Natural Language Engineering **18** (2): 147–153. © Cambridge University Press 2012
doi:10.1017/S135132491200006X

Special issue on statistical learning of natural language structured input and output

LLUÍS MÀRQUEZ[1] and
ALESSANDRO MOSCHITTI[2]
[1] *TALP Research Center, Technical University of Catalonia, Barcelona, Spain*
e-mail: lluism@lsi.upc.edu
[2] *University of Trento, Povo (TN), Italy*
e-mail: moschitti@disi.unitn.it

Abstract

During last decade, machine learning and, in particular, statistical approaches have become more and more important for research in Natural Language Processing (NLP) and Computational Linguistics. Nowadays, most stakeholders of the field use machine learning, as it can significantly enhance both system design and performance. However, machine learning requires careful parameter tuning and feature engineering for representing language phenomena. The latter becomes more complex when the system input/output data is structured, since the designer has both to (i) engineer features for representing structure and model interdependent layers of information, which is usually a non-trivial task; and (ii) generate a structured output using classifiers, which, in their original form, were developed only for classification or regression. Research in empirical NLP has been tackling this problem by constructing output structures as a combination of the predictions of independent local classifiers, eventually applying post-processing heuristics to correct incompatible outputs by enforcing global properties. More recently, some advances of the statistical learning theory, namely structured output spaces and kernel methods, have brought techniques for directly encoding dependencies between data items in a learning algorithm that performs global optimization. Within this framework, this special issue aims at studying, comparing, and reconciling the typical domain/task-specific NLP approaches to structured data with the most advanced machine learning methods. In particular, the selected papers analyze the use of diverse structured input/output approaches, ranging from re-ranking to joint constraint-based global models, for diverse natural language tasks, i.e., document ranking, syntactic parsing, sequence supertagging, and relation extraction between terms and entities. Overall, the experience with this special issue shows that, although a definitive unifying theory for encoding and generating structured information in NLP applications is still far from being shaped, some interesting and effective *best practice* can be defined to guide practitioners in modeling their own natural language application on complex data.

1 Introduction to the special issue

Machine learning, together with techniques borrowed from other related fields, such as statistics and optimization, has become indispensable for large part of

https://doi.org/10.1017/S135132491200006X Published online by Cambridge University Press

Computational Linguistics and Natural Language Processing (NLP) research and applications. On the one hand, these techniques have enhanced systems' performance and have significantly sped up some development and knowledge-acquisition phases. On the other hand, their use requires careful parameter tuning and, above all, engineering of machine-based representations of natural language phenomena, e.g., by means of features that sometimes detach from the common sense interpretation of such phenomena. These difficulties become more marked when the input/output data have a structured and relational form: The designer has both to engineer features for representing the system input, e.g., the syntactic parse-tree of a sentence, and devise methods for generating the output, e.g., by building a set of classifiers that provide boundaries and types (semantic argument, function, or concept type) of some of the parse-tree constituents.

Research in empirical NLP has been tackling these complexities since the early work in the field. For instance, word sense tagging is a problem in which the input, i.e., word sequences, and output, i.e., sense tag sequences, are structured. However, the designed models were mainly based on local information, e.g., the model in Yarowsky (1992) for word sense disambiguation. Other approaches attempted to improve locality by encoding rules capturing more global information. For example, Brill's POS tagger (Brill 1992) learned and applied correcting rules expressing some global view of the sequence. More recently, the idea of *global model* is encoded by post-processing heuristics, which aim at correcting incompatible output – see some semantic role-labeling systems in Carreras and Màrquez (2005). The use of such ad hoc solutions was mainly due to the lack of statistical and machine learning theory suggesting how models should be designed for capturing dependencies among the items in the structured data.

In contrast, recent work in machine learning has provided different paradigms to globally represent and process such data. The key ingredients of all these approaches are as follows: (i) They define how to score globally the appropriateness of complete output candidate structures (or *hypotheses*) for a given input; (ii) learning is performed to optimize a global objective function; (iii) features used to score the hypotheses can encode any aspect of the input and output structures (usually with some limitations on the output); and (iv) a search procedure (also known as *decoding* or *inference*) is defined over the space of possible output structures for finding the one with the highest score. For instance, we may find the following:

- *Graphical models*, which encode observations and labels as nodes of graphs, where the edges define dependencies among them, weighted with probabilities. Inference algorithms are defined over these graphs in order to find the most probable assignment to variables (i.e., defining the output structure) for a given input example. Learning consists in acquiring the weights of all model parameters. Conditional Random Fields (CRFs) are one of the most representative examples of such models (Lafferty, McCallum and Pereira 2001).

https://doi.org/10.1017/S135132491200006X Published online by Cambridge University Press

- *Non-probabilistic discriminative approaches*, including global linear models such as the simple structured perceptron (Collins and Duffy 2002) and other more sophisticated Max-Margin learning algorithms, such as SVMstruct (Tsochantaridis *et al.* 2005) and Max-Margin Markov Networks (Taskar, Guestrin and Koller 2004), which are extensions of the well-known Support Vector Machines (SVMs) (Vapnik 1998) into the structure prediction scenario. Interestingly, all these methods are instances of kernel machines. By using structural kernels (e.g., convolution kernels) one can encode complex input data, such as the examples appearing in NLP problems (Shawe-Taylor and Cristianini 2004).

However, the learning paradigms above may have drawbacks related to complexity and efficiency, which can limit their use in practical situations. For instance, parameter estimation and exact decoding can be computationally prohibitive for some real world models and applications.

In parallel with the previous research, alternative approaches have been studied. In particular, *constrained conditional models* and *re-ranking*. In the former the predictions of several locally trained classifiers are combined to score complete candidate solutions for a given input. Integer Linear Programming is usually used to efficiently find the best scored global solution under a set of structural and problem-specific constraints (e.g., Punyakanok *et al.* 2005; Chang *et al.* 2008). The latter are based on the efficient coupling of basic local models for generating lists of hypotheses for every input and classifiers (re-rankers), which are used to select the best output structure from the list of hypotheses. Noticeably, such classifiers can exploit the global information (the dependencies among observations and labels) available in the hypotheses to learn the ranking task.

In summary, on the one hand, there is an interesting theory that needs to be further explored both to improve accuracy and efficiency of the algorithms. On the other hand, there are more practical approaches, based on simple classifiers, which are typically very fast and can achieve state-of-the-art results when accurately designed for specific tasks (Collins 2000; Shen, Sarkar and Och 2004; Carreras and Màrquez 2005; Charniak and Johnson 2005; Koo and Collins 2005; Kudo, Suzuki and Isozaki 2005; Ge and Mooney 2006; Roark *et al.* 2006; Moschitti *et al.* 2007; Huang 2008; Punyakanok, Roth and Yih 2008; Dinarelli, Moschitti and Riccardi 2009; Nguyen, Moschitti and Riccardi 2009).

This special issue aimed at collecting and presenting extensions, findings (both empirical and theoretical), and new directions within the framework of the above research. In addition, we were also interested in the comparison of different learning paradigms for dealing with structured NLP data, as well as in the adaptation or generalization of already existing methods to novel NLP tasks or usages. As we will see in the following section, some of these aspects are effectively addressed by the papers presented in this special issue.

https://doi.org/10.1017/S135132491200006X Published online by Cambridge University Press

2 Content overview

The special issue call[1] received thirteen high-quality papers, out of which only five (for space reasons) could be included in the journal. Most of the excluded papers were recommended for publication in the *Journal of Natural Language Engineering* following its standard editorial process. Interestingly, the special issue call was coordinated with the TextGraphs 2011 workshop,[2] which also included a special theme on 'Graphs in Structured Input/Output Learning'. Out of the five accepted papers, two of them are expanded versions of papers previously published at TextGraphs. These two were selected as the two best contributions from the workshop, according to the special issue topic, and invited to submit to the special issue. The acceptance process consisted of two-review rounds,[3] in which every paper was reviewed by three members of the guest editorial board, and one more round for preparing the camera-ready versions. The five papers selected by this highly accurate review process, address various aspects of NLP and machine learning. Document ranking, syntactic parsing and sequence labeling, and relation extraction between terms and entities were tackled with different structured input/output approaches. These ranged from re-ranking to joint constraint-based global models as described in what follows.

In the first paper, Villatoro *et al.* introduce a novel method to re-rank the list of documents returned by an Information Retrieval (IR) system. This approach is based on a Markov Random Field (MRF) model, which classifies the retrieved documents as relevant or irrelevant. The proposed MRF combines (i) information provided by the base IR system, (ii) similarities among documents in the retrieved list, and (iii) relevance feedback. The experiments show that the joint model can improve on the mean average precision used by state-of-the-art retrieval engines.

In the second paper, Søgaard illustrates an innovative approach to unsupervised dependency parsing. This is based on the assumption that a dependency structure is also a partial order on the nodes in terms of centrality or saliency. By exploiting this property, the author sorts input words by means of a ranking model and obtains the dependency parse. For this purpose, a simple deterministic parsing algorithm is applied, which relies on the universal dependency rules defined by Naseem *et al.* (2010). The resulting unsupervised parser is shown to be competitive to state-of-the-art unsupervised dependency parsers in a variety of languages.

In the third paper, Zhang *et al.* approach a rich sequence-labeling task, namely, supertagging by exploiting, in addition to the standard adjacent label features, long-range information. Capturing these long-distance dependencies proves to be very important for the task. Authors propose word-to-word dependencies derived from a dependency parser and long-range features encoded as soft constraints

[1] See http://disi.unitn.it/moschitti/NLE-learning-of-NL-structured-io.html for more details.

[2] http://www.textgraphs.org/ws11

[3] Regarding the TextGraphs papers, the first round was considered to be that of the workshop, with an additional review by the guest editors. The second round of reviewing was run entirely by the special issue. This synergy was facilitated by the fact that some of the reviewers were included in both program committees.

https://doi.org/10.1017/S135132491200006X Published online by Cambridge University Press

in the training. The tagger is learnt in a grammar-satisfying space and uses a CFG filter to impose grammar constraints for the model parameter's update. The experiments show that the proposed structure-guided supertaggers improve on the baseline models. The structured model also results in an increase of the F-score of the final parser, achieving a competitive performance at higher parsing speed than state-of-the-art HPSG parsers.

The fourth paper by Do and Roth proposes a machine learning approach, which, by exploiting existing resources, such as Wikipedia, determines the taxonomic relation between pairs of terms, e.g., *car* is an ancestor of *Toyota Camry*. The model is based on a global constraint-based inference process that leverages an existing knowledge base to enforce relational constraints among terms. The experiments show that the global model significantly outperforms other systems built upon existing knowledge sources.

Finally, Li *et al.* describe topic models jointly used with entity relation detection. The text between pairs of named entities is used to define mini-documents, which are characterized by topic distributions. The resulting system is based on Maximum Entropy Discriminant Latent Dirichlet Allocation (MedLDA) with mixed membership for relation detection. Such membership formulation enables the system to incorporate heterogeneous features. The experiments on standard relation extraction corpus show that the proposed system improves the baselines, i.e., SVMs and LDA, respectively.

Acknowledgments

The guest editors would like to express their gratitude to all contributing authors, the executive editor, Ruslan Mitkov, the editors and the staff of the *Journal of Natural Language Engineering*. We are especially grateful to Stephen Clark and Hwee Tou Ng, who agreed to be the representatives of the JNLE editorial board in our guest editorial board for the special issue, and to Irina Temnikova and Natalia Konstantinova, who greatly helped us during the whole process. Last but not least, the high quality reviewing and selection process could not have been possible without the help of our expert guest editorial board, composed by Roberto Basili, Ulf Brefeld, Razvan Bunescu, Nicola Cancedda, Xavier Carreras, Stephen Clark, Trevor Cohn, Walter Daelemans, Hal Daumé, Jason Eisner, James Henderson, Liang Huang, Richard Johansson, Terry Koo, Mirella Lapata, Yuji Matsumoto, Ryan McDonald, Raymond Mooney, Hwee Tou Ng, Sebastian Riedel, Dan Roth, Mihai Surdeanu, Ivan Titov, Kristina Toutanova, Jun'ichi Tsujii, Antal van den Bosch, Scott Yih, Fabio Massimo Zanzotto, and Min Zhang.

Guest editors' work is partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant numbers 247758 (EternalS), 288024 (LiMoSINe), 247762 (FAUST), and 247914 (MOLTO).

References

Brill, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Applied Natural Language Processing*, Povo, Trento, Italy.

https://doi.org/10.1017/S135132491200006X Published online by Cambridge University Press

Carreras, X., and Màrquez, L. 2005. Introduction to the CoNLL-2005 shared task: semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pp. 152–64. Ann Arbor, MI, USA: Association for Computational Linguistics.

Chang, M., Ratinov, L., Rizzolo, N., and Roth, D. 2008 (July). Learning and inference with constraints. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Chicago, Illinois, USA, July 13–17.

Charniak, E., and Johnson, M. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, University of Michigan, Ann Arbor, USA, June 25–30, pp. 363–70.

Collins, M. 2000. Discriminative reranking for natural language parsing. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, Stanford, CA, USA, June 29–July 2, pp. 175–82.

Collins, M., and Duffy, N. 2002. New ranking algorithms for parsing and tagging: kernels over discrete structures, and the voted perceptron. In *Proceedings of the Association for Computational Linguistics*, Philadelphia, PA, USA, pp. 263–70.

Dinarelli, M., Moschitti, A., and Riccardi, G. 2009. Re-ranking models based-on small training data for spoken language understanding. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 1076–85. Singapore: Association for Computational Linguistics.

Ge, R., and Mooney, R. J. 2006. Discriminative reranking for semantic parsing. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Sydney, Australia, pp. 263–70. Sydney, Australia: Association for Computational Linguistics.

Huang, L. 2008. Forest reranking: discriminative parsing with non-local features. In *Proceedings of ACL-08: HLT*, Columbus, OH, USA, pp. 586–94. Columbus, OH, USA: Association for Computational Linguistics.

Koo, T., and Collins, M. 2005. Hidden-variable models for discriminative reranking. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, pp. 507–14. Vancouver, British Columbia, Canada: Association for Computational Linguistics.

Kudo, T., Suzuki, J., and Isozaki, H. 2005. Boosting-based parse reranking with subtree features. In *Proceedings of ACL'05*, New York City, USA.

Lafferty, J., McCallum, A., and Pereira, F. 2001 (June). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, Williamstown, MA, USA, June 28–July 1, pp. 282–89.

Moschitti, A., Quarteroni, S., Basili, R., and Manandhar, S. 2007. Exploiting syntactic and shallow semantic kernels for question/answer classification. In *Proceedings of ACL'07*, Prague, Czech Republic, June 23–30.

Naseem, T., Chen, H., Barzilay, R., and Johnson, M. 2010. Using Universal Linguistic Knowledge to Guide Grammar Induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics. pp. 1234–1244.

Nguyen, T-V. T., Moschitti, A., and Riccardi, G. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In *Proceedings of the 2009 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, Singapore.

Punyakanok, V., Roth, D., and Yih, W. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics* **34**(2): 257–87.

Punyakanok, V., Roth, D., Yih, W., and Zimak, D. 2005. Learning and inference over constrained output. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Edinburgh, Scotland, UK, July 30–August 5.

Roark, B., Liu, Y., Harper, M., Stewart, R., Lease, M., Snover, M., Shafran, I., Dorr, B., Hale, J., Krasnyanskaya, A., and Yung, L. 2006. Reranking for sentence boundary detection in conversational speech. In *ICASSP*, Toulouse, France, vol. 1, pp. 545–48.

Shawe-Taylor, J., and Cristianini, N. 2004. *Kernel Methods for Pattern Analysis*. Cambridge, UK: Cambridge University Press.

Shen, L., Sarkar, A., and Och, F. J. 2004. Discriminative reranking for machine translation. In *HLT-NAACL*, Boston, MA, USA, pp. 177–84.

Taskar, B., Guestrin, C., and Koller, D. 2004. Max-margin Markov networks. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems 16*. Cambridge, MA, USA: MIT Press.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. 2005. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research (JMLR)* **6**(September): 1453–84.

Vapnik, V. N. 1998. *Statistical Learning Theory*. New York, NY, USA: John Wiley.

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of Roget’s categories trained on large corpora. In *Proceedings of the 14th Conference on Computational linguistics* (*COLING ’92*), vol. 2, pp. 454–60. Stroudsburg, PA, USA: Association for Computational Linguistics.

https://doi.org/10.1017/S135132491200006X Published online by Cambridge University Press