A Mission for Computational Natural Language Learning

Walter Daelemans
CNTS Language Technology Group
University of Antwerp
Belgium
walter.daelemans@ua.ac.be

Abstract

In this presentation, I will look back at 10 years of CoNLL conferences and the state of the art of machine learning of language that is evident from this decade of research. My conclusion, intended to provoke discussion, will be that we currently lack a clear motivation or "mission" to survive as a discipline. I will suggest that a new mission for the field could be found in a renewed interest for theoretical work (which learning algorithms have a bias that matches the properties of language?, what is the psycholinguistic relevance of learner design issues?), in more sophisticated comparative methodology, and in solving the problem of transfer, reusability, and adaptation of learned knowledge.

1 Introduction

When looking at ten years of CoNLL conferences, it is clear that the impact and the size of the conference has enormously grown over time. The technical papers you will find in this proceedings now are comparable in quality and impact to those of other distinguished conferences like the *Conference on Empirical Methods in Natural Language Processing* or even the main conferences of ACL, EACL and NAACL themselves. An important factor in the success of CoNLL has been the continued series of shared tasks (notice we don't use terms like *challenges* or *competitions*) that has produced a useful set of benchmarks for comparing learning methods, and that has gained wide interest in the field. It should also be noted, however, that the success of the conferences is inversely proportional with the degree to which the original topics which motivated the conference are present in the programme. Originally, the people driving CoNLL wanted it to be promiscuous (i) in the selection of partners (we wanted to associate with Machine Learning, Linguistics and Cognitive Science conferences as well as with Computational Linguistics conferences) and (ii) in the range of topics to be presented. We wanted to encourage linguistically and psycholinguistically relevant machine learning work, and biologically inspired and innovative symbolic learning methods, and present this work alongside the statistical and learning approaches that were at that time only starting to gradually become the mainstream in Computational Linguistics. It has turned out differently, and we should reflect on whether we have become too much of a mainstream computational linguistics conference ourselves, a back-off for the good papers that haven't made it in EMNLP or ACL because of the crazy rejection rates there (with EMNLP in its turn a back-off for good papers that haven't made it in ACL). Some of the work targeted by CoNLL has found a forum in meetings like the workshop on *Psycho-computational models of human language acquisition*, the *International Colloquium on Grammatical Inference*, the workshop on *Morphological and Phonological Learning* etc. We should ask ourselves why we don't have this type of work more in CoNLL. In the first part of the presentation I will sketch *very* briefly the history of SIGNLL and

Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X),
pages 1–5, New York City, June 2006.

CoNLL and try to initiate some discussion on what a conference on Computational Language Learning should be doing in 2007 and after.

2 State of the Art in Computational Natural Language Learning

The second part of my presentation will be a discussion of the state of the art as it can be found in CoNLL (and EMNLP and the ACL conferences). The field can be divided into theoretical, methodological, and engineering work. There has been progress in theory and methodology, but perhaps not sufficiently. I will argue that most progress has been made in *engineering* with most often incremental progress on specific tasks as a result rather than increased understanding of how language can be learned from data.

Machine Learning of Natural Language (MLNL), or Computational Natural Language Learning (CoNLL) is a research area lying in the intersection of computational linguistics and machine learning. I would suggest that Statistical Natural Language Processing (SNLP) should be treated as part of MLNL, or perhaps even as a synonym. Symbolic machine learning methods belong to the same part of the ontology as statistical methods, but have different solutions for specific problems. E.g., Inductive Logic Programming allows elegant addition of background knowledge, memory-based learning has implicit similarity-based smoothing, etc.

There is no need here to explain the success of inductive methods in Computational Linguistics and why we are all such avid users of the technology: availability of data, fast production of systems with good accuracy, robustness and coverage, cheaper than linguistic labor. There is also no need here to explain that many of these arguments in favor of learning in NLP are bogus. Getting statistical and machine learning systems to work involves design, optimization, and smoothing issues that are something of a black art. For many problems, getting sufficient annotated data is expensive and difficult, our annotators don't sufficiently agree, our trained systems are not really that good. My favorite example for the latter is part of speech tagging, which is considered a solved problem, but still has error rates of 20-30% for the ambiguities that count, like verb-noun ambiguity. We are doing better than handcrafted linguistic knowledge-based approaches but from the point of view of the goal of robust language understanding unfortunately not that significantly better. Twice better than very bad is not necessarily any good. We also implicitly redefined the goals of the field of Computational Linguistics, forgetting for example about quantification, modality, tense, inference and a large number of other sentence and discourse semantics issues which do not fit the default classification-based supervised learning framework very well or for which we don't have annotated data readily available. As a final irony, one of the reasons why learning methods have become so prevalent in NLP is their *success* in speech recognition. Yet, there too, this success is relative; the goal of spontaneous speaker-independent recognition is still far away.

2.1 Theory

There has been a lot of progress recently in theoretical machine learning(Vapnik, 1995; Jordan, 1999). Statistical Learning Theory and progress in Graphical Models theory have provided us with a well-defined framework in which we can relate different approaches like kernel methods, Naive Bayes, Markov models, maximum entropy approaches (logistic regression), perceptrons and CRFs. Insight into the differences between generative and discriminative learning approaches has clarified the relations between different learning algorithms considerably.

However, this work does not tell us something general about machine learning *of language*. Theoretical issues that should be studied in MLNL are for example which classes of learning algorithms are best suited for which type of language processing task, what the need for training data is for a given task, which information sources are necessary and sufficient for learning a particular language processing task, etc. These fundamental questions all relate to learning algorithm *bias* issues. Learning is a search process in a hypothesis space. Heuristic limitations on the search process and restrictions on the representations allowed for input and hypothesis representations together define this bias. There is not a lot of work on matching properties of learning algorithms with properties of language processing

tasks, or more specifically on how the bias of particular (families of) learning algorithms relates to the hypothesis spaces of particular (types of) language processing tasks.

As an example of such a unifying approach, (Roth, 2000) shows that several different algorithms (memory-based learning, tbl, snow, decision lists, various statistical learners, ...) use the same type of knowledge representation, a linear representation over a feature space based on a transformation of the original instance space. However, the only relation to language here is rather negative with the claim that this bias is not sufficient for learning higher level language processing tasks.

As another example of this type of work, Memory-Based Learning (MBL) (Daelemans and van den Bosch, 2005), with its implicit similarity-based smoothing, storage of all training evidence, and uniform modeling of regularities, subregularities and exceptions has been proposed as having the right bias for language processing tasks. Language processing tasks are mostly governed by Zipfian distributions and high disjunctivity which makes it difficult to make a principled distinction between noise and exceptions, which would put *eager* learning methods (i.e. most learning methods apart from MBL and kernel methods) at a disadvantage.

More theoretical work in this area should make it possible to relate machine learner bias to properties of language processing tasks in a more fine-grained way, providing more insight into both language and learning. An avenue that has remained largely unexplored in this respect is the use of artificial data emulating properties of language processing tasks, making possible a much more fine-grained study of the influence of learner bias. However, research in this area will not be able to ignore the "no free lunch" theorem (Wolpert and Macready, 1995). Referring back to the problem of induction (Hume, 1710) this theorem can be interpreted that no inductive algorithm is universally better than any other; generalization performance of any inductive algorithm is zero when averaged over a uniform distribution of all possible classification problems (i.e. assuming a random universe). This means that the only way to test hypotheses about bias and necessary information sources in language learning is to perform empirical research, making a reliable experimental methodology necessary.

2.2 Methodology

Either to investigate the role of different information sources in learning a task, or to investigate whether the bias of some learning algorithm fits the properties of natural language processing tasks better than alternative learning algorithms, *comparative* experiments are necessary. As an example of the latter, we may be interested in investigating whether part-of-speech tagging improves the accuracy of a Bayesian text classification system or not. As an example of the former, we may be interested to know whether a relational learner is better suited than a propositional learner to learn semantic function association. This can be achieved by comparing the accuracy of the learner with and without the information source or different learners on the same task. Crucial for objectively comparing algorithm bias and relevance of information sources is a methodology to reliably measure differences and compute their statistical significance. A detailed methodology has been developed for this involving approaches like k-fold cross-validation to estimate classifier quality (in terms of measures derived from a confusion matrix like accuracy, precision, recall, F-score, ROC, AUC, etc.), as well as statistical techniques like McNemar and paired cross-validation t-tests for determining the statistical significance of differences between algorithms or between presence or absence of information sources. This methodology is generally accepted and used both in machine learning and in most work in inductive NLP.

CoNLL has contributed a lot to this comparative work by producing a successful series of shared tasks, which has provided to the community a rich set of benchmark language processing tasks. Other competitive research evaluations like senseval, the PASCAL challenges and the NIST competitions have similarly tuned the field toward comparative learning experiments. In a typical comparative machine learning experiment, two or more algorithms are compared for a fixed sample selection, feature selection, feature representation, and (default) algorithm parameter setting over a number of trials (cross-validation), and if the measured differences are statistically significant, conclusions are drawn about which algorithm is better suited to the problem

being studied and why (mostly in terms of algorithm bias). Sometimes different sample sizes are used to provide a learning curve, and sometimes parameters of (some of the) algorithms are optimized on training data, or heuristic feature selection is attempted, but this is exceptional rather than common practice in comparative experiments.

Yet everyone knows that many factors potentially play a role in the outcome of a (comparative) machine learning experiment: the data used (the sample selection and the sample size), the information sources used (the features selected) and their representation (e.g. as nominal or binary features), the class representation (error coding, binarization of classes), and the algorithm parameter settings (most ML algorithms have various parameters that can be tuned). Moreover,all these factors are known to interact. E.g., (Banko and Brill, 2001) demonstrated that for confusion set disambiguation, a prototypical disambiguation in context problem, the amount of data used dominates the effect of the bias of the learning method employed. The effect of training data size on relevance of POS-tag information on top of lexical information in relation finding was studied in (van den Bosch and Buchholz, 2001). The positive effect of POS-tags disappears with sufficient data. In (Daelemans et al., 2003) it is shown that the joined optimization of feature selection and algorithm parameter optimization significantly improves accuracy compared to sequential optimization. Results from comparative experiments may therefore not be reliable. I will suggest an approach to improve methodology to improve reliability.

2.3 Engineering

Whereas comparative machine learning work can potentially provide useful theoretical insights and results, there is a distinct feeling that it also leads to an exaggerated attention for accuracy *on the dataset*. Given the limited transfer and reusability of learned modules when used in different domains, corpora etc., this may not be very relevant. If a WSJ-trained statistical parser looses 20% accuracy on a comparable newspaper testcorpus, it doesn't really matter a lot that system A does 1% better than system B on the default WSJ-corpus partition.

In order to win shared tasks and perform best on some language processing task, various clever architectural and algorithmic variations have been proposed, sometimes with the single goal of getting higher accuracy (ensemble methods, classifier combination in general, ...), sometimes with the goal of solving manual annotation bottlenecks (active learning, co-training, semisupervised methods, ...).

This work is extremely valid from the point of view of computational linguistics researchers looking for any old method that can boost performance and get benchmark natural language processing problems or applications solved. But from the point of view of a SIG on computational natural language learning, this work is probably too much theory-independent and doesn't teach us enough about *language* learning.

However, engineering work like this can suddenly become theoretically important when motivated not by a few percentage decimals more accuracy but rather by (psycho)linguistic plausibility. For example, the current trend in combining local classifiers with holistic inference may be a cognitively relevant principle rather than a neat engineering trick.

3 Conclusion

The field of computational natural language learning is in need of a renewed mission. In two parent fields dominated by good engineering use of machine learning in language processing, and interesting developments in computational language learning respectively, our field should focus more on theory. More research should address the question what we can learn about language from comparative machine learning experiments, and address or at least acknowledge methodological problems.

4 Acknowledgements

There are many people that have influenced me, most of my students and colleagues have done so at some point, but I would like to single out David Powers and Antal van den Bosch, and thank them for making this strange field of computational language learning such an interesting and pleasant playground.

References

Michele Banko and Eric Brill. 2001. Mitigating the paucity-of-data problem: exploring the effect of train-

ing corpus size on classifier performance for natural language processing. In *HLT '01: Proceedings of the first international conference on Human language technology research*, pages 1–5, Morristown, NJ, USA. Association for Computational Linguistics.

Walter Daelemans and Antal van den Bosch. 2005. *Memory-Based Language Processing*. Cambridge University Press, Cambridge, UK.

Walter Daelemans, Véronique Hoste, Fien De Meulder, and Bart Naudts. 2003. Combined optimization of feature selection and algorithm parameter interaction in machine learning of language. In *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, Lecture Notes in Computer Science 2837, pages 84–95, Cavtat-Dubrovnik, Croatia. Springer-Verlag.

D. Hume. 1710. *A Treatise Concerning the Principles of Human Knowledge*.

M. I. Jordan. 1999. *Learning in graphical models*. MIT, Cambridge, MA, USA.

D. Roth. 2000. Learning in natural language: Theory and algorithmic approaches. In *Proc. of the Annual Conference on Computational Natural Language Learning (CoNLL)*, pages 1–6, Lisbon, Portugal.

Antal van den Bosch and Sabine Buchholz. 2001. Shallow parsing on the basis of words only: a case study. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 433–440, Morristown, NJ, USA. Association for Computational Linguistics.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

David H. Wolpert and William G. Macready. 1995. No free lunch theorems for search. Technical Report SFI-TR-95-02-010, Santa Fe, NM.