A Semantic Parser for Neuro-degenerative Disease Knowledge Discovery

I. Burak Ozyurt
Department of Psychiatry, UCSD
LOCI MC 9151-B 9500 Gilman Drive
La Jolla, CA 92093

Abstract

Ever increasing size of the biomedical literature makes tapping into implicit knowledge in scientific literature a necessity for knowledge discovery. In this paper, a semantic parser for recognizing semantic roles and named entities in individual sentences of schizophrenia related scientific abstracts is described. The named entity recognizer, CRFNER, outperforms ABNER in biological named entity recognition and achieves 82.5% micro-averaged F_1 on clinical psychology/neuroscience named entities. Support vector machine based semantic role labeling system achieves 75.3% micro-averaged F_1 for semantic role identification and classification on schizophrenia corpus.

Introduction

With the goal of automatically creating a structured knowledge base from unstructured scientific abstracts for neuro-degenerative disease knowledge discovery, a suite of NLP and machine learning tools is being currently developed. Extraction of useful information and relations from schizophrenia abstracts involves both identifying named entities (both common biological and cognitive psychology/neuroscience domain specific terms) and identifying and classifying semantic roles of a predicate in single sentences, i.e. semantically parsing sentences. While, discourse analysis is necessary for combining and relating extracted structured information across multiple sentences of an abstract, it is not addressed in this paper.

Named entities like dosage information, protein, drug and disease names are not only useful by themselves in building databases and/or controlled vocabularies, but they are also useful for higher order NLP tasks including semantic role labeling and question answering. Named entity recognition can be cast as relational learning task. A classification task of predicting outputs $\mathbf{Y}$ from provided inputs $\mathbf{X}$, can be approached, probabilistically, by estimating the conditional probability $P(\mathbf{Y}|\mathbf{X})$.

A Conditional Random Field (CRF) (Lafferty, McCallum, & Pereira 2001) is a Markov random field that is globally conditioned on input $\mathbf{X}$. One particular type of CRF model, particularly suitable for modeling natural language sequences is linear-chain CRF, which can be considered forming a discriminative-generative pair with hidden Markov models (HMM) (Sutton & McCallum 2006). Unlike HMMs and Maximum Entropy Markov models, CRFs don't suffer from label bias problem (Lafferty, McCallum, & Pereira 2001). A first-order linear-chain CRF is defined as

$$p_\Lambda(\mathbf{y}|\mathbf{x}) = \frac{\mathbf{1}}{\mathbf{Z(x)}} \exp(\sum_{\mathbf{t=1}}^{\mathbf{T}} \sum_{\mathbf{k=1}}^{\mathbf{K}} \lambda_\mathbf{k} \mathbf{f_k}(\mathbf{y_{t-1}}, \mathbf{y_t}, \mathbf{x}, \mathbf{t}))$$

In this log-linear model, $f_k(\mathbf{y_{t-1}}, \mathbf{y_t}, \mathbf{x}, \mathbf{t})$ is one of the k feature functions depending potentially on all input variables and only on the current and previous output values, $\Lambda = \{\lambda_k\}$ is the set of weights to be estimated and $Z(\mathbf{x})$ is partition function normalizing the clique potentials assuring proper conditional probabilities.

Conditional random fields are successfully applied to recognize titles, abstracts, authors, keywords etc. in computer science papers (Peng & McCallum 2004) and to recognize biological named entities in biomedical abstracts (Settles 2004). We present a CRF based named entity recognizer, CRFNER, extended with syntactic and semantic features that outperforms ABNER (Settles 2004) on the schizophrenia corpus for biological named-entity recognition (NER). It also allows for the recognition of ten additional named entities deemed to be important in building a structured knowledge base for neuro-degenerative disease knowledge discovery and question answering.

The Berkeley FrameNet project (Baker, Fillmore, & Lowe 1998) has created a large on-line English lexical database based on frame semantics. A semantic frame is a script-like conceptual structure describing a particular situation, object or event including its participants and properties. The frames are hierarchical and compositional. Complex interactions and situations can evoke multiple frames simultaneously. Even, within a single sentence multiple frames can be invoked. A frame is evoked by a lexical unit (LU) occurring in a sentence. A lexical unit is defined as a pairing of a word with a meaning, and not restricted to verbs. In this study, however, the only lexical units considered are verbs (predicates). The participating constituents of a sentence involved in an evoked frame are defined as frame elements (FE) or semantic roles. As of version 1.3, FrameNet database has 6000 fully annotated lexical units, nearly 800 frames and

more than 135,000 annotated sentences. Despite its size, the corpus coverage of FrameNet is limited in specialty areas, especially cognitive psychology and neuroscience. Therefore, based on semantic frames in FrameNet, schizophrenia abstracts are hand-annotated to extend FrameNet. The parse tree for a short annotated sentence from the schizophrenia corpus for the frame `Change_position_on_a_scale` is shown in Figure 1. The four semantic roles (or FEs) for this frame are namely, Item, Manner, Difference and Duration. The lemma of the predicate of this sentence is `decrease`. Based on the assumption that semantic structure of a sentence can be identified from syntactic structural features and limited semantic word knowledge information, role identification and classification task is taken as a classification task. A two stage support vector machine based role labeling system similar to the system described in Pradhan et al. (2005) is introduced. The main differences are cost model for argument (semantic role) identification, classifier ensembles for argument classification and extended feature sets including novel semantic features.

Datasets and Preprocessing

To proceed with the goal of creating a structured knowledge database for neuro-degenerative disease researchers from unstructured textual data, the first 50,000 abstracts returned from a PubMED (the National Library of Medicine's search service) search returned for the keyword 'schizophrenia' are selected as the unstructured corpus. The annotation effort is limited to the first 1000 abstracts, however. This dataset consists of abstract title and body; author and journal information is not used. Each abstract body is first separated into individual sentences by a sentence boundary detector. The implemented sentence boundary detector takes into account acronyms, decimal numbers etc. since they can be easily mistaken by naive sentence boundary detection relying only on punctuation for sentence endings, resulting in spurious sentences. The detected sentences are parsed using Charniak's syntactic parser (Charniak 2000), which also provides part-of-speech (POS) tags for the parsed sentences. All 50,000 abstracts, 370,950 sentences, are syntactically parsed. With the ultimate goal, knowledge discovery, in mind, to select the most promising predicates that will potentially evoke interesting frames which will help to identify important relations between entities of interest, a predicate frequency table is constructed. Under the assumption that the more frequent predicates in schizophrenia abstracts will be used to describe common areas of concern and will contain more information than the less frequent ones, the predicates of more-representative 50,000 schizophrenia abstracts are sorted by decreasing frequency. Starting from the most frequent predicate, the frames that can be evoked is selected from the FrameNet. A frame is selected if the sense of the frame evoking predicate is present in a random sample of abstract sentences with the same predicate and the frame has some annotated sentences. From the first hundred most frequent predicates, sixteen predicates are selected and 1960 sentences from the first 1000 abstracts are hand-annotated for semantic role labeling task. These sixteen predicates evoke thirteen frames namely, Assessing, Cause_change_of_position_on_a_scale, Change_position_on_a_scale, Communicate_categorization, Cure, Evidence, Inclusion, Inspecting, Reasoning, Research, Scrutiny, Statement and Supply. Important entity relationships including drug disease symptom relations, experiment and assessment method relationships and effect of experimental design parameters on the experimental results can be extracted from this set of frames. FrameNet 1.3, contains only 631 annotated sentences for those sixteen predicates with 1516 semantic roles. In total, the combined annotated corpus for training and testing consists of 2591 sentences and 5530 role labels.

A CRF based Named Entity Recognizer

Conditional random fields (CRF) are applied to a particular natural language processing (NLP) task, namely, named entity recognition (NER), to detect fifteen named entity types from schizophrenia abstracts.

Two sets of named entities of interest, first being biological named entities, namely, Protein, DNA, RNA, Cell Line and Cell Type; second being combination of generic named entities consisting of Time, Location, Organization, Nationality and Percentage, and named entities more specific to clinical psychology/neuroscience objectives, namely, Drug, Disease, Dosage, Age and Clinical Assessment. In total, fifteen named entities have to be recognized. For the biological named entity set, ABNER is used to bootstrap the hand labeling of 8800 sentences for the 1000 PubMED abstracts. The hand-labeling is performed by a biochemist/chemical engineer specialized in molecular biology. For the second named entity set, hand-crafted regular expressions are used to select a subset of sentences from the same 1000 PubMED abstracts for hand-tagging and bootstrapping. Separate sets of regular expressions for each named entity type is incrementally crafted in an iterative fashion. This iterative process is applied independently to each named entity type. The goal of each iteration is to increase the coverage of sentences selected for annotation by adding to or modifying the corresponding regular expression set. The iterations are stopped when no new sentences can be added. All the sentences selected for each named entity type are merged after this process. 3662 out of 8800 sentences are selected this way and hand-labeled by the author. In total 8866 named entities are annotated.

For named entity recognition (NER) task, two sets of binary features are used. The first set consists of mostly orthographic features commonly used in other NER systems and approaches, including in the identity of the word at time t in the sequence (sentence), if the word is all in uppercase, has a certain prefix/suffix etc. The second set of features, called for here on as the extended feature set, include syntactic features (e.g. POS tags, whether the word is part of a non-recursive noun phrase) and semantic ones (e.g. word/phrase is in the supplied country/region list, word has a hyponym of a more general semantic category in a lexico-semantic database). The feature set is further enhanced by conjunctions of orthographic features for the word at current time step in the sequence with the features of words at previous and next time steps in the sequence.

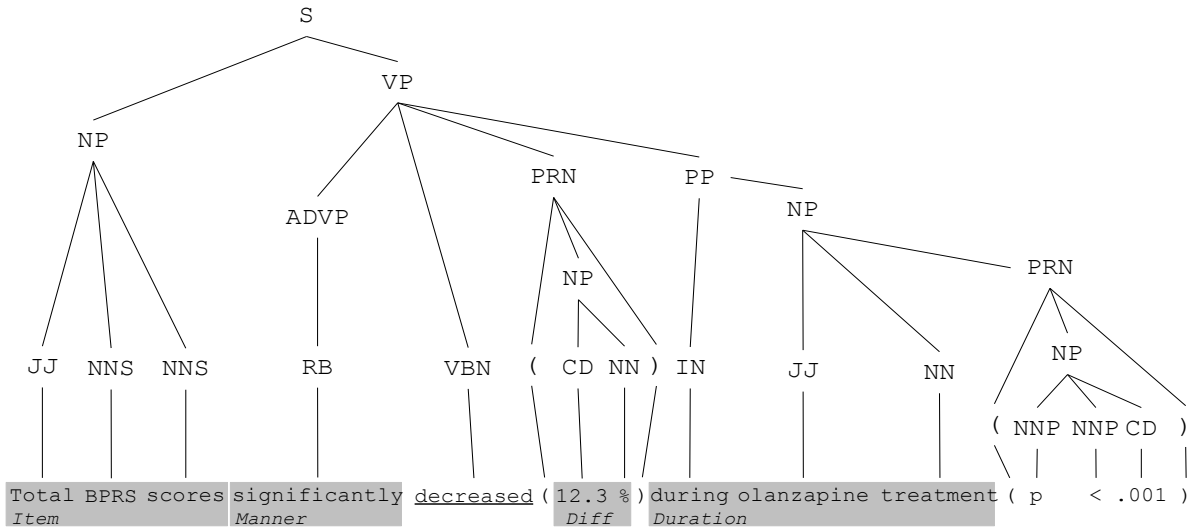

Figure 1: An example syntax tree from Charniak parser evoking Change_position_on_a_scale semantic frame with semantic roles highlighted.

Country and region lists are extracted from the online version of CIA World Factbook. The non-recursive noun phrases are tagged by using a transformation learning-based noun phrase chunker (Ngai & Florian 2001). As in ABNER (Settles 2004), words are also assigned to generalized word classes as used in (Collins 2002).

A named entity (NE) correctly identified and labeled increments the true positive count for that named entity type. All partial matches are counted as errors. The overall performance is measured by micro-averaged precision (P), recall (R) and F_1. For CRF training and labeling, as in ABNER, MALLET (McCallum 2002) is used.

For biological named entities, the first 5000 sentences of 8800 sentence corpus and for the second set, the first 1800 sentences of the 3662 sentence corpus are used for training. To test if ABNER's state of the art performance for biological named entities (F_1 around 70%) can be replicated in other domains, the first annotated set is labeled via ABNER. To compare ABNER and CRFNER, ABNER and CRFNER are both trained on the first 5000 sentences of the 8800 sentence schizophrenia corpus.

For the second set of named entity types, CRFNER is trained with baseline features and different combinations of features from the extended set. The results are summarized in Table 1.

Table 1: NER Performance Results

	P	**R**	**F_1**
BIO NE Set			
ABNER	40.94	58.95	48.32
ABNER (retrained)	62.15	43.33	51.06
CRFNER (POS+NPC)	68.21	46.61	55.38
Second NE Set			
CRFNER (Base)	88.26	76.9	82.19
CRFNER+LOC	88.4	77.3	82.48

Features for Semantic Role Labeling

In total, seven different feature sets are considered for semantic role labeling. The BASELINE feature set contains eight features, namely the lemma of the predicate, the voice of the sentence (active or passive), the head word of the constituent, the POS tag of head word, the position of the constituent relative to the predicate, phrase type of the constituent, the subcategorization frame and constituent path. This feature set is first proposed by (Gildea & Jurafsky 2002).

The EXTENDED feature set adds syntactic frame and surface distance to the predicate from the constituent as proposed by (Xue & Palmer 2004) and the lemma of the content word of the constituent and its POS tag using robust versions of the rules proposed in (Surdeanu *et al.* 2003).

The EXT_SEM feature set adds two novel ternary-valued semantic features to the EXTENDED feature set. The semantic features determine, if the noun head or content word of a constituent is animate , non-animate or unknown using WordNet 3.0 (Fellbaum 1998) as the common knowledge source. In case of multiple word senses, a sense-frequency weighted voting mechanism is used to determine if the noun is animate or not.

The FULL feature set adds 15 binary named entity features as described in previous section to the EXT_SEM feature set. If a portion of a constituent is matched to a named entity, the corresponding binary feature is set.

Specifically for argument identification, a subset of baseline features is proposed by Xue and Palmer (Xue & Palmer 2004). XUE_PALMER_IDENT feature set contains predicate lemma, head word and its POS tag, phrase type and constituent path. Two additional feature sets for argument identification are also introduced. The EXT_XUE_PALMER including content word and its POS tag and EXT_SEM_XUE_PALMER feature set including additional semantic features for noun head and content words.

Classifier

The semantic parsing task is handled by a two level cascaded SVM classifiers. All of the classifiers are binary one-against-all (OVA) SVMs. The classifiers can be configured to generate probabilistic outputs using Platt's method (Platt 2000), which is also used by Pradhan et al. (Pradhan *et al.* 2005) in their SVM based semantic role labeling system. The parameters of the sigmoids used for SVM output to probability mapping are estimated by using three-fold stratified 25% of the training data as holdout set. The first level classifier is used to filter out constituents which cannot be semantic roles.

The second level of classifiers consist of one SVM per semantic role. Unlike ProbBank (Palmer, Gildea, & Kingsbury 2005) role labels, FrameNet role labels (FEs) are fine-grained, reflecting the richness of the frame structure. Most of the frame elements are unique to a small set of related semantic frames and can be considered of forming cliques which can be handled independent of each other. Thus, OVA classifiers for argument (role) classification can be trained and tested on the set of training data evoked by the semantic frames of the classifier clique instead of using the whole training or testing set, reducing training and testing time dramatically. Within a classifier clique, the all in the one-against-all, for a particular argument classifier, means all the training instances destined for the clique minus the particular argument the classifier is trained. During classification, the predicate of a sentence that determines the set of possible frames that can be evoked by it, is used in the gating function to relay the data instance to corresponding classifier ensemble(s). The classifier ensembles are responsible for determining both the mostly likely frame from their list of frames and assigning labels to the corresponding constituents. The most likely frame is selected as the frame with maximum averaged score (or probability with probabilistic outputs) from each corresponding SVM classifier. After that, the label for a constituent is selected as the label of classifier with the maximum score (or probability) belonging to the frame. The system is implemented as components which can be chained together. A simple workflow manager system allows easy configuration and chaining of these components. For SVM learning and classification, SVM^{light} (Joachims 1998) package is used.

For argument identification, the actual number of arguments (roles) are usually an order of magnitude smaller than non-argument constituents. Also, a false positive is more desirable than a false negative, since a false negative will guarantee that the argument classification will fail because the data instance will be incorrectly filtered. During SVM training, assigning cost factors as additional constraints to argument training instances that are different from those assigned to non-argument training instances may overcome the unbalanced nature of the argument identification. Whenever a cost model is used in experiments, cost factors 15 and 1 are used for arguments and non arguments, respectively.

Results and Discussion

The hand-annotated 1960 sentence schizophrenia dataset is randomly split into two stratified sets of equal sizes. The training set for semantic role identification and classification is comprised of the combined 631 sentences from FrameNet 1.3 and the first of the split stratified schizophrenia dataset. The other set is reserved for testing. Using the generated parse trees, 23793 constituents are extracted. More than thirty feature set and parameter combinations are tested. Table 2 summarizes the most interesting combinations. The baseline run *run1* uses BASELINE feature set for both argument identification and classification, without probabilistic outputs for classifiers and no cost model for argument identification. The runs *run2*, *run3* and *run4* change the argument classification feature set to EXTENDED, EXTENDED_SEM and FULL, respectively, while keeping other parameters constant at baseline run level. The runs *run5*, *run6*, *run7*, *run8* and *run9* change the argument identification feature set from BASELINE to XUE_PALMER_IDENT, EXT_XUE_PALMER, EXT_SEM_XUE_PALMER, EXTENDED_SEM and FULL, respectively, while keeping other parameters constant at baseline run level. The run *run10* tests the effect of using probabilistic outputs for argument classification while keeping the other parameters at the baseline. The run *run11* tests the effect of using a cost model for argument identification while all other parameters are at the baseline. The run *run12* tests the combined effect of cost model for argument identification with EXTENDED_SEM feature set and probabilistic outputs for argument classification also with EXTENDED_SEM feature set. The runs *run13* and *run14*, both use EXTENDED_SEM feature set for argument identification and classification, with no probabilistic outputs and with and without cost model for argument classification, respectively. The argument classification performance is also tested with correct arguments provided simulating perfect argument identification as indicated by run *run15*. For argument classification, the best performing feature set is EXTENDED_SEM which includes novel semantic features introduced. For argument identification, the best performing feature set was also the EXTENDED_SEM maximizing final argument classification performance. Use of probabilistic outputs for either argument identification or classification has decreased the overall performance. However, incorporating cost factors slightly improved the overall role labeling performance with baseline features. The best achieved micro-averaged F_1 performance was 75.3% without cost modeling and using EXTENDED_SEM feature set for both

argument identification and argument classification. The training and classification time is dominated by argument identification step, while the argument classification classifier ensembles can all be trained in less than a minute on a Intel Dual Core T5500 1.66Ghz CPU laptop. To determine the performance of argument classification isolated from argument identification, perfect argument identification is simulated by providing only argument training instances to the classifier ensembles. The micro-averaged F_1 was 81.4% for this test.

An error analysis for false positive errors of argument classification with perfect argument identification to analyze argument classification errors only, revealed two major types of errors. The first being mix-up of certain role labels for Assessing, Evidence and Change_position_on_a_scale frames, and the second being (to a lesser extent) wrong frame selection. The pattern for the first type of errors was that some FE pairs, e.g. Attribute vs. Item FE for Change_position_on_a_scale, can only be differentiated by context, commonsense and/or domain knowledge. For some FEs like Means vs. Method of Assessing frame, the difference between their meanings is almost nonexistent but annotated as different FEs in FrameNet. Selecting wrong frame from a set of possible frames for a predicate is actually a word sense disambiguation error since each sense of a predicate is associated with a different frame. Syntactic parsing errors, which contribute to both argument identification and classification, are observed to be of predominantly PP attachment ambiguity type.

Conclusion and Future Directions

In this paper, a semantic parser consisting of a conditional random fields based named entity recognizer (CRFNER) and a cascaded two-stage SVM based semantic role labeling system is introduced. This parser allows automatic information/relation extraction to aid neuro-degenerative disease research community for determining the causes of diseases like schizophrenia, their early diagnosis and potential treatments.

The CRFNER tool outperformed both original ABNER and its retrained version on the schizophrenia corpus, thanks to its extended set of features. For the second named entity set, CRFNER achieved 82.5% micro-averaged F_1 using baseline features and country/region lexicon.

The FrameNet lexical database is augmented with schizophrenia specific annotated sentences to enable semantic role labeling on schizophrenia domain. Semantic role labeler has achieved 75.3% micro-averaged F_1 on combined semantic role identification and classification task using EXTENDED_SEM feature set including novel semantic features both for semantic label identification and classification.

The hand annotation for extending FrameNet into schizophrenia and other neuro-degenerative disease domains is further pursued. Templates for structured information extraction is under development with the cooperation of clinical psychologists for the ultimate goal of a question answering/semantic search engine system for neuro-degenerative diseases.

Table 2: Semantic Role Labeling Results

	ArgIdent			**ArgClass**		
Run	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Baseline						
run1	75.9	72.3	74.0	80.6	67.1	73.2
ArgClass Features varied						
run2	75.9	72.3	74.0	81.9	67.6	74.1
run3	75.9	72.3	74.0	82.2	67.7	74.2
run4	75.9	72.3	74.0	81.3	65.5	72.6
ArgIdent Features varied						
run5	77.2	77.9	77.6	82.3	68.2	74.6
run6	76.4	78.7	77.5	83.1	67.7	74.6
run7	76.9	78.5	77.7	82.8	67.4	74.3
run8	75.7	77.0	76.3	82.3	68.3	74.7
run9	75.5	74.9	75.2	81.8	67.4	73.9
ArgClass Prob. Outputs						
run10	75.9	72.3	74.0	82.3	63.8	71.9
ArgIdent Cost Model						
run11	77.1	71.0	73.9	80.7	68.2	74.0
Cost Model + Prob. Outputs						
run12	75.7	76.9	76.3	83.3	63.5	72.1
Best overall performance						
run13	75.7	76.9	76.3	83.3	68.4	75.1
run14	75.7	77.0	76.3	83.6	68.5	75.3
run15	Labels Provided			89.0	75	81.4

Acknowledgments

This research was supported by 1 U24 RR021992 to the Function Biomedical Informatics Research Network (BIRN, http://www.nbirn.net), that is funded by the National Center for Research Resources (NCRR) at the National Institutes of Health (NIH). Special thanks for Sinem Ozyurt MS Chem, MS ChE for annotating biological named-entities.

References

Baker, C.; Fillmore, C.; and Lowe, J. 1998. The Berkeley Framenet project. In *Proceedings of COLING-ACL-1998*.

Charniak, E. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL*, 132–139.

Collins, M. 2002. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *Proceedings of Association for Computational Linguistics Conference*, 489–496.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

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

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning* 137–142.

Lafferty, J.; McCallum, A.; and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting

and labeling sequence data. In *Proceedings of ICML-2001*, 282–289.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Ngai, G., and Florian, R. 2001. Transformation-based learning in the fast lane. In *Proceedings of North American ACL 2001*, 40–47.

Palmer, M.; Gildea, D.; and Kingsbury, P. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–106.

Peng, F., and McCallum, A. 2004. Accurate information extraction from research papers using conditional random fields. In *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT/NAACL-04)*.

Platt, J. 2000. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Smola, A.; Bartlett, P.; Schoelkopf, B.; and Schuurmans, D., eds., *Advances in Large Margin Classifiers*, 61–74.

Pradhan, S.; Hacioglu, K.; Krugler, V.; Ward, W.; Martin, J.; and Jurafsky, D. 2005. Support vector learning for semantic argument classification. *Machine Learning* 60:11–39.

Settles, B. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA)*.

Surdeanu, M.; Harabagiu, S.; Williams, J.; and Aarseth, P. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL-2003*.

Sutton, C., and McCallum, A. 2006. An introduction to conditional random fields for relational learning. In Getoor, L., and Taskar, B., eds., *Introduction to Statistical Relational Learning*. MIT Press.

Xue, N., and Palmer, M. 2004. Calibrating features for semantic role labeling. In *Proceedings of EMNLP-2004*.