Research Statement

Veselin Stoyanov

Overview

My research resides at the intersection of Natural Language Processing (NLP) and Machine Learning (ML). I am interested in developing novel methods for structured learning for problems in which interactions between small subsets of items gives rise to global properties. My interest in such models originates from previous experience in opinion analysis and coreference resolution, which are structured and can benefit from rigid models. I plan to apply the models that I develop to NLP problems such as studying the spread of opinions in social networks, performing coreference resolution and extracting information as well as other structured relational learning (SRL) problems such as label propagation and entity linking.

The structured prediction models that I am developing build upon the framework of probabilistic graphical models and extend it in three important ways: (i) provide a principled way of learning model parameters when the model is used with approximations, (ii) allow for the incorporation of priors, semi-supervised and unsupervised learning characteristic of Bayesian learning while maintaining the fast classification runtime characteristic of discriminative approaches, (iii) introduce a principled way of learning when data is missing non-homogeneously. The approaches that I am developing will contribute to the fields of NLP and ML by: (i) creating a new state-of-the-art for structured problems that are currently modeled with PGMs that require approximations, (ii) make principled probabilistic modeling possible for problems that cannot be modeled with current formalisms, and, (iii) make available a general modeling framework and a software implementation for the use of the community.

In the rest of this statement, I will first describe the general framework of structured learning that I am developing and then discuss the three applications that I am planning to explore.

Structured Learning: From Compatibilities to Global Decisions

Structured prediction refers to the task of labeling a set of related outputs with the goal of minimizing a loss function computed over the entire set. There are three important cases in which structured approaches are more desirable than standard supervised learning methods:

1. The loss function may not decompose over individual items in the set making it preferable to pick a globally optimal solution with respect to the loss.

2. By considering the set of output labels as a whole, a learner can model interactions between output labels. In this way, the learner can explore regularities in the output labels.

3. In the statistical relation learning (SRL) setting, a learner is presented with a single training example that contains both the data that is labeled (i.e., the training data) and the data that should be predicted (i.e., the test data). In this case, generalization and prediction can be performed only if inputs and outputs are modeled together. In other words, the SRL setting requires structured learning. Many NLP problems fall in the SRL setting.

Some examples of structured learning tasks include part-of-speech (POS) tagging – predicting a sequence of POS tags given a sequence of words; coreference resolution – deciding what text mentions refer to the same real-world entity, where the output is a set of globally coherent equivalence classes; information extraction – filling a template of structured information from unstructured text, where different mentions of particular information have to be consolidated in a single template; social network modeling – modeling different properties of participants in a social network and their interactions.
More precisely, I am interested in structured learning problems in which the interactions between items can be described in terms of compatibilities between small subsets of items. This setting is fairly general – it includes all of the examples above: POS tagging usually models interactions of each tag with a small subset of surrounding words and tags; coreference resolution systems typically model interactions between pairs on noun phrases (more structured approaches also model coherence of output clusters); information extraction systems consider a small number of contexts triggered by particular lexical items; and in social networks, interactions are described in terms of nodes that are linked together.

The models for structured prediction that I am developing build on the formalism of Probabilistic Graphical Models (PGMs). Next, I briefly describe PGMs and the methods that I am developing to perform structured prediction through the use of PGMs.

Probabilistic Graphical Models. PGMs (such as Bayesian Networks and Markov Random Fields) are tools for representing joint distributions of items (random variables). PGMs are suitable for my problem because they are represented in terms of interactions between subsets of items that give rise to joint distributions over all random variables. Additionally, the PGM formalism enjoys the availability of standard inference\(^1\) and learning algorithms (Koller and Friedman, 2009). General inference in PGMs is intractable, but there exist multiple approximate algorithms that exhibit good performance in practice.

Structured Prediction via Probabilistic Graphical Models. PGMs have been successfully used to model structured prediction problems such as POS tagging, for example. Nevertheless, my work extends the PGMs framework in three important ways:

- **Learning in the presence of approximations.** There exist well-established general training procedure for the cases in which the structure of the problem being modeled admits exact inference and decoding (Lafferty et al., 2001; Koller and Friedman, 2009).\(^2\) When the model requires approximations, however, these assumptions fall apart and the general setting can lead to degenerate solutions (Kulesza and Pereira, 2008).

  We argue that the presence of approximations motivates a different learning objective: one should directly seek the parameters that minimize the empirical risk of the entire system—the PGM together with the approximate inference and decoding procedures. We introduce an algorithm based on backpropagation and gradient-based optimization that can locally optimize the empirical risk for a given loss function, given inference and decoding procedures, in a PGM with arbitrary structure. We applied our algorithm on a range of synthetic-data problems and showed that minimum risk training significantly reduced loss on test data compared to the traditional training of CRFs (Stoyanov et al., 2011). Subsequently, we used our method to model three NLP tasks requiring approximate inference – predicting congressional votes, information extraction from semi-supervised texts and collective multi-label classification and found that minimum risk training significantly outperforms the traditional learning paradigm (Stoyanov and Eisner, 2011).

- **Bayesian/Discriminative Hybrids.** The learning procedure described in the previous bullet can be classified as discriminative learning – given some input \(x\), the goal is to predict an output \(y\) that achieves low loss on the training data set. An alternative view is postulated by Bayesian decision theory – we should impose a prior over the parameter values, model the posterior probability of the parameter values given the training data and pick the output that minimizes expected loss during testing. This view, known as Bayesian learning has three advantages over discriminative learning: it allows us to incorporate prior beliefs about the parameters of the model; it takes into account all uncertainty about the data in the model; and it allows semi-supervised and unsupervised learning to be approached naturally. Bayesian learning, however, requires integration over infinitely many parameters, which is intractable in the general case and computationally expensive to approximate.

---

1. Inference refers to the task of computing probabilities in PGMs. Note that PGMs are models of joint probability. The process of converting probabilities to decisions (e.g., given probabilities of POS tags we want to predict an output tag sequence) is known as decoding.

2. This is the case when the structure of the model is a tree or a graph with low tree-width. Decoding complexity depends on the loss function and is tractable when the function decomposes over output random variables.
I am developing Bayesian/discriminative hybrids that incorporate the advantages of the two learning approaches described above: the relatively fast testing times of discriminative learning and the advantages brought about by the more principled Bayesian approach. Our hybrids assume that we are willing to pay a higher computational cost during training. We rely on Markov Chain Monte Carlo methods such as Gibbs sampling to approximate the integrals required by the Bayesian principle during training. Through sampling, we build a large set of data points that represents an accurate approximation of the Bayesian distribution of the data. We then learn a classifier that minimizes the training loss for the sampled data. The result of this process is a discriminative classifier that achieves the lowest Bayes risk and can be run efficiently during testing. This approach is superior to the standard practice of picking the maximum a posteriori (MAP) parameter values because it takes into account the test loss and the model uncertainty during training.

• Missing Data. In many problems only some aspects of the data are observed. For example, in SRL we might be presented with a single example only parts of which are labeled. Our goal is to learn a set of parameters from the labeled part of the example and use those parameters to predict the unlabeled part of the example. In contrast, in the standard machine learning case data is missing homogeneously – it is assumed that always the same parts of training and test examples are unobserved.

To address the problem of data missing non-homogeneously, I am exploring methods to model the mechanism that censors the data. I am also developing methods that I can use to hide the values of some of the observed random variables and learn to reconstruct them.

In summary, I am developing methods to learn discriminative models by minimizing the risk of the Bayesian distribution while considering mechanisms that create missing data. This methods can accurately estimate parameters that minimize structured risk when the models are used with approximations and data is missing non-homogeneously. So far, we have applied our algorithms to three previously studied NLP applications and achieved state-of-the-art results. Current extensions to the models (bullets two and three) will make our models available to a larger range of NLP problems, including the SRL setting. Our methods will provide a principled way of approaching a number of existing problems with the hope of improving the accuracy of current systems.

Analyzing the Spread of Opinions in Social Networks

Social networks, the graphs of relationships and interactions among a set of individuals, play a fundamental role in understanding the flow of information, ideas and influence. Study of social networks emerged from theoretical interests in modeling social groups, but advances of the Internet and the Semantic Web have brought about unprecedented opportunities for applications that either work directly on explicit social networks or use social network technology to provide practical tools.

Of particular interest in the field of social network analysis is the study of how influence diffuses through the network. Propagation of ideas and influence through a social network has been studied in a number of areas, including medical and technological innovations and in game-theoretic settings. Models for the spread of influence in networks are of practical interest for the political and intelligence domains as well as the study of “word of mouth” marketing.

I plan to study the spread of opinions in social networks as an important component of the propagation of influence. This research will combine the previously discussed methods for structured learning with methods for linguistically modeling opinions. The latter include, among others, algorithms for automatically recognizing topical links between opinions and for summarizing opinions that I have previously developed.

Opinion topics. Modeling opinion topics is important in the context of studying the spread of opinions in social network because we need to understand when two nodes in the network express opinions on the same topic. As part of our work on opinion topic analysis, we provided a new, operational definition of opinion topic in which the topic of an opinion depends on the context in which its associated opinion expression occurs (Stoyanov and Cardie, 2008a; Stoyanov and Cardie, 2008b). Using this definition, we extended an existing fine-grained opinion corpus (the MPQA corpus (Wiebe et al., 2005)) with manual annotations that encode topic information (Stoyanov and Cardie, 2008a). This work also suggested a novel method
for general-purpose opinion topic identification that, following our new definition, treats the problem as an exercise in topic coreference resolution (Stoyanov and Cardie, 2008b). We evaluated our approach using the topic-annotated portion of the MPQA corpus.

**Opinion summaries.** While opinion information as extracted by existing opinion extraction systems (i.e., raw opinion information) is useful, analyzing how opinions spread in a social network requires that the raw information is consolidated in a meaningful way. I will use the term opinion summarization to describe the process of meaningfully aggregating opinions. An opinion summary combines opinions from the same source on the same topic, computes statistics for each source/topic, and aggregates multiple opinions from the same source on the same topic. In previous work we defined two general representations for opinion summaries, identified the problems that need to be addressed by opinion summarization systems, developed methods to address these problems, introduced novel quantitative evaluation metrics for opinion summaries and construct and evaluated full opinion summaries for the documents in a standard opinion-oriented corpus (the aforementioned MPQA corpus) (Stoyanov and Cardie, 2006; Stoyanov and Cardie, 2011).

**Noun Phrase Coreference Resolution**

*Noun phrase coreference resolution* or simply *coreference resolution* is defined as the problem of identifying all noun phrases (NPs) in a document that refer to the same real-world entity. Coreference resolution is an important NLP task and represents the kind of structured problem that can benefit from jointly modeling decisions concerning all NPs. For example, deciding that both the pronoun “he” as well as the mention “Hillary Clinton” co-refer with a mention of “Clinton” leads to a globally undesirable assignment that can be avoided by structured prediction.

In previous work, I initiated an effort that created a coreference resolution research platform, which: (i) implements the basic underlying architecture of contemporary coreference resolution systems, (ii) includes several state-of-the-art coreference algorithms, (iii) can run on the majority of the standard coreference resolution data sets, (iv) implements most standard coreference resolution scoring algorithms and, (v) is relatively fast, easy to configure and run and can be extended with new methods and features with little effort. With our system, Reconcile, in hand, we were then able to study the state of the art along three axes. First, we examined three subproblems that play a role in coreference resolution: named entity recognition, anaphoricity determination, and coreference element detection. We confirmed that certain assumptions regarding these subproblems in the evaluation methodology can dramatically simplify the overall task (Stoyanov et al., 2009). Second, we measured the performance of a state-of-the-art coreference resolver on several classes of anaphora and used these results to develop a quantitative measure for estimating coreference resolution performance on new data sets. Our coreference resolution platform, Reconcile, has been released and is being used by the research community.

The structured learning approaches that I am developing are directly applicable to the coreference resolution problem. Using these methods for coreference resolution will serve a dual purpose – validate the methods for structured learning and bring about new principled and more accurate approaches to model coreference resolution.

**Information Extraction**

Information extraction (IE) is the problem of automatically populating scenario templates for a given event or entity type. For example, an IE application targeting disasters may have to detect mentions of disasters and populate a template including fields such as the date, time, location and the damage caused by the disaster. Information extraction often deals with consolidating information from multiple parts of a document or from multiple documents. The information can be repeated, conflicting and/or incomplete.

My interest in information extraction is motivated by two reasons: First, I believe that IE is an important NLP task because the advent of the Semantic Web presents an unprecedented opportunity to develop IE approaches with immediate practical impact. Second, information extraction provides opportunities for structured learning and a natural testbed for the structured learning approaches that we are developing.
References


