workshop on BioNLP*, pages 74–82.

He Tan, Rajaram Kaliyaperumal, and Nirupama Benis. 2012. Ontology-driven construction of corpus with frame semantics annotations. In *CICLing 2012, Part I, LNCS 7181*, pages 54–65.

He Tan. 2010. A study on the relation between linguistics-oriented and domain-specific semantics. In *Proceedings of the 3rd International Workshop on SWAT4LS*.

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Richard Tzong-Han Tsai, Wen-Chi Chou, Ying-Shan Su, Yu-Chun Lin, Cheng-Lung Sung, Hong-Jie Dai, Irene Tzu-Hsuan Yeh, Wei Ku, Ting-Yi Sung, and Wen-Lian Hsu. 2006. Biosmile: adapting semantic role labeling for biomedical verbs: an exponential model coupled with automatically generated template features. In *Proceedings of the 2005 Workshop on BioNLP*.

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