Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields (SRL-2004), Banff, Canada, July 2004

Relational Markov Networks for Collective Information Extraction

Razvan C. Bunescu RAZVAN@CS.UTEXAS.EDU
Raymond J. Mooney MOONEY@CS.UTEXAS.EDU
Department of Computer Sciences, The University of Texas at Austin, 1 University Station C0500, Austin, TX 78712-0233

Abstract

Most information extraction (IE) systems treat separate potential extractions as independent. However, in many cases, considering influences *between* different potential extractions could improve overall accuracy. Statistical methods based on *undirected* graphical models, such as *conditional random fields* (CRFs), have been shown to be an effective approach to learning accurate IE systems. We present a new IE method that employs Relational Markov Networks, which can represent arbitrary dependencies between extractions. This allows for "collective information extraction" that exploits the mutual influence between possible extractions. Experiments on learning to extract protein names from biomedical text demonstrate the advantages of this approach.

1. Introduction

Information extraction (IE), locating references to specific types of items in natural-language documents, is an important task with many practical applications. Since IE systems are difficult and time-consuming to construct, most recent research has focused on empirical techniques that automatically construct information extractors by training on supervised corpora (Cardie, 1997; Califf, 1999). One of the current best empirical approaches to IE is *conditional random fields* (CRF's) (Lafferty et al., 2001). CRF's are a restricted class of *undirected graphical models* (Jordan, 1999) designed for sequence segmentation tasks such as IE, part-of-speech (POS) tagging (Lafferty et al., 2001), and shallow parsing (Sha & Pereira, 2003). In a recent follow-up to previously published experiments comparing a large variety of IE-learning methods (including HMM, SVM, MaxEnt, and rule-based methods) on the task of tagging references to human proteins in Medline abstracts (Bunescu et al., 2004), CRF's were found to significantly out-perform competing techniques.

As typically applied, CRF's, like almost all IE methods, assume separate extractions are independent and treat each potential extraction in isolation. However, in many cases, considering influences *between* extractions can be very useful. For example, in our protein-tagging task, repeated references to the same protein are common. If the context surrounding one occurrence of a phrase is very indicative of it being a protein, then this should also influence the tagging of another occurrence of the same phrase in a different context which is not indicative of protein references.

Relational Markov Networks (RMN's) (Taskar et al., 2002) can be seen as a generalization of CRF's that allow for *collective classification* of a set of arbitrarily related entities by integrating information from features of individual entities as well as the relations between them. Results on classifying connected sets of web pages have verified the advantage of this approach (Taskar et al., 2002). In this paper, we present an approach to *collective information extraction* using RMN's that simultaneously extracts all of the information from a document by exploiting the textual content and context of each relevant substring as well as the document relationships between them. Experiments on human protein tagging demonstrate the advantages of collective extraction on several annotated corpora of Medline abstracts.

2. The RMN Framework for Entity Recognition

Assume we are given a collection of training documents D where all named entities have been manually annotated. We associate with each document $d \in D$ a set of candidate entities $d.E$, in our case a restricted set of token sequences from the document. Each entity $e \in d.E$ is characterized by a set of boolean features $e.F$. This set of features is the same for all candidate entities, and it can be assimilated with the relational database definition of a table. One particular feature is $e.label$ which is set to 1 if e is considered a valid extraction, and 0 otherwise. In our document model, labels are the only hidden features, and the inference procedure will try to find a most probable assignment of values to labels, given the current model parameters.

Each document is associated with an undirected graphical model, with nodes corresponding directly to entity features, one node for each feature of each candidate entity in the document. The set of edges is created by matching *clique templates* against the entire set of entities $d.E$. A clique template is a procedure that finds all subsets of entities satisfying a given constraint and then connects feature nodes associated with the entities in each subset so that they form a clique.

Formally, there is a set of clique templates C, with each template $c \in C$ defined by:

1. A matching operator M_c for selecting subsets of entities.
2. A selected set of features $S_c = \langle X_c, Y_c \rangle$ for entities returned by the matching operator. X_c denotes the observed features, while Y_c refers to the hidden labels.
3. A clique potential ϕ_c that gives the compatibility of each possible configuration of values for the features in S_c, s.t. $\phi_c(s) \geq 0, \forall s \in S_c$.

Given a set, E, of nodes, $M_c(E) \subseteq 2^E$ consists of subsets of entities whose selected feature nodes S_c are to be connected in a clique. In previous applications of RMNs, the selected subsets of entities for a given template have the same size; however, our clique templates may match a variable number of entities. The set S_c may contain the same feature from different entities. Usually, for each entity in the matching set, its label is included in S_c. Depending on the number of hidden labels in Y_c, we define two categories of clique templates:

- **Local Templates** are all templates $c \in C$ for which $|Y_c| = 1$. They model the correlations between an entity's observed features and its label.
- **Global Templates** are all templates $c \in C$ for which $|Y_c| > 1$. They capture influences between multiple entities from the same document.

After the graph model for a document d has been completed with cliques from all templates, the probability distribution over the random field of hidden entity labels $d.Y$ given the observed features $d.X$ is computed as:

$$P(d.Y|d.X) = \frac{1}{Z(d.X)} \prod_{c \in C} \prod_{G \in M_c(d.E)} \phi_C(G.X_c, G.Y_c) \quad (1)$$

where $Z(d.X)$ is the normalizing partition function:

$$Z(d.X) = \sum_{Y} \prod_{c \in C} \prod_{G \in M_c(d.E)} \phi_C(G.X_c, G.Y_c) \quad (2)$$

3. Candidate Entities and Entity Features

Like most entity names, almost all protein mentions in our data are base noun phrases or parts of them. Therefore, such substrings are used to determine candidate entities. To avoid missing options, we adopt a very broad definition of base noun phrase.

Definition 1: A *base noun phrase* is a maximal contiguous sequence of tokens whose POS tags are from {*"JJ", "VBN", "VBG", "POS", "NN", "NNS", "NNP", "NNPS", "CD", "–"*}, and whose last word (the head) is tagged either as a noun, or a number.

Candidate extractions consist of base NPs, augmented with all their contiguous subsequences headed by a noun or number.

The set of features associated with each candidate is based on the feature templates introduced in (Collins, 2002), used there for training a ranking algorithm on the extractions returned by a maximum-entropy tagger. Many of these features use the concept of *word type*, which allows a different form of token generalization than POS tags. The word type is created by replacing any maximal contiguous sequences of capital letters with 'A', of lower-case letters with 'a', and of digits with '0'. For example, the word *TGF-1* would be mapped to type *A-0*. Consequently, each token position i in a candidate extraction provides three types of information: the word itself, its POS tag, and its type.

The left and right boundaries of a candidate extraction generate bigram and trigram features that combine words and word types. Other useful features are the head (last word), prefix and suffix lists of words from the candidate entity. A more detailed account of these feature templates is given in (Collins, 2002).

4. Local Clique Templates

Each feature template instantiates numerous features. For example, the candidate extraction `HDAC1 enzyme` has the head word *HD=enzyme*, the short type *ST=A0_a*, the prefixes *PF=A0* and *PF=A0_a*, and the suffixes *SF=a* and *SF=A0_a*. All other features depend on the left or right context of the entity. Feature values that occur less than three times are filtered out. If, after filtering, we are left with h distinct boolean features (f_i=v_j), we create h local (clique) templates $LT_1, LT_2, ..., LT_h$. Each template's matching operator is set to match any single-entity set. The collection of features S_i corresponding to template LT_i applied to the singleton entity set $\{e\}$ is $S_i = \langle X_i, Y_i \rangle = \langle \{ e.f_i{=}v_j \}, \{e.label\} \rangle$. The 2-node cliques created by all h templates around one entity are illustrated in Figure 1.

Each entity has a label node connected to its own set of h binary feature nodes. This leads to an excessive num-

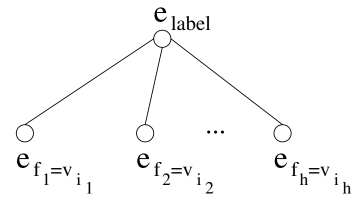

Figure 1. RMN generated by local templates.

ber of nodes in the model, most of which have the value zero. To reduce the number of nodes, we could remove the observed nodes from the graph, which then results in many one-node clique potentials (due to the observed features) being associated with the same label node. Because these clique potentials may no longer be distinguished in the MRF graph, in order to have all of them as explicit as possible in the graphical model, we decided to transform the relational Markov network into its equivalent *factor graph* representation. Factor graphs (Kschischang et al., 2001) are bipartite graphs that express how a global function of many variables (the probability $P(d.Y|d.X)$ in Equation 1) factors into a product of local functions (the potentials $\phi_C(G.X_c, G.Y_c)$ in Equation 1). Factor graphs subsume many different types of graphical models, including Bayesian networks and Markov random fields. The sum/max-product algorithms used for inference in factor graphs generalize a wide variety of algorithms including the forward/backward algorithm, the Viterbi algorithm, and Pearl's belief propagation algorithm (Pearl, 1988). To obtain the factor graph for a given Markov random field, we copy all nodes from the MRF, and create a new node for each instantiated clique potential. Each potential node is then linked to all nodes from the associated clique. However in this case, instead of creating a potential node for each feature-value pair as in the initial MRF model, we create a potential node only for the binary features that are 1 for the given entity. Correspondingly, the table associated with the potential will be reduced from 4 to 2 values. As an example, Figure 2 shows that part of the factor graph which is generated around the entity label for `HDAC1 enzyme`.

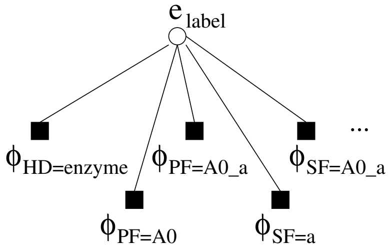

Figure 2. Factor Graph for local templates.

Note that the factor graph above has an equivalent RMN graph consisting of a one-node clique only, on which it is hard to visualize the various potentials involved. There are cases where different factor graphs may yield the same underlying RMN graph, which makes the factor graph representation preferable.

5. Global Clique Templates

Global clique templates enable us to model hypothesized influences between entities from the same document. They connect the label nodes of two or more entities, which, in the factor graph, translates into potential nodes connected to at least two label nodes. In our experiments we use three global templates:

Overlap Template (OT): No two protein names overlap in the text i.e if the span of one protein is $[s_1, e_1]$ and the span of another protein is $[s_2, e_2]$, and $s_1 \leq s_2$, then $e_1 < s_2$.

Repeat Template (RT): If multiple entities in the same document are repetitions of the same name, their labels tend to have the same value (i.e. most of them are protein names, or most of them are not protein names). Later we discuss situations in which repetitions of the same protein name are not tagged as proteins, and design an approach to handle this.

Acronym Template (AT): It is common convention that a protein is first introduced by its long name, immediately followed by its short-form (acronym) in parentheses.

5.1. The Overlap Template

The overlap template matches any two overlapping candidate entities and connects their label nodes through a potential node that forbids the case in which both have label-value 1, as illustrated in Table 1.

An alternative solution for the overlap template is to create a potential node for each token position that is covered by at least two candidate entities in the document, and connect it to their label nodes. The difference in this case is that the potential node will be connected to a variable number of entity label nodes. However this approach is better since it leads to fewer nodes being created in the document factor graph, which results in faster inference.

Table 1. Overlap Potential.

ϕ_{OT}	$e_1.label = 0$	$e_1.label = 1$
$e_2.label = 0$	1	1
$e_2.label = 1$	1	0

5.2. The Repeat Template

We could specify the potential for the repeat template in a similar 2-by-2 table, this time leaving the table entries to be learned, given that it is not a hard constraint. However we can do better by noting that the vast majority of cases where a repeated protein name is not also tagged as a protein happens when it is part of a larger phrase that *is* tagged. For example, `HDAC1 enzyme` is a protein name, therefore `HDAC1` is not tagged in this phrase, even though it was tagged previously in the abstract where it was not followed by `enzyme`. We need a potential that allows two entities with the same text to have different labels if the entity with label-value 0 is inside another entity with label-value 1. But a candidate entity may be inside more than one "including" entity, and the number of including entities may vary from one candidate extraction to another. We solve this problem, by introducing a logical OR clique template that matches a variable number of entities. When this template matches a subset of entities $e_1, e_2, ..., e_n$, it will create an auxiliary OR entity e_{or}, with a single feature $e_{or}.label$. The potential function is set so that it assigns a non-zero potential only when $e_{or}.label = e_1.label \vee e_2.label \vee ... \vee e_n.label$. The cliques are only created as needed, e.g. when the auxiliary OR variable is required by repeat and acronym clique templates.

Figure 3 shows the factor graph for a sample instantiation of the repeat template using the OR template. Here, u and v represent two same-text entities, u_1, u_2, ... , u_n are all entities that include u, and $v_1, v_2, ..., v_m$ are entities that include v. To avoid clutter, all entities in this and subsequent factor graphs stand for their corresponding label features. The potential function can either be preset to prohibit unlikely label configurations, or it can be learned to represent an appropriate soft constraint. In our experiments, it was learned since this gave slightly better performance.

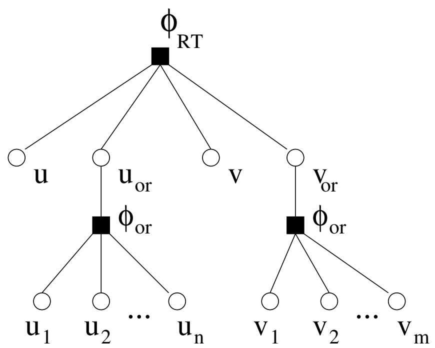

Figure 3. Repeat Factor Graph.

5.3. The Acronym Template

One approach to the acronym template would be to use an extant algorithm for identifying acronyms and their long forms in a document, and then define a potential function that would favor label configurations in which both the acronym and its definition have the same label. One such algorithm is described in (Schwartz & Hearst, 2003), achieving a precision of 96% at a recall rate of 82%. However, because this algorithm would miss a significant number of acronyms, we have decided to implement a softer version as follows: detect all situations in which a single word is enclosed between parentheses, such that the word length is at least 2 and it begins with a letter. Let v denote the corresponding entity. Let u_1, u_2, ..., u_n be all entities that end exactly before the open parenthesis. If this is a situation in which v is an acronym, then one of the entities u_i is its corresponding long form. Consequently, we use a logical OR template to introduce the auxiliary variable u_{or}, and connect it to v's node label through an acronym potential, as illustrated in Figure 4. For example, consider the phrase `the antioxidant superoxide dismutase - 1 ( SOD1 )`, where both `superoxide dismutase - 1` and `SOD1` are tagged as proteins. `SOD1` satisfies our criteria for acronyms, thus it will be associated with the entity v in Figure 4. The candidate long forms are u_1 = `antioxidant superoxide dismutase - 1`, u_2 = `superoxide dismutase - 1`, and u_3 = `dismutase - 1`.

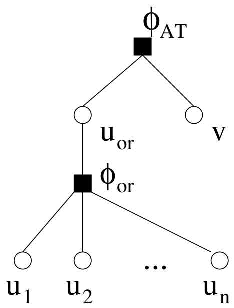

Figure 4. Acronym Factor Graph.

6. Inference in Factor Graphs

Given the clique potentials, the inference step for the factor graph associated with a document involves computing the most probable assignment of values to the hidden labels of all candidate entities:

$$Y^* = arg \max_Y P(d.Y|d.X) \tag{3}$$

where $P(d.Y|d.X)$ is defined as in Equation 1. A brute-force approach is excluded, since the number of possible label configurations is exponential in the number of candidate entities. The sum-product algorithm (Kschischang et al., 2001) is a message-passing algorithm that can be used for computing the marginal distribution over the label variables in factor graphs without cycles, and with a minor change (replacing the sum operator used for marginalization with a max operator) it can also be used for deriving the most probable label assignment. In our case, in order to get an acyclic graph, we would have to use local templates only. However, it has been observed that the algorithm often converges in general factor graphs, and when it converges, it gives a good approximation to the correct marginals. The algorithm works by altering the belief at each label node by repeatedly passing messages between the node and all potential nodes connected to it (Kschischang et al., 2001).

The time complexity of computing messages from a potential node to a label node is exponential in the number of label nodes attached to the potential. Since this "fan-in" can be large for OR potential nodes (and also for the second solution to overlap potential nodes), this step required optimization. Fortunately, due to the special form of the OR and overlap potentials, and the normalization before each message-passing step, we were able to develop a linear-time algorithm for these special cases. For example, the formulae for computing the OR messages for the sum-product algorithm are shown in Equations 4, with the relevant messages illustrated in Figure 5 (to avoid clutter, e and ϕ stand for e_{OR} and ϕ_{OR} respectively).

$$\begin{aligned}
\mu_{\phi \to e}(0) &= \prod_{i=1}^{n} \mu_{e_i \to \phi}(0) \\
\mu_{\phi \to e}(1) &= 1 - \mu_{\phi \to e}(0) \\
\mu_{\phi \to e_i}(0) &= \mu_{\phi \to e}(1) + \\
&\quad \frac{\mu_{\phi \to e}(0)}{\mu_{e_i \to \phi}(0)} \left(\mu_{\phi \to e}(0) - \mu_{\phi \to e}(1)\right) \\
\mu_{\phi \to e_i}(1) &= \mu_{\phi \to e}(1)
\end{aligned} \tag{4}$$

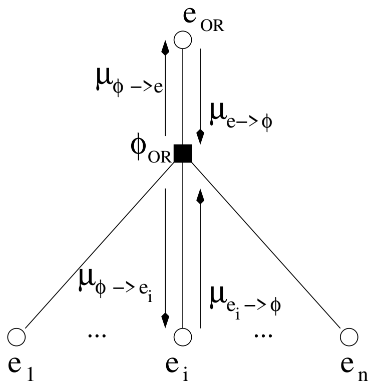

Figure 5. Messages in OR Factor Graph.

7. Learning Potentials in Factor Graphs

Following a maximum likelihood estimation, we shall use the log-linear representation of potentials:

$$\phi_C(G.X_c, G.Y_c) = exp\{\mathbf{w_c f_c}(G.X_c, G.Y_c)\}$$

Let $\mathbf{w}$ be the concatenated vector of all potential parameters $\mathbf{w_c}$. One approach to finding the maximum-likelihood solution for $\mathbf{w}$ is to use a gradient-based method, which requires computing the gradient of the log-likelihood with respect to potential parameters $\mathbf{w_c}$. It can be shown that this gradient is equal with the difference between the empirical counts of $\mathbf{f_c}$ and their expectation under the current set of parameters $\mathbf{w}$. This expectation is expensive to compute, since it requires summing over all possible configurations of candidate entity labels from a given document. To circumvent this complexity, we use the Collins' voted perceptron approach (Collins, 2002), which approximates the full expectation of $\mathbf{f_c}$ with the $\mathbf{f_c}$ counts for the most likely labeling under the current parameters, $\mathbf{w}$. In all our experiments, the perceptron was run for 50 epochs, with a learning rate set at 0.01.

8. Experimental Results

We have tested the RMN approach on two datasets that have been hand-tagged for human protein names. The first dataset is Yapex[1] which consists of 200 Medline abstracts. The second dataset is Aimed[2] which has been previously used for training the protein interaction extraction systems in (Bunescu et al., 2004). It contains 225 Medline abstracts, of which 200 are known to describe interactions between human proteins, while the other 25 do not refer to any interaction. We compared the performance of three systems: **LT-RMN** is the RMN approach using local templates and the overlap template, **GLT-RMN** is the full RMN approach, using both local and global templates, and **CRF**, which uses a CRF for labeling token sequences. We used the CRF implementation from (McCallum, 2002) with the set of tags and features used by the Maximum-Entropy tagger described in (Bunescu et al., 2004). All Medline abstracts were tokenized and then POS tagged using Brill's tagger (Brill, 1995). Each extracted protein name in the test data was compared to the human-tagged data, with the positions taken into account. Two extractions are considered a match if they consist of the same character sequence in the same position in the text. Results are shown in Tables 2 and 3 which give average precision, recall, and F-measure

[1]URL: www.sics.se/humle/projects/prothalt/
[2]URL: ftp.cs.utexas.edu/mooney/bio-data/

using 10-fold cross validation.

Table 2. Extraction Performance on Yapex.

Method	Precision	Recall	F-measure
LT-RMN	70.79	53.81	61.14
GLT-RMN	69.71	65.76	67.68
CRF	72.45	58.64	64.81

Table 3. Extraction Performance on Aimed.

Method	Precision	Recall	F-measure
LT-RMN	81.33	72.79	76.82
GLT-RMN	82.79	80.04	81.39
CRF	85.37	75.90	80.36

These tables show that, in terms of F-measure, the use of global templates for modeling influences between possible entities from the same document significantly improves extraction performance over the local approach (a one-tailed t-test for statistical significance results in a p value less than 0.01 on both datasets). There is also a small improvement over CRF's, with the results being statistically significant only for the Yapex dataset, corresponding to a p value of 0.02. We hypothesize that further improvements to the LT-RMN approach would push the GLT-RMN performance even higher. The tagging scheme used by CRFs, in which each token is assigned a tag, is essentially different from the RMN approach, where candidate extractions are either rejected or accepted. In the tagging approach used by CRFs, extracted entities are available only after tagging is complete, thereby making it difficult to account for influences between them during tagging.

9. Related Work

There have been some previous attempts to use global information from repetitions, acronyms, and abreviations during extraction. In (Chieu & Ng, 2003), a set of global features are used to improve a Maximum-Entropy tagger; however, these features do not fully capture the mutual influence between the labels of acronyms and their long forms, or between entity repetitions. In particular, they only allow earlier extractions in a document to influence later ones and not vice-versa. The RMN approach handles these and potentially other mutual influences between entities in a more complete, probabilistically sound manner.

10. Conclusions and Future Work

We have presented an approach to collective information extraction that uses Relational Markov Networks to reason about the mutual influences between multiple extractions. A new type of clique template – the logical OR template – was introduced, allowing a variable number of relevant entities to be used by other clique templates. Soft correlations between repetitions and acronyms and their long form in the same document have been captured by global clique templates, allowing for local extraction decisions to propagate and mutually influence each other. Experimental results showed that a collective approach to extraction significantly improves performance.

Regarding future work, a richer set of features for the local templates would likely improve performance. Currently, LT-RMN's accuracy is still significantly less than CRF's, which limits the performance of the full system. Another limitation is the approximate inference used by both RMN methods. The number of factor graphs for which the sum-product algorithm did not converge was non-negligible, and our approach stopped after a fix number of iterations. Besides exploring improvements to loopy belief propagation that increase computational cost (Yedidia et al., 2000), we intend to examine alternative approximate-inference methods such as Gibbs sampling, and other Monte Carlo algorithms.

A natural next step is to integrate IE subtasks like named entity recognition and coreference resolution, such that decisions made in one subtask influence decisions made in the other. The context of a pronoun referring to an entity can help in disambiguating the class of that entity through the use of a general repeat template. Recent work in anaphora resolution using RMNs (McCallum & Wellner, 2003) and the joint solving of two different NLP tasks using dynamic CRFs (McCallum et al., 2003) show the benefit of an integrated, collective approach.

11. Acknowledgements

This work was partially supported by grants IIS-0117308 and IIS-0325116 from the NSF.

References

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, *21*, 543–565.

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., & Wong, Y. W. (2004). Comparative experiments on learning information extractors for proteins and their interactions. *Special Issue in the Journal Artificial Intelligence in Medicine on Summarization and Information Extraction from Medical Documents*. To appear.

Califf, M. E. (Ed.). (1999). *Papers from the AAAI-1999*

Workshop on Machine Learning for Information Extraction. Orlando, FL: AAAI Press.

Cardie, C. (1997). Empirical methods in information extraction. *AI Magazine*, *18*, 65–79.

Chieu, H. L., & Ng, H. T. (2003). Named entity recognition with a maximum entropy approach. *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)* (pp. 160–163). Edmonton, Canada.

Collins, M. (2002). Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-02)* (pp. 489–496). Philadelphia, PA.

Jordan, M. I. (Ed.). (1999). *Learning in graphical models*. Cambridge, MA: MIT Press.

Kschischang, F. R., Frey, B., & Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, *47*, 498–519.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of 18th International Conference on Machine Learning (ICML-2001)* (pp. 282–289). Williams College, MA.

McCallum, A., Rohanimanesh, K., & Sutton, C. (2003). Dynamic conditional random fields for jointly labeling multiple sequences. *NIPS-2003 Workshop on Syntax, Semantics, Statistics*. Whistler, Canada.

McCallum, A., & Wellner, B. (2003). Toward conditional models of identity uncertainty with application to proper noun coreference. *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web* (pp. 79–86). Acapulco, Mexico.

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

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo,CA: Morgan Kaufmann.

Schwartz, A. S., & Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Proceedings of the 8th Pacific Symposium on Biocomputing* (pp. 451–462). Lihue, HI.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. *Proceedings of Human Language Technology and the Meeting of the North American Association for Computational Linguistics* (pp. 134–141). Edmonton, Canada.

Taskar, B., Abbeel, P., & Koller, D. (2002). Discriminative probabilistic models for relational data. *Proceedings of 18th Conference on Uncertainty in Artificial Intelligence (UAI-02)* (pp. 485–492). Edmonton, Canada.

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2000). Generalized belief propagation. *Advances in Neural Information Processing Systems 12* (pp. 689–695). Denver, CO.