Semantic Role Labeling for Protein Transport Predicates

Steven Bethard, Zhiyong Lu,
James H. Martin
and Lawrence Hunter
University of Colorado

We present a model for identifying the semantic roles of protein transport predicates in GeneRIF data. GeneRIFs are sentences which briefly describe the function of a particular gene or protein. The protein transport predicates in these GeneRIFs are primarily nouns and thus offer unique challenges for semantic role labeling. Since most state-of-the-art syntactic parsers give little or no structure to noun phrases, we approach this problem with a word-chunking paradigm and train support vector machine classifiers to classify words as being at the beginning, inside or outside of a protein transport role. We train these models with the features of previous word-chunking models, features adapted from other types of models, and features derived from analysis of our data. Our models are able to label protein transport semantic roles with a precision of 90.4 and a recall of 78.7 using gold-standard protein boundaries, and a precision of 88.4 and a recall of 71.7 using automatic ones.

1. Introduction

With the advent of resources like FrameNet (Filmore, Wooters, and Baker 2001) and PropBank (Kingsbury and Palmer 2002; Palmer, Gildea, and Kingsbury 2005), automatic semantic role labeling has had a flurry of activity in recent years. Much of this work has focused on the arguments of verbs, and because PropBank is annotated on top of Wall Street Journal text, much of the work has been trained and evaluated on newswire text (Surdeanu et al. 2003; Xue and Palmer 2004; Pradhan et al. 2005; Toutanova, Haghighi, and Manning 2005; Punyakanok et al. 2005).

As a variety of research groups have reported success on these corpora, recent work has turned to transferring these results to different kinds of predicates and different genres of text. In this article, we show that automatic semantic role labeling can be transferred to the biomedical domain. Biomedical text differs widely from the text of both FrameNet and PropBank, both in the style of the written text and the predicates involved. Biomedical text often prefers light verbs and nominal predicates, so text like example 1, where all but one predicate is a nominal form, is quite common.

(1) [PREDICATE Truncation] of up to 44 C-terminal amino acids from the putatively cytoplasmic C-terminal hydrophilic domain [SUPPORT-VERB left] transport function [PREDICATE unimpaired], but [PREDICATE deletion] of the adjacent STAS (sulfate transporter anti-sigma factor antagonist) domain [PREDICATE abolished] function.

The predicates used in biomedical text are also quite unlike those of other corpora. Predicates like *endocytosis*, *exocytosis*, *internalize*, *traffic* and *translocate*, though common in texts describing protein transport, are completely absent from both the FrameNet and PropBank data.

Other researchers have explored the difficulties of porting semantic role labeling technologies to new domains and have encountered the same two basic problems: differences in text style and differences in predicates. The CoNLL 2005 shared task investigated semantic role labeling systems that were trained on the Wall Street Journal and tested on the Brown corpus (Carreras and Màrquez 2005). They found that "all systems experienced a severe drop in performance (about 10 F1 points)" when compared to their results on Wall Street Journal Data, and attributed this drop to the poorer performance of sub-components like part-of-speech taggers and syntactic parsers. In an investigation of how differences in predicates affect semantic role labeling, Pradhan et. al. (2004) investigated the arguments of nominalized predicates, using a selection of FrameNet predicates and manually annotated nominalizations from the Penn Chinese TreeBank. They found automatically identifying such roles to be much more difficult than annotating verbal roles, reporting F-scores in the low 50s and 60s. Research efforts like these suggest that porting semantic role labeling to biomedical text will offer some interesting challenges.

The remainder of this article discusses our approach to this. First we describe the biomedical resources employed. Then we discuss some of the differences between this data and the PropBank data, and explain the approach we take to address these differences. Next, we translate the task of identifying biomedical predicate arguments into a classification task, and describe the features used to train our classifiers. Finally, we evaluate the performance of our classifiers, and discuss the implications of these results.

2. Data

The data we consider in this article is a set of predicates and their semantic roles annotated on top of biomedical text. The National Library of Medicine (NLM) began a Gene Indexing initiative on April 1, 2002, the goal of which was to link any article about the basic biology of a gene or protein to the corresponding entry in Entrez Gene[1], the National Center for Biotechnology Information's gene database. The result was an entry within Entrez Gene called a Gene Reference Into Function (GeneRIF)[2], which acts as an important textual source of the functional annotation of genes (Wheeler et al. 2006; Rubinstein and Simon 2005). Our predicates and roles have been annotated on top of a subset of these GeneRIFs, for example:

(2) IRS-3 expression blocked glucose/IGF-1 induced [PATIENT IRS-2] [PREDICATE translocation] from the [ORIGIN cytosol] to the [DESTINATION plasma membrane].

GeneRIFs have been used in a variety of natural language processing projects on biomedical text, including projects to automate alerts for new findings (Mitchell et al. 2003) and to extract summaries of PubMed/MEDLINE records (Hersh and Bhupatiraju

1 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
2 http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html

2003; Bhalotia et al. 2003; Jelier et al. 2003; Lu, Cohen, and Hunter 2006). Most relevant to the research at hand is (Lu et al. 2006b), which describes OpenDMAP (Open-access Direct Memory Access Parser), a system which combines pattern matching with a domain specific ontology to build an application that can extract AGENT, PATIENT, ORIGIN and DESTINATION semantic roles for the predicate *translocation* with 75.4% precision and 71.1% recall.

We use data similar to that of (Lu et al. 2006b), and we focus in particular on predicates that describe protein transport. Protein transport is the biological process of moving proteins from one cellular component to another by various sorting mechanisms. For example, in order for extracellular signals to be transduced to the nucleus to activate specific genes, an essential step is translocating transcription factors into the nucleus. Understanding the mechanisms of protein transport has been a central theme in cell biology and has been studied for decades (Dalbey and Von Heijne 2001). However, while natural language processing technologies have generally shown success in facilitating biomedical research (Horn, Lau, and Cohen 2004; Muller, Kenny, and Sternberg 2004; Alfarano et al. 2005; Shah et al. 2005), there is currently very little work that has focused on applying NLP techniques to the protein transport domain.

For this project, a set of 837 GeneRIFs was first collected from two sources: GeneRIFs containing genes known to be involved in protein transport (e.g. *Src*, a tyrosine kinase playing critical roles in signaling) and GeneRIFs containing predicates known to express transport (e.g. *translocation* or *export*). The predicates in each of these GeneRIFs were annotated with the roles AGENT, PATIENT, ORIGIN and DESTINATION by domain experts following the annotation guidelines of (Lu et al. 2006a). GeneRIFs that did not express protein transport (e.g. because they expressed some other type of transport) were discarded.

Initially, this produced 1009 predicate annotations with 1803 labeled roles. However we adjusted the annotation scheme of (Lu et al. 2006a) in one small way. In (Lu et al. 2006a), if a predicate has a role containing a conjunction, e.g. *HopO1-1, HopS1, and HopS2*, the predicate would have been annotated three times, one for each conjoined element. Instead, we treat the phrase containing the conjoined elements as a single multi-word role, and only annotate the predicate once. After this change, the data contained 911 predicate annotations with 1543 labeled roles. We selected at random 200 GeneRIFs for the test set, reserving the remaining 637 for training. This gave us a test set with 215 annotated predicates and 371 labeled roles and a training set with 696 annotated predicates and 1172 labeled roles.

3. Protein Transport Role Labeling

In considering how to approach the task of identifying our protein transport argument roles, we considered the previous work in semantic role labeling, much of which has been based on PropBank, a million-word corpus annotating Wall Street Journal verbs with predicate argument structure (Kingsbury and Palmer 2002; Palmer, Gildea, and Kingsbury 2005). The most successful of the models trained on this corpus treated finding argument roles as a constituent identification task — sentences were first syntactically parsed, and then, given a verbal predicate, each syntactic constituent was classified as either being a role or not being a role of that predicate (Surdeanu et al. 2003; Xue and Palmer 2004; Pradhan et al. 2005; Toutanova, Haghighi, and Manning 2005; Punyakanok et al. 2005). This approach took advantage of the fact that PropBank is annotated on top of syntactic trees, and that most verbal arguments could be associated with a single syntactic constituent.

Though this approach has proven to be successful for identifying the semantic roles of some verbal predicates in biomedical data (Tsai et al. 2006), it is problematic for predicates in protein transport data because our protein transport predicates are predominantly (about 85%) nouns. Nouns introduce a number of difficulties that do not appear with verbs. First, many semantic relations for nouns are expressed by noun compounding, where many of the syntactic cues that were useful for verbs are unavailable. For example, there is no subject/object distinction for nouns, so that a two-noun compound can be formed just as easily using the verbal equivalent's subject, object or prepositional object. So for example, given the phrase *The transporter translocates GLUT-4 to the nucleus*, paraphrases using the nominalization *translocation* could look like any of examples 3, 4 and 5.

(3) [AGENT transporter] translocation

(4) [PATIENT GLUT-4] translocation

(5) [DESTINATION nuclear] translocation

Nouns are also more difficult due to the way most automatic syntactic parsers handle noun phrases. The Penn TreeBank (Marcus, Santorini, and Marcinkiewicz 1994), on which most modern parsers are trained, gives a very flat structure to noun phrases. So for both for long compound nouns, as in example 6, and conjoined nouns, as in example 7, all words would be concatenated into a single flat NP constituent.

(6) [NP [PATIENT fatty acid transport protein] [PREDICATE translocation]]

(7) [NP the actin cytoskeleton and [PATIENT ERK] [PREDICATE translocation]]

Thus, we cannot approach finding transport predicate arguments as a syntactic constituent classification problem as there is all too often no constituent corresponding to the predicate argument. In our data, 20% of roles match no constituent boundaries at all, and nearly 50% match only single-word constituents.

Previous work has approached these problems in a number of different ways. The first is by focusing on a smaller subtask that only attempts to identify the arguments of a nominalization in a noun compound. The latter is to use a word-chunking approach that classifies individual words instead of syntactic constituents.

3.1 Noun Compound Approaches

Lapata (2002) considered only compounds of a nominalization and a single modifier, and tried to identify whether the modifier was subject-like or object-like. For example, the phrases *child behavior* and *car lover* would be correctly parsed as in examples 8 and 9.

(8) [SUBJECT child] behavior

(9) [OBJECT car] lover

Lapata used statistical information about verbal subjects and objects to estimate model parameters for the nominalizations corresponding to those verbs. The model

Sales	B_ARG0
declined	O
10	B_ARG2
%	I_ARG2
to	O
$	B_ARG4
251.2	I_ARG4
million	I_ARG4
from	O
$	B_ARG3
278.7	I_ARG3
million	I_ARG3
.	O

Table 1
Role chunk labels for *Sales declined 10% to $251.2 million from $258.7 million.*

trained with this information was able to make the subject/object distinction correctly 86.1% of the time. Later, Girju et. al. followed on to this work by expanding the labels from SUBJECT and OBJECT to a set of 35 more semantically oriented labels, including AGENT, TEMPORAL, LOCATION and THEME. By applying support vector machine classifiers and a variety of lexicon-based features (Girju et al. 2004), Girju et. al. were able to achieve F-scores in the 60s and 70s on this harder task.

The success of both Lapata's and Girju et. al.'s approaches points out the advantage of sharing information between verbal predicates and nominal ones. Nonetheless, Lapata's and Girju et. al.'s work addresses only two-word noun compounds, and does not present a clear path for extension to multi-word noun compounds or to arguments occurring in other constituents. Thus we turn instead to a different approach: word-chunking.

3.2 Word Chunking Approaches

To be able to handle both multi-word compounds and arguments in other constituents, we adopted the approach of Hacioglu et. al., which considers argument identification as a word-chunking problem (Hacioglu and Ward 2003; Hacioglu 2004; Hacioglu et al. 2004). The word-chunking formulation converts a phrase identification problem into a word classification problem by selecting appropriate labels for each word in the phrase. For argument role identification, appropriate labels can be derived through a combination of a B(eginning), I(nside) or O(utside) prefix that indicates the location of the word within the role, and a role suffix that indicates the type of the role containing the word. So for example, given the sentence *Sales declined 10% to $251.2 million from $258.7 million*, its words would be labeled as in Table 1.

The success of such an approach was further demonstrated in the CoNLL 2004 Shared Task, which presented semantic role labeling of PropBank as a word-chunking task (Carreras and Màrquez 2004). The classifiers trained on this data used low-level features like part-of-speech tags and base-phrase chunks, and though performance was typically lower than that of constituent-based classifiers, these classifiers neither

to	O
induce	O
the	O
nuclear	B_DESTINATION
translocation	O
of	O
NF-kappaB	B_PATIENT
transcription	I_PATIENT
factor	I_PATIENT

Table 2
Role chunk labels for the phrase *to induce the nuclear translocation of NF-kappaB transcription factor*

required a syntactic parse to extract features nor chose their role boundaries based on one.

4. Classification Methods

Because the word-chunking approach of Hacioglu et. al. and the CoNLL 2004 Shared Task avoids the dependency on a syntactic parse, we adopt that approach in this work. So, for example, in trying to identify the roles of *translocation* in the phrase *to induce the nuclear translocation of NF-kappaB transcription factor*, we would attempt to label the words as in Table 2.

In order to be able to train classifiers that can perform such a labeling, we need to both select an appropriate machine-learning algorithm and determine an appropriate set of features. For the machine-learning algorithm, we select Support Vector Machines as they have been previously shown to perform well on similar tasks (Kudo and Matsumoto 2001; Hacioglu et al. 2004). We use the YamCha package (Kudo and Matsumoto 2001) which wraps the TinySVM[3] Support Vector Machine implementation with the appropriate logic for word-chunking problems.

For our features, we begin with the simple features used in the word-chunking model of (Hacioglu et al. 2004) [4]:

- The text of the word.
- The stem of the predicate.
- The part-of-speech of the word.
- The BIO tag for the phrase that includes the word, e.g. B_VP or I_NP.
- The brace tag indicating how many clauses start and end at the word, e.g. $(*)$ or $))$

However, this is a small feature space that misses some important characteristics of the task, so to augment our feature space, we turn to Hacioglu et. al.'s phrase-chunking

3 http://chasen.org/ taku/software/TinySVM/
4 We omit the feature giving the BIO tag for people, organizations and locations as such named entities generally do not occur in GeneRIF data

to	TO	O
induce	VB	O
the	DT	O
nuclear	JJ	B_DESTINATION
translocation	NN	O
of	IN	O
NF-kappaB	NN	B_PATIENT
transcription	NN	I_PATIENT
factor	NN	I_PATIENT

Table 3
Feature window for the classification of *NF-kappaB* in the phrase *to induce the nuclear translocation of NF-kappaB transcription factor*

model. The model itself is inappropriate for our task because, just as the constituent classification models, the phrase classification model tries to classify spans of words that are larger than most of our roles. The model's features, however, have a relatively straightforward translation to word-level features instead of phrase-level features and thus that can be modified for use with our model.

The full list of features we derived from (Hacioglu et al. 2004) is split into three sections for presentation purposes: features that describe the word to be classified, features that describe the predicate, and features that describe the path between the word and the predicate. All features in all sections rely on the same sub-components:

- Word stems are determined by a lookup table from the University of Pennsylvania of around 300,000 words.
- Part-of-speech tags are identified by the MXPOST part-of-speech tagger (Ratnaparkhi 1996).
- Syntactic phrases are determined by an in-house YamCha-based chunking systems trained on the CoNLL 2000[5] text chunking data
- Clause boundaries are determined by an in-house YamCha-based chunking system trained on the CoNLL 2001[6] clause identification data.

In addition to the features themselves, we use a windowing strategy to give some additional context to the word classifiers. For each word, we include not only the features of that word, but the features of some words before and after it. So for example, if our only features were the word itself and its part-of-speech, and we were considering a window of one word on each side of the one being classified, the feature window around *NF-kappaB* in the phrase *to induce the nuclear translocation of NF-kappaB transcription factor* would look like Table 3.

The following sections describe the individual word-level features in more detail.

5 http://www.cnts.ua.ac.be/conll2000/chunking/
6 http://www.cnts.ua.ac.be/conll2001/clauses/

Word-based features. These features characterize the word currently being classified, independent of the predicate whose arguments we are looking for.

- The text of the word.
- 2, 3, and 4 character suffixes of the word.
- The part-of-speech of the word.
- The BIO tag for the phrase that includes the word, e.g. B_VP, I_NP.
- The brace tag indicating how many clauses start and end at the word.

Predicate Features. These features characterize the predicate whose arguments we are looking for. For a given predicate in a given sentence, these features will be the same for all words in that sentence.

- The text of the predicate.
- The stem of the predicate.
- The part-of-speech of the predicate.
- The number of predicates in the sentence.
- The part-of-speech of the word before the predicate.
- The part-of-speech of the word after the predicate.
- The two phrase types preceding the predicate, e.g. NP, NP.
- The two phrase types following the predicate.

Path Features. These features characterize the path between the word being classified and the predicate whose arguments we are looking for.

- The location of the word relative to the predicate, either BEFORE, AFTER or PREDICATE.
- The distance between the predicate and the word in number of phrases.
- The distance between the predicate and the word in number of VPs.
- The phrasal path between the predicate and the word, e.g. NN>NP>PP>IN.
- The clause boundaries between the predicate and the word, e.g. ()).
- The clause boundaries between the sentence boundary and the word.

5. Adapting to Protein Transport Roles

Preliminary experiments[7] showed that our models were having difficulties with a few different areas of our data: the boundaries of protein names, conjoined predicates and arguments tied to a predicate through coreference.

7 These experiments were carried out as cross-validations on the training data so as to keep our test set clean.

We noticed early on that our models were having trouble determining when a phrase immediately preceding a predicate should be identified as a PATIENT. For example, our early models identified *GLUT4 requires* instead of *GLUT4* as the PATIENT in example 10, and couldn't find any PATIENT at all in example 11.

(10) These results suggest that [PATIENT GLUT4] requires [PREDICATE translocation]...

(11) ...involved in [PATIENT eNOS] [PREDICATE translocation]...

The system had learned a strategy that identified as the PATIENT everything from the last "proper noun" up to the predicate. In these two examples, the part-of-speech tagger identified only GLUT4 as a proper noun, and so not only did the system incorrectly include *requires* as part of the PATIENT in example 10, but it also failed to include the PATIENT *eNOS* in example 11. These errors indicated that our models were having trouble identifying the boundaries of protein names.

Our models were also having trouble with conjoined predicates, particularly when an argument was present for the first but elided for the second. So, for instance, in example 12, *protein* is the PATIENT of both *folding* and *translocation*, and in example 13, *Tir* is the PATIENT of both *secretion* and *translocation*. In both of these examples, our early models failed to identify *protein* and *Tir* as PATIENT roles of the *translocation* predicates.

(12) ...for ERdj5 in [PATIENT protein] folding and [PREDICATE translocation]...

(13) ...for efficient [PATIENT Tir] secretion and [PREDICATE translocation]...

Though our models were given a window of features around the word classified, this window was generally no more than two words before of after the word[8]. Thus words like *protein* and *Tir* above were too distant from the predicate to be considered as arguments, and so our models failed on them.

Finally, our models had trouble with the annotation style of (Lu et al. 2006a) in that it annotates some roles that are tied to the predicate only through a coreference chain. Example 14 shows such a role.

(14) a rapid activation of the [PATIENT acid sphingomyelinase] correlating with its microtubule- and microfilament-mediated [PREDICATE translocation]

In this example, the predicate *translocation* is contained within the prepositional phrase *with its... translocation*. PropBank-style annotation would thus likely annotate *its* as the PATIENT of *translocation*. However, the annotation style of (Lu et al. 2006a) allows for implicitly following up the coreference chain to conclude that *its* actually refers to *acid sphingomyelinase*, and then annotating *acid sphingomyelinase* as the PATIENT instead. These sorts of annotation decisions typically distance the argument from its predicate and make it difficult for our system to find the role.

To address these three issues — unidentified proteins, conjoined predicates and coreference chains — we introduced the following additional features:

8 We experimented with larger window widths, but these models only performed worse.

- A set of orthographic features that capture some of the irregularities of protein names. These included:
  - The capitalization class of the word; one of INITIAL-UPPER, ALL-UPPER, ALL-LOWER, MIXED-UPPER-LOWER or OTHER
  - The numeric class of the word; one of YEAR-DIGITS, DIGITS, ALPHANUMERIC, SOME-DIGITS, ROMAN-NUMERAL or OTHER
  - The punctuation class of the word; one of PUNCT-ONLY, INITIAL, POSSIBLE-INITIAL, ACRONYM, or HAS- plus one or more of DOT, DASH, SLASH or COMMA for each contained in the word.
- A protein BIO-chunk label of the word, i.e. B_PROTEIN, I_PROTEIN or O. We examined both gold-standard proteins, to give us an idea of the maximum possible performance, and proteins annotated automatically by ABNER (Settles 2005), a model based on conditional random fields and orthographic and gazetteer-based features.
- A feature indicating whether or not the word was in a base-phrase conjoined with the base-phrase of the predicate, and which conjunction was conjoining them, e.g. *and* or a comma.
- A feature indicating whether or not the word was part of the last protein before a pronoun. This is essentially a poor-man's coreference resolution scheme.

Our results using these features and the features introduced previously are discussed in the next section.

6. Results

Using our word-chunking approach and the features discussed above, we prepared to train models on our training data. YamCha, our SVM-based machine learning algorithm, requires a number of different parameters to be specified: the cost of misclassification, the degree of the polynomial and the width of the feature window. To determine the best set of these parameters, we first ran a number of cross-validations on the training set, varying each parameter over a number of possible values. The parameter settings that performed best on the training set were then used to train the models we evaluated on the test set[9].

We trained the models on the following feature sets:

Baseline The basic feature set of (Hacioglu et al. 2004).

Phrasal The Baseline features plus the features we derived from the phrase-classification model in (Hacioglu et al. 2004).

Protein The Phrasal features plus our features inspired by analysis of the protein transport data: the protein BIO chunk label, our orthographic features and the conjunction and coreference features.

We evaluate these models in terms of precision, recall and F-measure. Precision gives an idea of how often our system is right when it predicts that there should be a role. It is defined as the ratio of the number of roles correctly annotated to the number

9 For all models, the best cost was 1.0, the best polynomial degree was 2 and the best window size was 2 words before and after

	Unlabeled		Labeled		
	Precision	Recall	Precision	Recall	F-measure
Baseline	80.1	66.0	79.7	65.8	72.1
Phrasal	87.9	72.2	87.5	72.0	79.0
Protein (ABNER)	88.7	72.0	88.4	71.7	79.2
Protein (gold)	90.7	79.0	90.4	78.7	84.1

Table 4
Precision, recall and F-measure for various feature sets. Unlabeled precision and recall indicates our performance when labels are ignored and only the boundaries are considered. The Protein feature set is listed twice; once when proteins were determined automatically by ABNER, and once when proteins were determined from the manually annotated gold standard.

of roles our system predicted. Recall gives an idea of how many of the real roles out there in the data our system was able to find. It is defined as the ratio of the number of roles correctly annotated to the number of roles present in the test data. F-measure is the geometric mean of precision and recall, i.e. $\frac{2\times p\times r}{p+r}$. The unlabeled versions of precision and recall simply ignore the label and only check that the boundaries of the roles are correct.

Table 4 gives precision, recall and f-measure values for our models[10]. The model trained using only the simple features of (Hacioglu et al. 2004) is able to achieve a precision of 79.7 and a recall of 65.8. Adding in the features we derived from the phrase-chunking model, we achieve about an 8 point absolute improvement in precision (to 87.5) and about a 6 point improvement in recall (to 72.0). This indicates that our feature translation was effective, and our expanded word-level features capture a number of important indicators of a word's role type.

Our extended features, which were tailored to address some of the difficulties particular to protein transport predicates, make little difference when using the proteins automatically annotated by ABNER. However, when using gold standard proteins, we see about a three point gain in precision (to 90.4) and nearly a seven point gain in recall (to 78.7). This suggests that knowing the correct protein boundaries is crucially important in identifying protein transport roles. As automatic protein identification systems like ABNER improve, we should see similar improvement in the performance of our role labeler.

7. Error Analysis and Discussion

We observed that if our models were able to find a role, they typically had little trouble identifying the type of that role. The unlabeled precision and recall columns in Table 4 show this — for all models, there is only a very slight degradation in performance between when we calculate our performance solely on the boundaries (the unlabeled precision and recall) and when we calculate performance normally (the labeled precision and recall). This indicates that the most difficult part of the task was in identifying which words should be part of a semantic role.

10 Unfortunately, we cannot directly compare the numbers here to those of OpenDMAP, which reported precision of 75.4 and a recall of 71.1 (Lu et al. 2006b). They evaluated only the predicate *translocation*, calculated precision and recall at the sentence level instead of the individual role level, and included identifying the predicate as part of the task.

	Precision	Recall	F-measure	% of Roles
AGENT	100.0	0.0	0.0	1
PATIENT	89.6	75.6	82.0	55
ORIGIN	91.4	80.0	85.3	11
DESTINATION	91.3	85.4	88.2	33

Table 5
Precision, recall, F-measure and percent of the total roles for each role type. These were calculated based on the result of the Protein (gold) model.

To get an idea of how difficult the different types of roles were to identify, we calculated precision, recall and F-measure on each role type for our best model, Protein (gold). The results are shown in Table 5. Our models miss all three AGENT examples in our testing data due to data sparsity issues — AGENT roles make up less than 1% of the roles in protein transport predicates. Our models perform best on DESTINATION roles, taking advantage of the fact that *nuclear* in *nuclear translocation* is almost always a DESTINATION role, and that occurrences of this pattern account for 25% of DESTINATION roles. ORIGIN roles, on which our model performs second best, have the advantage of being consistently quite close to the predicate. Only 18% of words in ORIGIN roles are more than six words away from the predicate, compared to 38% of words in DESTINATION roles and 44% of words in PATIENT roles. Roles that are closer to the predicate are easier for our system to identify because they appear within the word window our models consider during classification.

We also took a closer look at some of the mistakes that our best model was making and identified a few broad classes of errors. For about 15% of the errors, it looked like having a complete syntactic parse would have helped. In these errors, one role was often separated from the predicate by something like an appositive. In example 15, the PATIENT *protein* was missed because it was separated from the predicate by *overexpressed in prostate cancer*. Figure 1 shows that with a syntactic parse, *protein* is the head noun of the predicate's NP complement, essentially only two constituents away, compared to the six word distance without the syntactic parse[11].

(15) This [PATIENT protein], overexpressed in prostate cancer, [PREDICATE shuttles] between the cytoplasm and the nucleus.

Another 20% of the errors were due to boundary mismatches, where our system predicted shorter or longer arguments. Most such errors appeared to be due to errors in part-of-speech tagging, which was relatively common since our part-of-speech taggers were trained on Wall Street Journal text, not biomedical text. In example 16, *substrate* was tagged as a verb instead of a noun, and so our system identified only *1* as a patient, instead of the full *Insulin receptor substrate 1*.

(16) [PATIENT Insulin receptor substrate 1] [PREDICATE translocation] to the [DESTINATION nucleus]

11 We count punctuation as words as well with the word-chunking approach

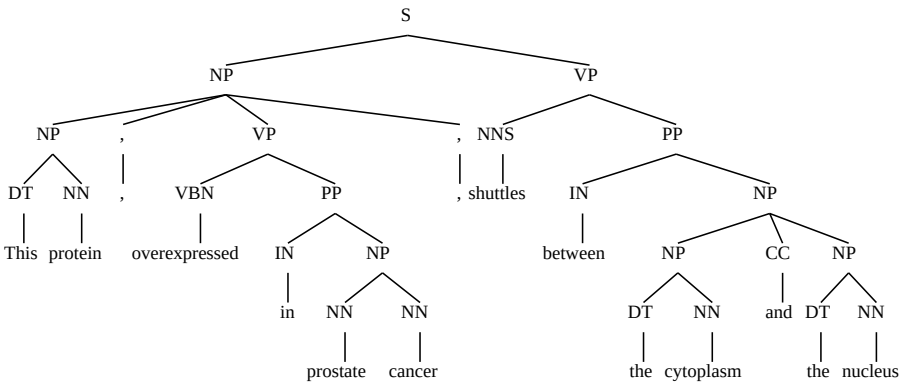

Figure 1
Syntax tree for the sentence *This protein, overexpressed in prostate cancer, shuttles between the cytoplasm and the nucleus*

Finally, another 40% of the model's errors could be attributed to trouble with roles that required tracing some sort of chain to find the argument. As discussed earlier, the scheme of (Lu et al. 2006a) allows roles to be annotated in distant parts of the sentence if some sort of coreference chain links the predicate and the distant argument. So for instance, *p53* in example 17 and *Daxx* in example 18 are marked as arguments instead of the closer pronoun *its*. Our system missed the distant PATIENT roles in both of these examples.

(17) Serine 392 exerts important effects upon [PATIENT p53] stability via the inhibition of its [ORIGIN nuclear] [PREDICATE export] mechanism.

(18) Tryptophan 521 and serine 667 residues of [PATIENT Daxx] regulate its [ORIGIN nuclear] [PREDICATE export] during glucose deprivation

These three classes of errors accounted for 75% of the errors made by our system[12]. They suggest that future research on protein transport roles could benefit from features derived from full syntactic parses, from training part-of-speech taggers and other components on biomedical text, and by better characterizing the ways in which coreference may link arguments to their predicates.

8. Conclusions

We have presented a model for identifying the semantic roles of protein transport predicates. Because the protein transport predicates considered here appear most frequently as nouns, and because most state-of-the-art syntactic parsers give little or no structure to noun phrases, our model could not follow the constituent classification paradigm typically used for semantic role labeling. Instead, we based our model on a word-chunking paradigm and trained support vector machine classifiers to classify words as being at the beginning, inside or outside of a role. We trained these models using

12 The remaining 25% of the errors were harder to diagnose. We leave analysis of them for future work.

the features of previous word-chunking models, features adapted from other types of models, and features derived from analysis of our protein transport data.

In the end, our models were able to achieve a precision of 90.4 and a recall of 78.7 using gold-standard protein boundaries, and a precision of 88.4 and a recall of 71.7 using automatic ones. Analysis of our errors suggested that future research should focus on adding more syntactic features, improving the performance on biomedical text of components like part-of-speech taggers and protein identifiers, and creating features that can help identify roles that are linked to their predicates through coreference.

9. Acknowledgments

This research was performed under an appointment of the first author to the Department of Homeland Security (DHS) Scholarship and Fellowship Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and DHS. ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-AC05-06OR23100. The third and fourth authors worked under National Library of Medicine grant 5R01LM008111-03. Computer time was provided by NSF ARI Grant #CDA-9601817, NSF MRI Grant #CNS-0420873, NASA AIST grant #NAG2-1646, DOE SciDAC grant #DE-FG02-04ER63870, NSF sponsorship of the National Center for Atmospheric Research, and a grant from the IBM Shared University Research (SUR) program. All opinions expressed in this article are the author's and do not necessarily reflect the policies and views of DHS, DOE, ORAU/ORISE or other sponsors.

References

Alfarano, C., C. E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M. J. Dumontier, M. R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J. P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J. J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B. F. Ouellette, and C. W. Hogue. 2005. The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res*, 33(Database issue), January.

Bhalotia, G., P. I. Nakov, A. S. Schwartz, and M. A. Hearst. 2003. Biotext team report for the trec 2003 genomics track. In *Proceedings of the Twelfth Text REtrieval Conference*, pages 612–621.

Carreras, Xavier and Lluís Màrquez. 2004. Introduction to the conll-2004 shared task: Semantic role labeling. In *CoNLL-2004 Shared Task*.

Carreras, Xavier and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *CoNLL-2005 Shared Task*.

Dalbey, R. E. and Gunnar Von Heijne. 2001. *Protein Targetting, Transport, and Translocation*. Academic Press, September.

Filmore, Charles J., Charles Wooters, and Collin F. Baker. 2001. Building a large lexical databank which provides deep semantics. In *The Pacific Asian Conference on Language, Information and Computation*.

Girju, Roxana, Ana-Maria Giuglea, Marian Olteanu, Ovidiu Fortu, Orest Bolohan, and Dan Moldovan. 2004. Support vector machines applied to the classification of semantic relations in nominalized noun phrases. In *The HLT/NAACL Workshop on Computational Lexical Semantics*.

Hacioglu, Kadri. 2004. A lightweight semantic chunking model based on tagging. In *HLT/NAACL-04*.

Hacioglu, Kadri, Sameer Pradhan, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2004. Semantic role labeling by tagging syntactic chunks. In *CoNLL-2004 Shared Task*.

Hacioglu, Kadri and Wayne Ward. 2003. Target word detection and semantic role chunking using support vector machines. In *HLT/NAACL-03*.

Hersh, W. and R. T. Bhupatiraju. 2003. Trec genomics track overview. In *Proceedings of The Twelfth Text REtrieval Conference*.

Horn, F., A. L. Lau, and F. E. Cohen. 2004. Automated extraction of mutation data from the literature: application of mutext to g protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*, 20(4):557–568, March.

Jelier, B., Schwartzuemie M., C. van der Fijk, M. Weeber, E. van Mulligen, and B. Schijvenaars. 2003. Searching for generifs: concept-based query expansion and bayes classification. In *Proceedings of The Twelfth Text REtrieval Conference*.

Kingsbury, Paul and Martha Palmer. 2002. From treebank to propbank. In *Language Resources and Evaluation*.

Kudo, Taku and Yuji Matsumoto. 2001. Chunking with support vector machines. In *North American Chapter of the Association for Computational Linguistics*.

Lapata, Maria. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.

Lu, Z., M. Bada, P. Ogren, K. B. Cohen, and L. Hunter. 2006a. Improving biomedical corpus annotation guidelines. In *Proceedings of the Joint BioLINK and 9th Bio-Ontologies Meeting, Fortaleza, Brazil*.

Lu, Z., K. B. Cohen, and L. Hunter. 2006. Finding generifs via go annotations. In *Proceedings of PSB 2006*, pages 52–61.

Lu, Z., J. Firby, W. Jr. Baumgartner, K. B. Cohen, P. Ogren, and L. Hunter. 2006b. Ontology-driven analysis of complex relationships in biomedical text: Extracting protein transport information from generifs. submitted.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Mitchell, J. A., A. R. Aronson, J. G. Mork, Folk L. C., S. M. Humphrey, and J. M. Ward. 2003. Gene indexing: characterization and analysis of nlm's generifs. In *Proceedings of AMIA 2003 Symposium*.

Muller, H. M., E. E. Kenny, and P. W. Sternberg. 2004. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11), November.

Palmer, Martha, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Pradhan, Sameer, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.

Pradhan, Sameer, Honglin Sun, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2004. Parsing arguments of nominalizations in english and chinese. In *The Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL-2004)*.

Punyakanok, Vasin, Peter Koomen, Dan Roth, and Wen tau Yih. 2005. Generalized inference with multiple semantic role labeling systems. In *CoNLL-2005 Shared Task*.

Ratnaparkhi, Adwait. 1996. A maximum entropy part-of-speech tagger. In *The Empirical Methods in Natural Language Processing Conference*.

Rubinstein, R. and I. Simon. 2005. Milano – custom annotation of microarray results using automatic literature searches. *BMC Bioinformatics*, 6:12.

Settles, Burr. 2005. Abner: an open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

Shah, P. K., L. J. Jensen, S. BouÃľ, and P. Bork. 2005. Extraction of transcript diversity from scientific literature. *PLoS Comput Biol*, 1(1), June.

Surdeanu, Mihai, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *ACL 2003*.

Toutanova, Kristina, Aria Haghighi, and Christopher D. Manning. 2005. Joint learning improves semantic role labeling. In *ACL 2005*.

Tsai, Richard Tzong-Han, Wen-Chi Chou, Yu-Chun Lin, Cheng-Lung Sung, Wei Ku, Ying-Shan Su, 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 *BioNLP-2006*.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. 2006. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 34(Database issue), January.

Xue, Nianwen and Martha Palmer. 2004. Calibrating features for semantic role labeling. In *EMNLP-2004*.