Exploring Relational Features and Learning under Distant Supervision for Information Extraction Tasks

Ajay Nagesh
Dept. of Computer Science and Engineering, IIT Bombay.
Faculty of Information Technology, Monash University.
ajaynagesh@cse.iitb.ac.in

Abstract

Information Extraction (IE) has become an indispensable tool in our quest to handle the data deluge of the information age. IE can broadly be classified into Named-entity Recognition (NER) and Relation Extraction (RE). In this thesis, we view the task of IE as finding patterns in unstructured data, which can either take the form of features and/or be specified by constraints. In NER, we study the categorization of complex relational[1] features and outline methods to learn feature combinations through induction. We demonstrate the efficacy of induction techniques in learning : i) rules for the identification of named entities in text – the novelty is the application of induction techniques to learn in a very expressive declarative rule language ii) a richer sequence labeling model – enabling optimal learning of discriminative features. In RE, our investigations are in the paradigm of *distant supervision*, which facilitates the creation of large albeit noisy training data. We devise an inference framework in which constraints can be easily specified in learning relation extractors. In addition, we reformulate the learning objective in a max-margin framework. To the best of our knowledge, our formulation is the first to optimize multi-variate non-linear performance measures such as F_β for a latent variable structure prediction task.

1 Introduction

Most of the content that we come across in the digital media in the form of emails, blogs, webpages, enterprise data and so on are authored in natural language and have very little structure to them. With the dawn of the information age, we produce a colossal amount of unstructured data everyday. This presents an enormous challenge for machines to process, curate, search and reason in such data.

The process of automatically identifying and disambiguating entities, their attributes and relationships in unstructured data sources is termed as *Information Extraction (IE)*. *IE* facilitates a rich and structured representation of data, enabling downstream applications to process unstructured documents like a standard database. The richness present in natural language text, presupposition of world knowledge and the rapid rate of content creation makes IE a highly challenging task. As a result, it has been a very active area of research in the computational linguistics community for over two decades (Sarawagi, 2008).

A few of the challenges faced when performing information extraction: (i) *Entity Disambiguation*: `Jeff Bezos` and `Bezos` refer to the same entity. `Washington` could be either a city, a state, or a person depending on the context. (ii) *Scope Resolution*: Certain Entities such as `Washington` in "`Washington Post`" should not be labeled as a location name because the entire textual span is an organization name (iii) *Type Disambiguation*: In the sentence, *" England beat Australia 2 - 0"*. `England` and `Australia` are sports organizations. (iv) *Relation mention detection*: The co-occurrence of `Obama` and `US` in a sentence is not a sure indication that the `President` relation (obtained from a database of facts) is expressed in it.

1.1 Contributions of the thesis

The problem of Information Extraction can be viewed as that of finding patterns in the data. These patterns can either take the form of features or can be specified as constraints on the search space.

Data-driven Patterns : Feature Combinations

Let us suppose that we are given a set of basic features (*e.g.* `Caps` - *a capitalized token*; `LastName` - *occurrence in a dictionary of last-names*). Named-

[1] Terminology is borrowed from *logic*, where relational logic is more powerful than propositional logic with the inclusion of quantifiers, but is a subset of first-order logic

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 40–47,
Denver, Colorado, June 1, 2015.

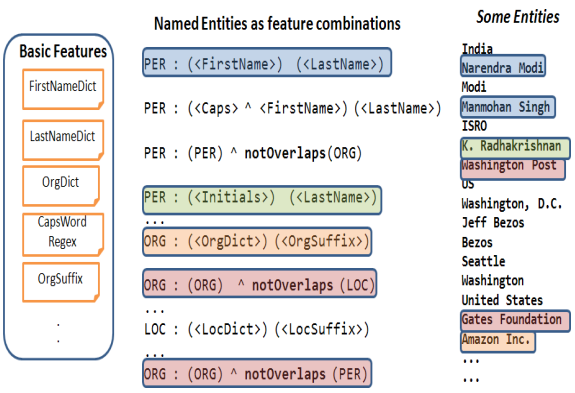

Figure 1: Patterns as Feature Combinations

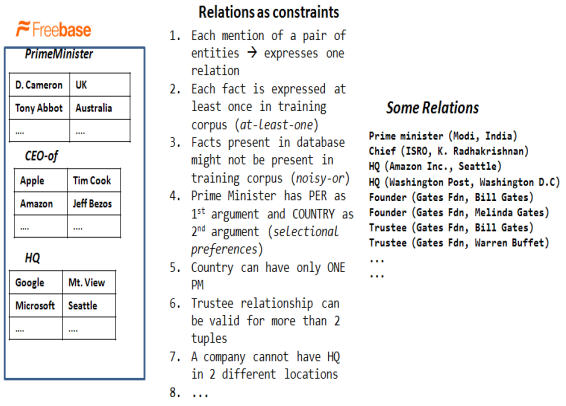

Figure 2: Patterns as Constraints

entities can be discovered by learning combinations of such features. For instance, *"if a span of text contains two tokens,* `Caps` *followed by* `LastName`*, then it is most probably a person named entity"*. We consider the previous statement as a pattern, leading to a named-entity.

Figure 1 depicts some of the basic features, a number of patterns (basic feature combinations) and the entities in text that can potentially match with these patterns. Named-entity recognition (NER) can immensely benefit from such patterns, some of which are domain-specific and others, domain-independent. Several patterns are non-trivial combinations of basic features. For instance, *"if a* `location` *name overlaps with an* `organization`*, then it is not a* `location` *named-entity"*. (*e.g.* `Washington` in `Washington Post`).

These patterns are very large in number and we could define them as feature classes. The set of features defined by them form a feature space. Since the number patterns are many and we are not sure which ones are triggered in a given piece of text, we would like to learn / induce such patterns.

In this thesis, we study the categorization of the feature classes. We also define various methods to learn feature combinations through induction. The features induced are consumed by a rule-based NER system to learn compact and "interpretable" rules that have a reasonable accuracy. We also demonstrate the use of these features in max-margin based sequence labeling models.

User-Specified Patterns : Constraints

Consider the problem of identifying relationships between entities in text. Here we can look at patterns as constraints that need to be enforced on relations extracted. Some of these are listed in Figure 2. They are few compared to the entity recognition case and can be specified by the user to restrict the search space.

For instance, we would like to enforce the following constraint: *For a* `Prime-minister` *relation, the first argument has to be a* `person` *and the second argument has to be a* `country` .

In this thesis, we look at a specific paradigm of relation extraction called *distant supervision* (Mintz et al., 2009). The goal is to learn relation extraction models by aligning facts in a database (Figure 2) to sentences in a large unlabeled corpus. Since the individual sentences are not hand labeled, the facts in the database act as "weak" or "distant" labels, and hence, the learning scenario is termed as distantly supervised. We look at ways in which constraints can be specified while learning relation extractors in this setting. We formulate an integer linear programming-based framework to facilitate the addition of constraints.

Existing distant supervision-based systems are often trained by optimizing performance measures (such as conditional log-likelihood or error rate) that are not directly related to the task-specific non-linear performance measure, *e.g.*, the F_1-score. We present a novel max-margin learning approach to optimize non-linear performance measures for distantly su-

pervised relation extraction models.

2 Learning for Named-Entity Extraction

Several problems in Machine Learning are immensely benefited from a rich structural representation of the data (Flach and Lachiche, 1999; Roth and Yih, 2001). Specifically, the tasks in Information Extraction are relation-intensive and the usage of relational features has been shown to be quite effective in practice (Califf, 1998; Roth and Yih, 2001). In this section, we define categories of predicates and discuss the complexity-based classification of relational features followed by techniques to induce features in several of these categories.

Feature Space Categorization

The relational features are in a language that is similar in expressive power as *first order definite clauses* (Horn, 1951). Predicates are defined on textual spans. The head predicate is the class label of a textual span.

We define two types of body predicates, namely, *relation* and *basic feature* predicates. A *relation* predicate is a binary predicate that represents the relationship between two *spans* of text. *E.g.* `overlaps(X,Y)`. A *basic feature* predicate is an assertion of a situation or a property of a *span* or a *subspan*. For example, `FirstName(`X`)` states that the span of text X occurs in a dictionary of first names. We illustrate each of these feature classes with an example of a typical definite clause belonging to the feature class.

1. **Simple Conjuncts (SCs)**:
   `Org(X) :- OrgGazeteer(X),CapsWord(X).`
   e.g. Microsoft
2. **Candidate Definition Features (CDs)** : These consist of the two following feature classes.
   (a) **Absolute Features (AFs)**: non-overlapping evidence predicates chained by relation predicates.
   `person-AF(X) :- contains(X, X1), FirstNameDict(X1), CapsWord(X1), before(X1,X2), contains(X, X2), CapsWord(X2).` *e.g.:* Sachin Tendulkar
   (b) **Composite Features (CFs)**: Defined as a conjunction of two AFs that share the same head predicate.
   `person(X) :- person-AF (X), leftContext(X, 1, L2), Salutation(L2).` *e.g.:* Mr. Sachin Tendulkar (note the presence of contextual clues such as salutation)
3. **Candidate Refinement Features (CRs)**: The body of the clause is defined by head predicates that belong to different class labels, and can contain negations in the body (hence, not a definite clause)
   `Loc(X) :- Loc1(X),org1(Y),¬overlaps(X,Y).`
   A span of text is a location, "*if it matches a location feature and does not overlap with an organization feature*". *e.g.:* Washington in "Washington Post" will not be marked as a location, due to this feature.

2.1 Feature Induction in a Rule-based Setting

Rule-based systems for NER achieve state-of-the-art accuracies (Chiticariu et al., 2010). However, manually building and customizing rules is a complex and labor-intensive process. In this work, we outline an approach that facilitates the process of building customizable rules for NER through rule induction. Given a set of basic feature predicates and an annotated document collection, our goal is to generate with reasonable accuracy an initial set of rules that are interpretable and thus can be easily refined by a human developer. Our contributions include (i) an efficient rule induction process in a declarative rule language, (ii) usage of induction biases to enhance rule interpretability, and (iii) definition of extractor complexity as a first step to quantify the interpretability of an extractor. We present initial promising results with our system and study the effect of induction bias and customization of basic features on the accuracy and complexity of induced rules. We demonstrate through experiments that the induced rules have good accuracy and low complexity, according to our complexity measure.

Our induction system is modeled on a four-stage manual rule development process since the overall structure of the induced rules must be similar in spirit to that which a developer who follows best practices would write. The stages of rule development and the corresponding phases of induction are summarized in Figure 3. In our system, we combine several induction techniques such as *least general generalization (LGG)*, iterative clustering, propositional rule learning in order to induce NER rules in a declarative rule language known as *Annotation Query Language (AQL)*. A brief overview of the salient aspects of our induction system is presented

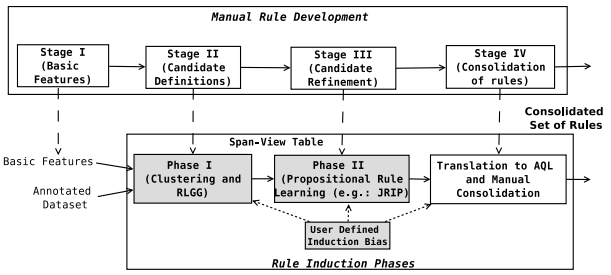

Figure 3: Correspondence between Manual Rule development and Rule Induction

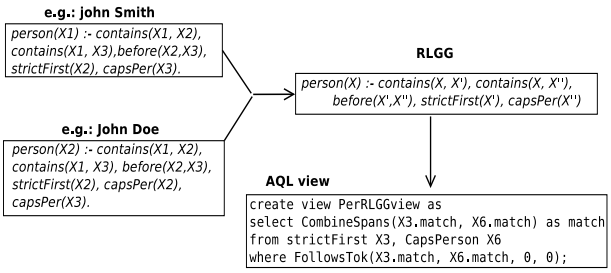

Figure 4: Relative Least General Generalization

in the following paragraphs.

Background Knowledge. We represent each example in the form of *first order definite clauses*, in conjunction with relevant background knowledge. This background knowledge will serve as input to our induction system.

Clustering and RLGG. The first phase of induction uses a combination of *clustering* and *relative least general generalization (RLGG)* techniques (Nienhuys-Cheng and Wolf, 1997; Muggleton and Feng, 1992). Using clustering, we group the examples based on the similarity of their background knowledge. This process is interleaved by RLGG where we take a set of examples and find their generalization that is analogous to the *least upper bound*. We recursively find pairwise-RLGGs of all examples in a cluster. At the end of this phase, we have a number of CD features.

The representation of an example and the RLGG procedure is shown in Figure 4.

Propositional Rule Learning. In the second phase, we begin by forming a structure known as the *span-view table*. Broadly speaking, this is an attribute-value table formed by all the features induced in the first phase along with the textual spans generated by them. The attribute-value table is used as input to a propositional rule learner such as *JRIP* to learn accurate compositions of a useful (as determined by the learning algorithm) subset of the CD features. This forms the second phase of our system. The rules learnt from this phase are in the space of CR features.

Induction Biases. At various phases, several induction biases are introduced to enhance the interpretability of rules. These biases capture the expertise gleaned from manual rule development and constrain the search space in our induction system.

Extractor Complexity. Since our goal is to generate extractors with manageable complexity, we must introduce a quantitative measure of extractor complexity, in order to (1) judge the complexity of the extractors generated by our system, and (2) reduce the search space considered by the induction system. To this end, we define a simple complexity score that is a function of the number of rules, and the number of predicates in the body of each rule of the extractor. Our simple notion of rule length is motivated by existing literature in the area of database systems.

AQL and SystemT : Advantages. The *hypothesis language* of our induction system is *AQL*, and we employ *SystemT* as the *theorem prover*. SystemT provides a very fast rule execution engine and is crucial to our induction system because we test multiple hypotheses in the search for the more promising ones. AQL provides a very expressive rule representation language that has proven to be capable of encoding all the paradigms that any rule-based representation can encode. The dual advantages of *rich rule-representation* and *execution efficiency* are the main motivations behind our choice.

We experimented with three different starting sets of basic feature predicates (with increasing accuracy and complexity) and observed that the complexity of the final set of induced rules is directly proportional to that of the initial set, both in terms of accuracy and complexity. We compared our induced set of rules with the manual rules. We achieve upto 75% accuracy of the state-of-the-art manual rules with a decrease in extractor complexity of upto 61%. For a more detailed exposition of the system and discussion of experiments, please refer to our work (Nagesh et al., 2012).

2.2 Feature Induction in a Max-margin Setting

In this piece of work, we view the problem of NER from the perspective of sequence labeling. The goal is to investigate the effectiveness of using relational features in the input space of a max-margin based sequence labeling model. Our work is based on StructHKL (Nair et al., 2012) and standard StructSVM formulations. We propose two techniques to learn a richer sequence labeling model by using relational features discussed above.

In one technique, we leverage an existing system that is known to learn optimal feature conjunctions (SCs) in order to learn relational features such as AFs and CFs. To achieve this, we propose a two-step process : (i) enumerate a good set of AFs using existing induction techniques (ii) use the StructHKL framework, which learns optimal conjunctions to learn CFs.

In the other technique, we leverage the StructSVM framework. We define a subsequence kernel to implicitly capture the relational features and reformulate the training objective.

Our experiments in sequence labeling tasks reinforce the importance of induction bias and the need for interpretability to achieve high-quality NER rules, as observed in the experiments of our previous work on rule induction.

3 Learning for Relation Extraction

In the second part of the thesis, we investigate another important problem in IE, namely, *relation extraction*. The task of extracting relational facts that pertains to a set of entities from natural language text is termed as *relation extraction*. For example, given a natural language sentence, *"On Friday, President Barack Obama defended his administration's mass collection of telephone and Internet records in the United States"*, we can infer the relation, `President(Barack Obama, United States)` between the entities `Barack Obama` and `United States`.

Our framework is motivated by distant supervision for learning relation extraction models (Mintz et al., 2009). Prior work casts this problem as a multi-instance multi-label learning problem (Hoffmann et al., 2011; Surdeanu et al., 2012). It is multi-instance because for a given entity-pair, only the label of the bag of sentences that contains both entities (aka mentions) is given. It is multi-label because a bag of mentions can have multiple labels. The inter-dependencies between relation labels and (hidden) mention labels are modeled by a Markov Random Field (Hoffmann et al., 2011).

3.1 Constrained Distant Supervision

Various models have been proposed in recent literature to align the facts in the database to their mentions in the corpus. In this work, we discuss and critically analyze a popular alignment strategy called the *"at least one"* heuristic. We provide a simple, yet effective relaxation to this strategy.

Our work extends the work by Hoffmann et al. (2011). We formulate the inference procedures in training as integer linear programming (*ILP*) problems and implement the relaxation to the *"at least one "* heuristic through a soft constraint in this formulation. This relaxation is termed as *"noisy-or"*. The idea is to model the situation where a fact is present in the database but it is not instantiated in the text.

Additionally, our inference formulation enables us to model additional type of constraints such as selectional preferences of arguments. Empirically, we demonstrate that this simple strategy leads to a better performance under certain settings when compared to the existing approaches. For additional details, please refer to our paper (Nagesh et al., 2014).

3.2 Distant Supervision in a Max-margin Setting

Rich models with latent variables are popular in many problems in natural language processing. For instance, in IE, one needs to predict the relation labels that an entity-pair can take based on the hidden relation mentions, *i.e.*, the relation labels for occurrences of the entity-pair in a given corpus. These models are often trained by optimizing performance measures (such as conditional log-likelihood or error rate) that are not directly related to the task-specific non-linear performance measure, *e.g.*, the F_1-score. However, better models may be trained by optimizing the task-specific performance measure while allowing latent variables to adapt their values accordingly.

Large-margin methods have been shown to be a

compelling approach to learn rich models detailing the inter-dependencies among the output variables. Some methods optimize loss functions decomposable over the *training instances* (Taskar et al., 2003; Tsochantaridis et al., 2004) compared to others that optimize non-decomposable loss functions (Ranjbar et al., 2013; Tarlow and Zemel, 2012; Rosenfeld et al., 2014; Keshet, 2014). They have also been shown to be powerful when applied to latent variable models when optimizing for decomposable loss functions (Wang and Mori, 2011; Felzenszwalb et al., 2010; Yu and Joachims, 2009).

In this work (Haffari et al., 2015), we describe a novel max-margin learning approach to optimize non-linear performance measures for distantly-supervised relation extraction models. Our approach can be generally used to learn latent variable models under multivariate non-linear performance measures, such as F_β-score.

Our approach involves solving the hard-optimization problem in learning by interleaving Concave-Convex Procedure with dual decomposition. Dual decomposition allowed us to solve the hard sub-problems independently. A key aspect of our approach involves a local-search algorithm that has led to a speed-up of 7,000 times in our experiments over an exhustive search baseline proposed in previous work (Ranjbar et al., 2012; Joachims, 2005).

Our work is the first to make use of max-margin training in distant supervision of relation extraction models. We demonstrate the effectiveness of our proposed method compared to two strong baseline systems which optimize for the error rate and conditional likelihood, including a state-of-the-art system by Hoffmann et al. (2011). On several data conditions, we show that our method outperforms the baseline and results in up to 8.5% improvement in the F_1-score.

4 Conclusion

Our thesis can be summarized as shown in Figure 5. The broad theme of each work along with its publication forum is indicated. In the entity extraction setting, we work in the paradigm of *relational feature space exploration*, and in the relation extraction setting, our research has been in the paradigm

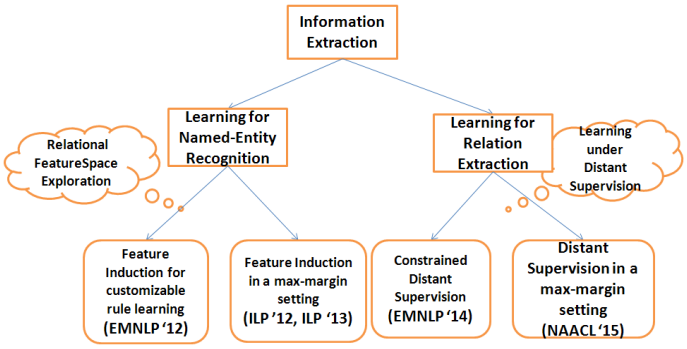

Figure 5: Thesis Summary

of *learning under distant supervision*.

The design of our feature induction approach is aimed at producing accurate rules that can be understood and refined by humans, by placing special emphasis on low complexity and efficient computation of the induced rules. According to our complexity measure, the induced rules have good accuracy and low complexity. While our complexity measure informs the biases in our system and leads to simpler, smaller extractors, it captures extractor interpretability only to a certain extent. Therefore, we believe more work is required to devise a more comprehensive quantitative measure for interpretability. Another interesting direction of future work, is the designing of human-computer interaction experiments, to present the induced rules to a manual rule-developer and evaluating the quality of rules induced.

In the distantly supervised relation extraction, our ILP formulation provides a good framework to add new types of constraints to the problem. In the future, we would like to experiment with other constraints such as modeling the selectional preferences of entity types.

Our max-margin framework for distant supervision provided a way to optimize F_1 score while training the model. Although we solved the hard optimization problem with an efficient dual-decomposition formulation, our algorithms do not scale very well to large datasets. As part of future work, we would like to investigate distributed optimization algorithms as an extension to our solutions. In addition, we would like to maximize other performance measures, such as area under the curve, for information extraction models. We would also

like to explore our approach for other latent variable models in NLP, such as those in machine translation.

References

Mary Elaine Califf. 1998. *Relational Learning Techniques for Natural Language Information Extraction*. Ph.D. thesis, Department of Computer Sciences, University of Texas, Austin, TX, August. Also appears as Artificial Intelligence Laboratory Technical Report AI 98-276.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *EMNLP*.

Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. 2010. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645.

Peter Flach and Nicolas Lachiche. 1999. 1bc: a first-order bayesian classifier. In *Proceedings Of the 9th International workshop on Inductive Logic Programming, Volume 1634 of Lecture Notes in Artificial Intelligence*, pages 92–103. Springer-Verlag.

Gholamreza Haffari, Ajay Nagesh, and Ganesh Ramakrishnan. 2015. Optimizing multivariate performance measures for learning relation extraction models. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - Jun 5, 2015, Denver, Colorado, USA*. The Association for Computational Linguistics.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 541–550, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alfred Horn. 1951. On sentences which are true of direct unions of algebras. *The Journal of Symbolic Logic*, 16(1):pp. 14–21.

T. Joachims. 2005. A support vector method for multivariate performance measures. In *International Conference on Machine Learning (ICML)*, pages 377–384.

Joseph Keshet. 2014. Optimizing the measure of performance in structured prediction. In Sebastian Nowozin, Peter V. Gehler, Jeremy Jancsary, and Christoph H. Lampert, editors, *Advanced Structured Prediction*. The MIT Press.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephen Muggleton and C. Feng. 1992. Efficient induction in logic programs. In *ILP*.

Ajay Nagesh, Ganesh Ramakrishnan, Laura Chiticariu, Rajasekar Krishnamurthy, Ankush Dharkar, and Pushpak Bhattacharyya. 2012. Towards efficient named-entity rule induction for customizability. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 128–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ajay Nagesh, Gholamreza Haffari, and Ganesh Ramakrishnan. 2014. Noisy or-based model for relation extraction using distant supervision. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October. Association for Computational Linguistics.

Naveen Nair, Amrita Saha, Ganesh Ramakrishnan, and Shonali Krishnaswamy. 2012. Rule ensemble learning using hierarchical kernels in structured output spaces. In *AAAI*.

Shan-Hwei Nienhuys-Cheng and Ronald de Wolf. 1997. *Foundations of Inductive Logic Programming*.

Mani Ranjbar, Arash Vahdat, and Greg Mori. 2012. Complex loss optimization via dual decomposition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2304–2311.

Mani Ranjbar, Tian Lan, Yang Wang, Stephen N. Robinovitch, Ze-Nian Li, and Greg Mori. 2013. Optimizing nondecomposable loss functions in structured prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(4):911–924.

Nir Rosenfeld, Ofer Meshi, Amir Globerson, and Daniel Tarlow. 2014. Learning structured models with the auc loss and its generalizations. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

D. Roth and W. Yih. 2001. Propositionalization of relational learning: An information extraction case study. Number UIUCDCS-R-2001-2206.

Sunita Sarawagi. 2008. Information extraction. *Found. Trends databases*, 1(3):261–377, March.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Meth-*

ods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12, pages 455–465, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daniel Tarlow and Richard S Zemel. 2012. Structured output learning with high order loss functions. In *Proceedings of the 15th Conference on Artificial Intelligence and Statistics*.

Benjamin Taskar, Carlos Guestrin, and Daphne Koller. 2003. Max-margin markov networks. In *NIPS*.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 104–, New York, NY, USA. ACM.

Yang Wang and Greg Mori. 2011. Hidden part models for human action recognition: Probabilistic versus max margin. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(7):1310–1323.

Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, page 147.