Introduction to the Special Issue on *Natural Language and Learning Machines*

Dan Roth*
University of Pennsylvania.

Roberto Basili**
Università di Roma, Tor Vergata

1. Introduction

The interaction between machine learning and natural language processing (NLP) research underlies most of the progress made in NLP for the last few decades (Cardie and Mooney 1999; Fung and Roth 2005). Machine Learning has been the common framework for the birth and development of most paradigms, discoveries and achievements in statistical natural language processing. At the international level the AAAI Fall symposiums in 1990 (Jacobs 1990) and 1992 (Goldman 1992) and the IBM TJ Watson paper on statistical Machine Translation (Brown et al. 1988) established firm roots for the use of Bayesian modeling and data-driven algorithms for complex computational linguistic tasks. At that time several Italian research groups were already working on machine learning methods for tasks such as natural language parsing and lexical acquisition. A relevant event was the Workshop *Apprendimento Automatico e Linguaggio Naturale* organized at the University of Torino, whose decisive inspiration was contributed by Leonardo Lesmo and Piero Torasso that pioneered NLP research in Italy ((Lesmo 1997)). One of the topics at the workshop was "Are syntactic representations and parsing still central in current NLP and Information Extraction tasks, given the role that shallow features combined with complex learning algorithms play in achieving significant results over several benchmarks?". As we know, some of these issues and challenges are still relevant today, and these questions still trigger many empirical studies and debates from heterogeneous intellectual positions.

In current research, the aforementioned issues are still open research issues, possibly formulated using a different jargon. Are parsing algorithms still relevant given the growing success demonstrated by recurrent neural networks in tasks that were believed to require parsing? Are linguistic aspects of the problem (e.g. traditional categories such as root vs. lemma distinctions, agreement or verbal aspects) still important given the ability to induce intermediate representations that seems to capture these notions? At the same time one needs to consider ways in which neural networks are currently being trained, mostly counting on vast amounts of task specific annotated data and, consequently, the generality of the representations thus induced.

2. Learning and Language Processing

It has been clearly shown that, in general, (natural language) inference can be formulated as a joint constrained optimization task done over learned components (Roth and

* Department of Computer and Information Science, University of Pennsylvania.
E-mail: danroth@seas.upenn.edu
** Dept. of Enterprise Engineering - Via del Politecnico 1, 00133 Rome, Italy.
E-mail: basili@info.uniroma2.it

Yih 2004, 2007; Chang, Ratinov, and Roth 2012). By “Inference" we refer here to the assignment of values to a collection of interdependent variables. The “optimal" decision model can then be arrived at by satisfying a set of constraints imposed on the final assignment of values to variables, the output decision. For example, the satisfaction of some constraints on the distribution of semantic roles can be jointly optimized with the interpretation of the reference target predicate. As a consequence, learning here corresponds to the learning of (the coefficients of) an objective function that combines a target function ψ (the cost of the proper assignment of values y to variables $\mathcal{Y}$) to be minimized together with the set of theory-driven constraints C, whose individual violations tend to increase the cost, i.e.:

$$y = \text{argmin}_{y \in \mathcal{Y}} w^T \phi(x, y) + u^T C(x, y) \quad (1)$$

where w and u are weights matrices to be learned through annotated data (Roth and Yih 2004; Chang, Ratinov, and Roth 2012), either jointly, or in a decomposed fashion. While $w^T \phi(x, y)$ thus correspond to the decision that the (data-driven) linguistic inference must produce, e.g. semantic role labels y for the individual word sequence x, the $u^T C(x, y)$ component constrains any choice of $y \in \mathcal{Y}$: it is thus helpful in judging the quality of alternative solutions y and ranks them. Joint optimization allows learning to proceed (i.e. carry out the labeling) by maximizing the satisfaction of all constraints.

The above general setting is important in NLP for multiple reasons:

- The decision function corresponds in a more or less direct way to a complex linguistic inference whose nature is in general semantic: it makes a bridge between the observable linguistic symbols x and the operational context (i.e. the world) in which the decision is immersed. For example, the joint assignment of predicate and roles to the incoming sentence x in semantic role labeling.
- The constraints C can be used to express linguistic principles, that embody forms of agreement that natural languages must convey between speakers and hearers. It reflects the expectations one has from an interpretation y of x, whatever the current natural language decision problem (x, y) is. In natural language such agreement is a strongly social phenomena, established across time and possibly through repeated attempts. Constraint optimization just expresses this approximation process.
- Some of the constraints might be derived from the reference world, where properties usually correspond to sound (although simple) theories. These model semantic aspects as well as other formal properties of the decision, are obtained to satisfy external (e.g. domain) knowledge.

The power of natural language results from its variability and its ambiguity. This is also what makes it a highly subjective phenomenon and makes it difficult to process and understand automatically. As is the experience of human subjects, we can say that subjectivity is the ontological status of natural language practices. However, we can foster an objective epistemology of even such highly subjective phenomenon. Machine learning is crucial in this sense. We can say that the increasing success of machine learning in NLP stands as a proof that an epistemologically objective approach to natural language is possible. Machine learning and its mathematics provides sound modeling tools for a vague problem. The constraint optimization model expressed by

Eq. 1 or the convergence properties of the learning algorithms used to model data-driven decisions based on the risk minimization principle are just examples of this contribution. Linguistic inference (as an incremental and iterative agreement process between speakers and hearers) is more easily mapped into a learning (and inference) process that resembles the nature of the language acquisition process. In other words, machine learning, as the ability of machines to develop decision functions, out from examples, and from (being told) constraints, seems a nice way to characterize language processing capabilities as those emerging from linguistic practices.

All the papers collected in this special issue follow, in a more or less tight fashion, the above mathematical setting, although under the umbrella of alternative paradigms, such as deep learning or distributional semantic analysis: they all make strong use of linguistic constraints to control the reference machine learning model. The variety of the tasks and the ways linguistic principles are adopted in the representational hypothesis and in the architectures proposed show the richness of methodologies and open aspects that still inspire research on machine learning for NLP.

3. Overview of the Issue

The first paper by Madotto and Attardi presents a neural network architecture for two tasks, *Reasoning Question Answering* and *Reading Comprehension*. Memory Networks (Weston, Chopra, and Bordes 2014) are employed in order to recognize entities and their relations to answers in a target text. A focus attention mechanism and an independent memory is adopted as an extension of a Recurrent Neural Network. The proposed model, Question Dependent Recurrent Entity Network (QDREN), exploits information and properties of the question during the memorization process and uses them to decide the correctness of one or more proposed answers. The extended network architecture is evaluated on synthetic as well as real datasets with improved accuracy levels and competitive results in both tasks.

In the paper by Passaro and colleagues, a corpus-driven approach to the acquisition of the lexical affective values used in sentiment analysis systems is presented. The acquisition of emotive embeddings for lexical items is realized by co-occurrence analysis with negative expressions. The proposed distributional semantic analysis is a form of bootstrapping for emotional lexicons, built around eight basic emotion categories. In this way, the authors show how to use positive vs. negative lexical valences to model behavioral data.

In the paper by Basile and colleagues presents a complex Deep Learning architecture for the joint learning of several Natural Language Processing tasks for Italian. The architecture is based on state of the art models and exploits both word-level and character-level representations through the integration of Long Short Term Memory (LSTM) networks, Convolutional Neural Networks (CNN) as well as Conditional Random Fields (CRF). The architecture, that provided state of the art performance in several sequence labeling tasks on English datasets, is applied to the Italian language with a multi-task learning paradigm, in particular, targeting PoS-tagging and sentiment analysis. State of the art performance is shown in all the tasks.

In the paper by Bonadiman and colleagues a deep neural network (DNN) for multi-task learning as applied to (three tasks in) the community Question Answering (cQA) process[1]. The latter task, i.e. the new question-old comment similarity estimation, is

1 The CQA process targeted is equivalent to the one proposed in the SemEval-2016 Task 3, i.e., question-comment similarity, question-question similarity and new question-comment similarity.

the task where multi-task learning provides the best contribution. The proposed DNN is jointly trained on all the three cQA tasks and avoids any use of manually designed features and it is shown to approach the state of the art established with methods that make heavy use of feature engineering. It learns to encode questions and comments into a single vector representation shared across the multiple tasks. The results on the official test sets show that the integrated neural network produces higher accuracy and faster convergence rates than the individual one.

... a Closing Remark

The collection of papers in this special issue provides further evidence for the need for stronger and often task specific representations to benefit machine learning for natural language processing. In all papers, complex architectures are obtained either by integrating different learning tasks in one joint training stage or by extending existing architectures. Example of the first approach are the multi-task learning of the individual community Question Answering subproblems in the Bonadiman paper or the joint multi-task learning proposed in Basile and colleagues for POS tagging and sentiment analysis. An example of the second is obtained through the memorization of the input question integrated with multiple sentence embeddings, as proposed by Madotto et al. in the QDREN architecture proposed for Reasoning Question Answering.

The interesting results collected here seem to be all moving in one general direction: the combination of local, i.e. task specific, evidence with general constraints usually derived from a theory of the target linguistic phenomena. As Equation 1 seems to definitively suggest, local (i.e. example specific) constraints should always be combined with theory-driven or expectation-driven constraints (e.g. the attempt to satisfy relational associations between a question and its reference input text). Language studies and linguistic principles thus still seem to have a relevant role in the research towards learning machines that address intelligence.

References

Brown, Peter F., John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, Robert L. Mercer, and Paul S. Roossin. 1988. A statistical approach to language translation. In *Proceedings of the 12th International Conference on Computational Linguistics, COLING '88, Budapest, Hungary, August 22-27, 1988*, pages 71–76.

Cardie, Claire and Raymond J. Mooney. 1999. Guest editors; introduction: Machine learning and natural language. *Mach. Learn.*, 34(1-3):5–9, February.

Chang, Ming-Wei, Lev Ratinov, and Dan Roth. 2012. Structured learning with constrained conditional models. *Machine Learning*.

Fung, Pascale and Dan Roth. 2005. Guest editors introduction: Machine learning in speech and language technologies. *Machine Learning*, 60(1-3):5–9.

Goldman, Robert. 1992. *Working notes of the 1992 AAAI Fall Symposium "Intelligent Probabilistic Approaches to Natural Language"*. AAAI Press, Menlo Parl, California, USA.

Jacobs, P.S:. 1990. *Working notes, 1990 AAAI Spring Symposium on Text-Based Intelligent Systems, appeared as Text-based intelligent systems: current research in text analysis, information extraction, and retrieval, Report 90CRD198*. General Electric R. & D. Centre, Schenectady, NY.

Lesmo, Leonardo. 1997. *Incontro dei Gruppi di Lavoro dell'Associazione Italiana per l'Intelligenza Artificiale su Apprendimento Automatico e Linguaggio Naturale*. Universita' di Torino, Torino.

Roth, D. and W. Yih. 2007. Global inference for entity and relation identification via a linear programming formulation. *Introduction to Statistical Relational Learning. Editors: Lise Getoor and Ben Taskar, 2007.*

Roth, Dan and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *CoNLL 2004*, pages 1–8.

Weston, Jason, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *CoRR*, abs/1410.3916.