

Machine Learning in Natural Language Processing

Andrew Roberts

16th October 2003

Machine learning is concerned with acquiring knowledge from an environment in a computational manner, in order to improve the performance. (See [15] for overview of ML.) Many forces over the past couple of decades have seen the blending of ML and NLP become increasingly common. The ever expanding availability of large corpora; more powerful computing resources; and a greater demand for natural language based applications, are three important factors.

Many important tasks within NLP require a substantial amount of knowledge, usually provided manually by practitioners within the field. However, with a shift to automatic processing of large amounts of language data, empirical methods have sought to reduce the dependence of manually embedding knowledge into NLP systems, and to let learning algorithms acquire the knowledge from available data. Data classification is a thoroughly active research topic within ML, and much research where ML and NLP have been combined is aimed at transforming a common NLP problem so that it can be represented as a classification problem. As a result, it is not uncommon nowadays to see that most well known ML techniques have been applied to just about every possible NLP task, where feasible. See [24] for an excellent review on the marrying of these two areas, and [12] for more in depth coverage.

Linear classifiers come in various flavours, but are all relatively simple to understand and are computationally efficient, particularly with 2-class problems. For example, linear threshold algorithms can calculate a weighted sum from input features, and then depending whether that sum is greater or smaller than a certain threshold, a decision can be made as its class. Thresholds are determined through the use of training data. Such classifiers have been applied to text categorisation [8, 19], shallow parsing [21] and POS tagging [26].

Decision trees are used to partition large samples of data into a hierarchical structure. Commonly associated as a tool for classification, they are also capable of generalising seen data into sets of rules. They are generic enough to be applied to many sub-tasks of NLP, and have been found in POS tagging [23], machine

translation [29], parsing [9], text categorisation [31] and word-sense disambiguation [3].

Clustering is a powerful method as it is one of the few that is fully unsupervised. It aims to discover natural partitions within data by grouping entities into clusters based on their similarity. Utilised in syntactic [25, 10] and semantic classification and information retrieval [11].

Memory-based learning (MBL)¹ — as the name suggests — is a method of classifying by having a full memory of previously seen examples at its disposal. The essence of this method is that we learn by comparing similarities of past experiences with new ones, as opposed to rule-based models. Traditionally, what is referred to as the *performance stage* of a MBL system, i.e., the classification stage, is often descended from the simple *k-nn* (*k* nearest neighbours) algorithm [5].

The TiMBL system, developed at Tilberg University has enjoyed large success in this domain. This system is more complex than the basic description of MBL above. It employs an algorithm to compress examples into a decision-tree like structure. This optimisation step means that the classification process does not have to examine all examples in memory, and instead only focuses only nodes with important feature relevant to the instance being classified. A brief summary of all tasks successfully tackled by TiMBL can be found in [7], and for a more in depth review of TiMBL applied to speech processing, POS tagging and phrase chunking, see [6].

Inductive logic programming (ILP) involves the application of machine learning to gain knowledge from a given domain, and then convert this knowledge into a set of facts and rules, which are represented in first-order logic. Examples of this discipline include research from the University of Texas, culminating in the CHILL system, which has been applied to grammar inference tasks [32, 33]. ILP has also been utilised in information extraction [28] and facilitating natural languages queries to databases [30].

Neural networks (NNs) are a well established and popular within AI. It is therefore not surprising to see how they have been used in a wide variety of problems within NLP. NNs consist of units that are connected by links. The links are assigned a value, known as a weight. Units perform simple computations based on the inputs, and pass on their output. Networks of these units can be connected together in a suitable topology for a given task. Units in between in the input and output units are called hidden units. Once layers of hidden units are present in a NN, it is possible to represent complex problems. The learning step is achieved through training, whereby using an algorithm such as back-propagation, it is pos-

¹Memory-based learning is one of several names that this approach has been labelled with. Alternatives include: instance-based, case-based, example-based, similarity-based and lazy learning.

sible to update the weights within the NN to boost its performance. Applications of NNs within NLP include grammar inference [17, 18], POS tagging [22, 27], speech processing [1] and parsing [4].

Genetic algorithms/evolutionary computing has proved to be a successful technique in optimising solutions, especially for difficult problems with large search spaces. Rather than a brute force approach (which is often computationally unfeasible), GAs emulate natural selection to a certain degree, in an attempt to *evolve* towards an optimal solution. GAs are prone to getting stuck at local maxima, as it is a greedy algorithm that evolves for the largest short term gain. That said, it is still a very powerful technique, and for its abilities to cope with large problems, has been applied to information retrieval [20], morphology [13], dialogue systems [2] and grammar inference [16].

See [14] for a survey on how NNs and GAs have been applied to NLP problems, including combinatory approaches of the two techniques.

References

- [1] Y. Bengio, A connectionist approach to speech recognition, *International Journal on Pattern Recognition and Artificial Intelligence* 7 (1993).
- [2] M. Blasband, GAG: Genetic Algorithms for Grammars, Tech. rep., CompuLeer (1998).
- [3] P. Brown, S. Della Pietra, V. Della Pietra and R. Mercer, Word sense disambiguation using statistical methods, in: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)* (1991), pp. 264–270.
- [4] F. Buø and A. Waibel, FeasPar: A feature structure parser learning to parse spoken language, in: *Proceedings of 16th International Conference on Computational Linguistics (COLING-96)* (Copenhagen, Denmark, 1996), pp. 188–193.
- [5] T. Cover and P. Hart, Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* 13 (1967) 21–27.
- [6] W. Daelemans, A. Van den Bosch, J. Zavrel, J. Veenstra, S. Buchholz and G. Busser, Rapid development of NLP modules with memory-based learning, in: *In Proceedings of ELSNET in Wonderland* (1998), pp. 105–113.
- [7] W. Daelemans, J. Zavrel and K. van der Sloot, TiMBL: Tilburg Memory—Based Learner. Version 4.3. Reference guide,

- Tech. Rep. ILK 02-10, Induction of Linguistic Knowledge Research Group, Tilberg University, The Netherlands (2002), URL <http://ilk.uvt.nl/downloads/pub/papers/ilk.0210.ps>.
- [8] I. Dagan, Y. Karov and D. Roth, Mistake-driven learning in text categorization, in: *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Brown University, Providence, Rhode Island, 1997).
- [9] M. Haruno, S. Shirai and Y. Ooyama, Using decision trees to construct a practical parser, in: *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (Montréal, Canada, 1998), pp. 1136–1142.
- [10] J. Hughes, *Automatically Acquiring a Classification of Words*, Ph.D. thesis, School of Computing, University of Leeds (1994).
- [11] O. Ibrahimov, I. Sethi and N. Dimitrova, Clustering of imperfect transcripts using a novel similarity measure, in: *Proceedings of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications* (2001).
- [12] D. Kazakov, *Natural Language Processing Applications of Machine Learning*, Ph.D. thesis, Czech Technical University, Prague (1996), URL <ftp://ftp.cs.york.ac.uk/pub/aig/Papers/dimitar.kazakov/content.ps.gz>.
- [13] D. Kazakov, Unsupervised learning of naive morphology with genetic algorithms, in: *Workshop Notes of the ECML/MLnet workshop on empirical learning of Natural Language Processing Tasks* (1997).
- [14] A. Kool, Literature survey (1999), URL <http://cnts.uia.ac.be/kool/litsurv.ps>, unpublished manuscript, University of Antwerp.
- [15] P. Langley, *Elements of Machine Learning* (Morgan Kaufmann, 1996).
- [16] M. M. Lankhorst, Breeding grammars: Grammatical inference with a genetic algorithm, Tech. Rep. CS-R 9401, Department of Computer Science, University of Gronigen, The Netherlands (1994), URL <http://www.cs.rug.nl/info/reports/csr9401.ps.gz>.
- [17] S. Lawrence, S. Fong and C. L. Giles, Natural language grammatical inference: a comparison of recurrent neural networks and machine learning methods, in: *Connectionist, Statistical, and Symbolic Approaches to Learning*

for *Natural Language Processing*, eds. S. Wermter, E. Riloff and G. Scheler (Springer Verlag, Berlin, 1996), vol. 1040 of *LNAI*, pp. 33–47, URL <http://www.neci.nj.nec.com/homepages/lawrence/papers/nl-book96/nl-book96.pdf>

- [18] S. Lawrence, C. L. Giles and S. Fong, Natural language grammatical inference with recurrent neural networks, *IEEE Transactions on Knowledge and Data Engineering* 12 (2000) 126–140, URL <http://www.neci.nec.com/lawrence/papers/nl-tkde98/nl-tkde98.pdf>.
- [19] D. Lewis, R. Schapire, J. Callan and R. Papka, Training algorithms for linear text classifiers, in: *Proceedings of the 19th International Conference on Research and Development in Information Retrieval, SIGIR* (Zurich, Switzerland, 1996), pp. 298–306.
- [20] R. M. Losee, Learning syntactic rules and tags with genetic algorithms for information retrieval and filtering: An empirical basis for grammatical rules, *Information Processing & Management* 32 (1996) 185–197, URL <http://www.ils.unc.edu/losee/genel.pdf>.
- [21] M. Muñoz, V. Punyakanok, D. Roth and D. Zimak, A learning approach to shallow parsing, in: *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC)* (1999).
- [22] Q. Ma and H. Isahara, A multi-neuro tagger using variable lengths of contexts, in: *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (Montréal, Canada, 1998), pp. 802–806.
- [23] L. Màrquez, *Part-of-Speech Tagging: A Machine Learning Approach based on Decision Trees*, Ph.D. thesis, Dep. Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya (1999).
- [24] L. Màrquez, Machine Learning and Natural Language Processing, Seminar: Industrias de la lengua / La ingeniería Lingüística en la sociedad de la información (2000), URL <http://www.lsi.upc.es/lluism/cursos/docSoria00/soria00.ps.gz>.
- [25] A. Roberts, Automatic acquisition of word classification using distributional analysis of content words with respect to function words, Tech. rep., School of Computing, University of Leeds (2002).

- [26] D. Roth and D. Zelenko, Part of speech tagging using a network of linear separators, in: *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (Montréal, Canada, 1998), pp. 1136–1142.
- [27] H. Schütze, Part-of-speech induction from scratch, in: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)* (Columbus, Ohio, 1993), pp. 251–258.
- [28] S. Soderland, D. Fisher, J. Aseltine and W. Lehnert, Crystal: Inducing a conceptual dictionary, in: *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (1995), pp. 1314–1319.
- [29] H. Tanaka, Decision tree learning algorithm with structural attributes: Application to verbal case frame acquisition, in: *Proceedings of the 16th International Conference on Computational Linguistics (COLING)* (Copenhagen, Denmark, 1996), pp. 943–948.
- [30] C. A. Thompson, R. J. Mooney and L. R. Tang, Learning to parse natural language database queries into logical form, in: *Workshop on Automata Induction, Grammatical Inference and Language Acquisition* (1997).
- [31] S. Weiss, C. Apte, F. Damerau, D. Johnson, F. Oles, T. Goetz and T. Hampp, Maximizing text-mining performance, *IEEE Intelligent Systems* 14 (1999) 63–69.
- [32] J. M. Zelle and R. J. Mooney, Learning semantic grammars with constructive inductive logic programming, in: *Proceedings of the 11th National Conference on Artificial Intelligence* (AAAI Press/MIT Press, Washington, D.C., 1993), pp. 817–822, URL <ftp://ftp.cs.utexas.edu/pub/mooney/papers/chill-aaai-93.ps.Z>.
- [33] J. M. Zelle and R. J. Mooney, Inducing deterministic Prolog parsers from treebanks: A machine learning approach, in: *Proceedings of the 12th National Conference on Artificial Intelligence* (AAAI Press/MIT Press, Seattle, WA, 1994), pp. 748–753, URL <ftp://ftp.cs.utexas.edu/pub/mooney/papers/chill-aaai-94.ps.Z>.