RESEARCH STATEMENT – Roi Reichart

(roiri@cs.huji.ac.il, roiri@csail.mit.edu, www.cs.huji.ac.il/∼roiri, http://people.csail.mit.edu/roiri)

Natural Language processing (NLP) is a field that combines linguistics, cognitive science, statistical machine learning and other computer science areas in order to compile intelligent computer systems that can understand human languages. NLP has various applications, among which are machine translation, question answering and search engines.

The field of NLP has, in the past two decades, come to simultaneously rely on and challenge the field of machine learning. Statistical methods now dominate NLP, and have moved the field forward substantially, opening up new possibilities for the exploitation of data in developing NLP components and applications.

Many state of the art natural language algorithms are based on supervised learning techniques. In this type of learning, a corpus consisting of texts annotated by human experts is compiled and used to train a learning algorithm. While supervised learning has made substantial contribution to NLP, it faces some significant challenges.

Many fundamental NLP tasks, such as syntactic parsing, part-of-speech (POS) tagging and machine translation, involve structured prediction and sequential labeling. For tasks of this nature, an annotation bottleneck exists - compiling annotated corpora is costly and error prone due to the complex nature of annotation.

A closely related challenge is that of *domain adaptation*. Supervised algorithms usually perform well when the training data and the data for which they should provide predictions (the test data) are drawn from similar domains. When an algorithm trained with data from one domain is to provide predictions for data taken from a substantially different domain, its performance markedly degrades. Again,the burden involved in manually annotating data dims it unfeasible to create sufficient examples for a wide range of domains.

Supervised natural language learning is also challenged by methodological problems. Annotation schemes for tasks such as syntactic parsing are often based on arbitrary decisions. Such schemes often provide a detailed description of certain structures while addressing others only briefly. Many applications would benefit from different annotation decisions.

My research focuses on developing machine learning techniques which deal with these challenges. Its main theme is utilizing the plentiful amounts of raw text available nowadays, for creating state of the art algorithms that use little to no manually annotated data. Cases where little amounts of manually annotated text is utilized are referred to as *semi-supervised learning* [1]; while cases where only raw text is used are known as *unsupervised learning* [2, 3]. I have explored semi-supervised and unsupervised techniques for NLP tasks which involve structured prediction and sequence labeling, mainly syntactic parsing and POS tagging.

Combining Labeled and Unlabeled Data: Semi-Supervised Learning

My research of semi-supervised learning has followed four main directions:

1. Utilizing raw text to enhance the performance of syntactic parsers when only small amounts of labeled data are available [4, 14]. When parsers are used by applications in the highly variable web, this is the rule rather than the exception.

2. Selecting high quality text samples for statistical parsers training, whose annotation requires minimal human effort. Active learning [5] has traditionally been used for this task [6]. My

research focuses on finding both reliable measures for the human annotation efforts and developing algorithms for selecting the samples [7, 8].

3. Developing reliable confidence scores for the output of statistical parsers [9, 10]. The output of statistical parsers is often used by NLP applications such as question answering and machine translation. A confidence score can provide these applications with an indication about the effect that a given syntactic analysis will have on their performance. In addition, confidence scores play an important role in many semi-supervised learning protocols [11].

4. A main theme of my post-doctoral research is developing structured prediction algorithms, based on approximate inference techniques, that enable high quality semi-supervised learning through joint learning and inference. Sharing information between test sentences through joint inference and learning several highly-correlated tasks together can lead to significant performance improvement when training data is sparse.

For the first direction, various semi-supervised techniques have been explored throughout the years. Among these are active-learning, co-training and ensemble methods. Our work was the first to show that self-training, a technique where the parser is trained on its own output, can substantially enhance parser performance in the small annotated training data case. The self-training protocol we proposed, currently achieves the best results for this scenario.

For the second direction, we have addressed two scenarios that have not been explored in the literature before. One case is where a corpus is to be built for multiple linguistics annotation schemes. In our work we developed algorithms for simultaneous annotation with syntactic trees and name entities.

The other case is that of building a syntactically annotated corpus where the cognitive effort of the human annotator is to be minimized while the corpus quality as training material for statistical parsers is to be maximized. In this work we proposed measures for the human cognitive efforts in syntactically analyzing a sentence and developed the algorithms accordingly. Our algorithm has been shown to have better performance on the task when compared to previously proposed algorithms. The paper that summarizes this research received the **best paper award** in CoNLL 2009.

For the third direction, we developed state of the art confidence measures for both supervised and unsupervised parsers. In the supervised case the measure is based on an ensemble technique while in the unsupervised case it utilizes a cluster-based tree representation. In both cases, the proposed score has been shown to outperform other existing scores.

In addition to its value for applications, the score can indicate if the tree induced by the parser can be used as training data for supervised and unsupervised parsers. This potential application of quality scores is especially important in the unsupervised case, since high quality trees produced by unsupervised algorithms can provide training material for supervised parsers with no human efforts. Our results indicate that the unsupervised quality score can be used in a way that substantially improves the performance of both unsupervised syntactic parsers [12] and unsupervised semantic parsers [13].

For the last direction, the new dual decomposition algorithm we presented [14] can be integrated to any of-the-shelf sentence-level algorithm and make inference over a set of sentences possible. Such inference takes advantage of the regularities of linguistic structure and has been shown to significantly improve the performance of state-of-the-art dependency parsers and part-of-speech taggers when small amounts of labeled training data are available. In my current research I have

extended my efforts to the field of information extraction. We have developed a joint learning algorithm for the tasks of event segmentation (inferring which event each part of the text describes) and word tagging (annotating each word in the text with labels that describe the information it provides regarding the event it describes). We intend to publish our results in the next major NLP conference.

Language Acquisition from Plain Text: Unsupervised Learning

In the unsupervised direction, my research has focused on sequential clustering tasks such as POS tagging and the induction of labeled parse trees. We have explored both algorithmic approaches and the unique evaluation challenges of the field.

Labels induced by unsupervised algorithms receive arbitrary names. Evaluating their quality against a gold standard therefore requires a correspondence scheme between the induced and gold labels. We have explored mapping-based and information-theoretic-based approaches for this challenge and analyzed their properties [15, 16].

For the computation of mapping-based evaluation schemes we explored graph-based optimization methods such as the Kuhn-Munkres algorithm. For the information-theoretic evaluation schemes [17] we developed measures based on conditional and unconditional entropies and explored the aspects of the similarity between the induced and the gold standard clustering they reflect. We used these measures to analyze existing algorithmic approaches to the aforementioned sequential clustering problems, and to suggest novel techniques.

A property shared by most algorithmic approaches to the aforementioned tasks is that the optimal clustering is defined to be the one that optimizes a non-convex function of the data. Optimization algorithms for such functions are well-known to converge to local maxima which yield clustering solutions of variable quality. Finding a final, high quality solution is a difficult problem that has hardly been addressed in the NLP literature. We developed a method for identifying the high quality runs of unsupervised POS induction algorithms [18].

We proposed various algorithmic techniques which deal with this problem. For the induction of labeled parse trees we developed clustering techniques based on a Bayesian minimum description length (MDL) principle [20] and on ensembles of Dirichlet Process mixture model experts. For POS induction we developed both a deterministic algorithm based on linguistics distributional and morphological considerations and a method that imposes existing stochastic methods to converge to a high quality solution [19, 18]. All these methods output a high quality, deterministic final clustering (best reported results for the addressed problems at the time the papers were published and still very competitive today).

My research on semi-supervised and unsupervised learning follows a few characteristics guidelines. It explores methods for the application of unlabeled text in a statistically robust manner and aims at developing evaluation measures that reflect the various statistical and cognitive dimensions of natural language acquisition.

Research Agenda: Directions for Future Work

The NLP research community has only taken its first steps in its journey towards building computer algorithms that are able to acquire human languages from plain text without the guidance of a human expert.

On the algorithmic level, efficient algorithms that are based on sophisticated statistical models and can be applied to plentiful amounts of raw text are to be developed. In the future I intend to devote a significant portion of my research efforts to the development of formal statistical models and of efficient inference and learning techniques with formal guarantees for these models. Recently I extended my research to deal with an especially challenging problem that requires such algorithms – the problem of joint learning and inference.

Almost any system for natural language understanding must recover linguistic structure at many different levels: parts of speech, syntactic dependencies and phrases, named entities, and so on. For example, modern question answering systems require question type classification, parsing, named entity tagging, semantic role labeling, and often other tasks, many of which are dependent on one another and must be pipelined together. Pipelined systems are ubiquitous in NLP: in addition to the above examples, parsers and named entity recognizers commonly use part of speech tags and chunking information, and also word segmentation. A major challenge in modern NLP research is how to perform such tasks jointly to avoid the cascading of errors from one application to another.

In a recent paper [14] we have proposed a novel algorithm for novel corpus-level inference that was demonstrated to improve the performance of state of the art models for syntactic parsing and part-of-speech tagging. In my current research we develop a novel algorithm for joint learning and inference in information extraction. In the future I intend to extend these research efforts.

Exploring the connection between the different linguistic annotation layers should not be limited to joint learning and inference. In my research I explored the application of syntactic acquisition algorithms to semantic acquisition tasks such as ontology induction [21], verb argument identification [22], learning the syntactic/semantic frame of a verb [23] and understanding the role of commas in sentences [24]. This research can potentially cast light on the role of syntax in defining the semantic meaning of sentences. Moreover, it can yield novel measures for the output quality of syntactic acquisition algorithms based on the value of their output for semantic acquisition. Deeper understanding of the connection between syntax and semantics is an important topic for future research.

On the linguistic side, our understating of what are human languages is to be enhanced. One step in this direction is using unsupervised methods such as those developed in my research for the design of data-driven tag sets for NLP tasks, hopefully reflecting the nature of the data better than manually designed tag sets.

My PhD and post-doc research have lead to a large number of papers that were published in the major international NLP conferences (ACL, EMNLP, COLING and CoNLL). In CoNLL 2009 one of these papers received the **best paper award**.

Most of my papers are the fruit of a cooperation with at least one co-author (that is, in addition to my PhD and post-doc advisers, or other senior Professors with whom I have worked), and the total number of co-authors with which I cooperated is 11 (From Israel,US and Germany). Cooperating with other people is a fundamental aspect of my research style. It enhances my work and enables me to face more difficult challenges and address more complex issues. I also find it to be a very enjoyable way of work.

References

[1] Steven Abney, "Semisupervised Learning for Computational Linguistics.", *Chapman and HAll, 2007.*

[2] Dan Klein, "The unsupervised learning of natural language structure.", *PhD Thesis, Stanford University, 2005.*

[3] Alexander Clark, "Unsupervised Language Acquisition: Theory and Practice.", *PhD Thesis, University of Sussex, 2001.*

[4] **Roi Reichart** and Ari Rappoport, "Self-Training for Enhancement and Domain Adaptation of Statistical Parsers Trained on Small Datasets.", *ACL '07.*

[5] David Cohn, Les Atlas and Richard Ladner. "Improving generalization with active learning.", *Machine Learning*, 15(2):201–221.

[6] Rebecca Hwa, "Sample selection for statistical parsing.", *Computational Linguistics*, 30(3):253–276.

[7] **Roi Reichart**, Katrin Tomanek, Udo Hahn and Ari Rappoport, "Multi-Task Active Learning for Linguistic Annotations.", *ACL '08.*

[8] **Roi Reichart** and Ari Rappoport, "Sample Selection for Statistical Parsers: Cognitively Driven Algorithms and Evaluation Measures.", *CoNLL '09.*

[9] **Roi Reichart** and Ari Rappoport, "An Ensemble Method for Selection of High Quality Parses.", *ACL '07.*

[10] **Roi Reichart** and Ari Rappoport, "Automatic Selection of High Quality Parses Created By a Fully Unsupervised Parser.", *CoNLL '09.*

[11] Avrim Blum and Tom Mitchell, "Combining labeled and unlabeled data with co-training.", *COLT '98.*

[12] **Roi Reichart** and Ari Rappoport, "Improved Fully Unsupervised Parsing With Zoomed Learning.", *EMNLP '10.*

[13] Dan Goldwasser, **Roi Reichart**, James Clark and Dan Roth, "Confidence Driven Unsupervised Semantic Parsing.", ACL 2011.

[14] Alexander Rush, **Roi Reichart**, Michael Collins and Amir Globerson, "Efficient and Exact Inter-Sentence Decoding for Natural Language Processing.", Submitted.

[15] **Roi Reichart** and Ari Rappoport, "The NVI Clustering Evaluation Measure.", *CoNLL '09.*

[16] **Roi Reichart**, Omri Abend and Ari Rappoport, "Type Level Clustering Evaluation: New Measures and a POS Induction Case Study.", *CoNLL '10.*

[17] Marina Meila, "Comparing clustering – an information based distance.", *Journal of Multivariate Analysis*, 98:873–895.

[18] **Roi Reichart**, Raanan Fattal and Ari Rappoport, "Improved Unsupervised POS Induction Using Intrinsic Clustering Quality and a Zipfian Constraint.", *CoNLL '10.*

[19] Omri Abend, **Roi Reichart** and Ari Rappoport, "Improved Unsupervised POS Induction through Prototype Discovery.", *ACL '10.*

[20] **Roi Reichart** and Ari Rappoport, "Unsupervised Induction of Labeled Parse Trees by Clustering with Syntactic Features.", *COLING '08.*

[21] Dmitry Davidov, **Roi Reichart** and Ari Rappoport, "Superior and Efficient Fully Unsupervised Pattern-based Concept Acquisition Using an Unsupervised Parser.", *CoNLL '09.*

[22] Omri Abend, **Roi Reichart** and Ari Rappoport, "Unsupervised Argument Identification for Semantic Role Labeling.", *ACL '09.*

[23] Omri Abend, **Roi Reichart** and Ari Rappoport, "A Supervised Algorithm for Verb Disambiguation into Verb-Net Classes.", *COLING '08.*

[24] Vivek Srikumar, **Roi Reichart**, Mark Sammons, Ari Rappoport and Dan Roth, "Extraction of Entailed Semantic Relations Through Syntax-based Comma Resolution.", *ACL '08.*