Machine Learning in Human Language Technology

Nikos D. Fakotakis and Kyriakos N. Sgarbas

Wire Communications Laboratory, Electrical and Computer Engineering Department,
University of Patras, GR-265 00 Rio, Greece
{fakotaki,sgarbas}@wcl.ee.upatras.gr

1 Introduction

The undoubted usefulness of present-day information systems is only moderated by the fact that people have to invest substantial effort and training time in order to learn how to use them. Even modern applications with Graphical-User Interfaces (which are considered user-friendly), built-in wizards and on-line context-sensitive help, require a considerable self-training period, thus discouraging most people from fully exploiting their capabilities. In the years to come we expect that information systems will gradually become more and more complex and since the training period is usually proportional to the system complexity, with the usual Human Computer Interaction methods less and less people will have the time to learn how to use a new piece of software.

The obvious solution to the aforementioned problem is to make the Human Computer Interface intelligent enough to understand human language. That is one vital objective of Human Language Technology (HLT). HLT is an emerging field, which is expected not only to facilitate future applications as the natural improvement of today's Graphical User Interfaces, but also to provide applications of significant importance like machine translation systems and tele-services. HLT will enable casual, not technically inclined users to benefit from modern information technology. HLT will also enable the retrieval of information from resources, which are not organized for computer access, like newspapers and audio recordings.

HLT is further subdivided into Speech Processing (SP) and Natural Language Processing (NLP). SP covers all tasks concerning spoken language input and output, including speech recognition and synthesis, signal representation, speaker identification/verification and several performance issues, like robustness, adaptation and real-time response. NLP covers grammars and parsing techniques, lexicon representations, syntax, semantics and dialogues. The two areas (SP and NLP) are generally complementary to each other although some "gray areas" do exist, in which overlaps occur frequently, e.g. SP systems may have a language model independent of any NLP module that may follow, or NLP systems may consider phonetic transcriptions in order to become more versatile.

HLT is an intriguing field of research where some very interesting problems arise; problems that researchers in the Machine Learning (ML) field would like to try solving using their own methods. In fact, HLT researchers already use some well known machine learning techniques like hidden Markov models (HMMs), neural

G. Paliouras, V. Karkaletsis, and C.D. Spyropoulos (Eds.): ACAI '99, LNAI 2049, pp. 267-273, 2001.

networks (NNs), probabilistic context-free grammars (PCFGs), etc., although they prefer the term "training" instead of "learning" to denote the adaptation of their systems. But most of the HLT systems, especially in the NLP area, are built by handcrafted rules derived from linguistic theories and syntax textbooks. There is nothing wrong about that of course, especially if the systems perform acceptably in practice, except that systems produced by this method are highly language dependent and very likely application dependent too. Considering the time and effort needed to build such systems for several languages and different applications, the need of language independent and application independent methods of language representation becomes imperative. Thus one of the most promising subfields of NLP is the so-called "corpus-based NLP" which aims to extract lexicons, linguistic rules and related knowledge from large text corpora automatically, usually by statistic methods.

We believe that if the communication between the HLT community and the ML community becomes stronger, then even more sophisticated ML methods will be applied to HLT systems and new ML techniques will be devised having the HLT requirements in mind. The use and evaluation of existing and innovative machine learning techniques in the field of HLT will affect significantly the overall performance of HLT systems and will indicate new directions for research and development. The ML field is mature enough to provide solutions and methodologies from which an emerging technology like HLT will certainly benefit. Fortunately, the scientific community has realized the importance of the cooperation between ML and HLT [8]. It remains to see to what extent these two fields can cooperate and what promises lie towards this direction.

This chapter stresses the importance for cooperation in research between the field of Machine Learning and the field of Human Language Technology, arguing that this cooperation can be of benefit to both scientific areas in terms of trends, aims and methods used. This argument is backed up by the results of the *Workshop on Machine Learning in Human Language Technology* (held on July 7-8, 1999, Chania, Greece, in the framework of ACAI '99), by introducing the papers presented in the Workshop and pointing out how these papers advance the research status of both fields.

2 Brief History

Although ML-oriented ideas like training the system to a set of rules (grammar) have been researched quite early in the SP field (see for example [1]), analogous attempts in NLP have flourished mostly in the nineties with a renewed interest towards corpus-based research, after the rule-based techniques used so far proved inefficient to confront real-world texts [3]. The new techniques were first applied as enhancements to rule-based methods (like [17]) thus producing hybrid components, and later were applied as stand-alone methods with satisfactory robust results. The applied ML-techniques varied from statistical/probabilistic methods [9], [23] to transformation-based learning [5] and n-gram models [7]. The good news with the statistical/corpus-based/ML methods was that they were mainly language independent [11], robust [4]

and application independent [6]. The new techniques have also indicated new interesting directions for research, for example adaptation of existing systems to different domains [26]. For a thorough historic review on the subject see [21], [12] and [10].

3 The ACAI '99 Workshop on ML in HLT

The *Workshop on Machine Learning in Human Language Technology* was intended to promote the cooperation between the two scientific fields. The Workshop was held on July 7-8, 1999, at Chania, Greece, in the framework of the ECCAI Advanced Course on Artificial Intelligence for 1999 (ACAI '99), whose goal was to present the current state-of-the-art in Machine Learning, and to show the potential of Machine Learning in a variety of problems. The aim of the Workshop was the presentation of the current Machine Learning research activities in the area of Human Language Technology (i.e. Written and Spoken Language Recognition, Understanding, Generation, etc.).

The papers presented in the Workshop [27] addressed several HLT issues, like *Part-of-Speech Tagging, Word Sense Disambiguation, Unknown Word Acquisition, Sentence Boundary Disambiguation, etc.*, using a wide range of ML techniques, like *Decision Tree Induction, Transformation-Based Learning, Self-Organization, Memory-Based Learning, etc*. Descriptions of these papers are presented in the following section.

4 Overview of the Papers

The invited talk by Kodratoff [14] examined the potential of Knowledge Discovery in Texts (KDT) as a new scientific topic emerging from Knowledge Discovery in Databases (KDD or Data Mining, as often called). The author compared KDT with classical Natural Language Processing (NLP) and explained the different nature of the KDT-extracted knowledge arguing that the knowledge obtained by KDT, since it is extracted from a large number of texts, constitutes an absolutely new type of knowledge, often complementary to the knowledge obtained by NLP. The author provided a detailed example which presented the different types of knowledge that the NLP system TROPES and the KDD system CLEMENTINE extract from the same text corpus. Since Data Mining is an always-hot topic with a very broad range of applications, we can expect that KDT techniques will certainly develop and flourish in the years to come.

The paper by Orphanos et al. [18] presented a machine learning approach to the problems of disambiguation and unknown word guessing based on decision trees. Three induction algorithms were introduced; two producing generalized trees and one producing binary trees. All three algorithms use set value attributes in decision tree induction in a linguistic context obtained by some extensions to the basic model. The

authors used a Modern Greek corpus of more than 130,000 tokens with wide thematic coverage to obtain performance results for the aforementioned three algorithms in POS disambiguation and POS guessing. The results were very positive: the authors reported 93-95% accuracy in POS disambiguation and 82-88% in guessing the POS of unknown words. These results are presented in detail in the paper and compared to each other, according to the algorithms used.

Megyesi [16] presented the implementation of a Brill tagger for Hungarian. The author has adapted and tested the tagger for the particular language, by automatically acquiring rules from a training corpus. The method used was based on Transformation-Based Error-Driven Learning, but the results obtained at first did not reach the high accuracy levels of the English implementation. Due to the rich morphology, high inflectionality and free word order of the Hungarian language, the method obtained only 83% accuracy. In order to improve the overall performance of the system, the author explained how she augmented the tagger's rule generating mechanism with extended lexical templates, obtaining an impressive 14% increase in accuracy for a final 97%.

Another method to enhance the performance of a Brill tagger was presented in the paper by Petasis et al. [19]. The authors used Transformation-Based Error-Driven Learning to resolve the POS ambiguity in Modern Greek texts. They presented experimental results from two different test cases, one based on a 65,000-word corpus on "management succession events" and the other based on a 125,000-word general-purpose corpus, obtaining 95% accuracy. One additional interesting result of these experiments was evidence that the performance of the method is independent of the thematic domain of the corpus used in the training process. The authors have also estimated the optimal size for the training corpus with respect to performance over training effort, i.e. the size after which the increase in performance is too small to justify further training. This optimal size was measured to 18,000 words for the Greek language.

On the task of word sense disambiguation, Levinson [15] presented a fully unsupervised method based on pairwise similarity clustering. The proposed method needs just a sufficiently large monolingual corpus in order to be applied and it can be used for both morphological disambiguation and for discrimination between different meanings of the same part of speech. Since it does not rely on any linguistic characteristics but only on the manner in which information is organized and conveyed, the method proves to be language independent. The author presented the successful application of his method to English and Hebrew, two languages with very different structures, and examined several variations of the main method.

The paper by Basili et al. [2] presented a conceptual clustering method for learning subcategorisation frames from a corpus and described an original architecture for tuning the lexicon to a specific sublanguage in order to improve syntactic analysis. A lexicalised shallow parser was used for the evaluation of the proposed techniques. The authors described RGL, a conceptual clustering system for learning verb subcategorisation frames, and CHAOS, a robust parser for information extraction,

combined in a bootstrap architecture tested with a collection of tagged trees from a large newspaper corpus.

On the issue of unknown word acquisition, Kameda et al. [13] presented a rule-based method applied to Japanese language. The authors defined three classes of unknown words: heterographs, compound words and out-of-dictionary words. Especially for compound words, which are characterised by an underlying syntactic structure, the authors defined a surface layer and a deep layer utilizing a four-stage acquisition process. The paper described the corresponding algorithms for each class of unknown words and presented evaluation results for a system implementing the proposed method. The reported results were 96.8% for class-1 (heterographs), 76.4% (words) and 73.4% (lexemes) for class-2 (compound words).

In the paper by Thomas [24] an inductive machine learning method was presented for the construction of wrappers from semi-structured documents (such as WEB-pages). The proposed method learns how to construct T-wrappers for multi-slot extraction in an automatic way from positive examples, which consist of text-tuples occurring in the document. The technique is based on a modified version of least general generalization (TD-Anti-Unification) for a subset of feature structures (tokens). The author has tested the method on web pages with various types of structured information.

Using a “house design” task as an example, the paper by Sakurai et al. [20] described an algorithm for self-organization of a task model from word sequences. The proposed algorithm uses six basic rules and two types of background knowledge: “sequence changeable words knowledge” and “identical words knowledge”. The authors described a structure called an Extended Object Automaton and they defined an inference method for constructing automatically Extended Object Automata based on a self-organization procedure. The proposed method has been successfully applied to a multi-modal system performing a house design task. The authors report 90% acceptance ratio for 175 sequences, 90% activity ratio and negative acceptance ratio less than 0.1%.

The problem of sentence boundary disambiguation was addressed by Stamatatos et al. [22]. The authors proposed a variation of Transformation-Based Learning (TBL) applied to the observation of preceding and following words and punctuation marks in order to automatically extract sentence boundary disambiguation rules from real-world texts. The proposed method has been tested for Modern Greek texts downloaded from the World Wide Web consisting of more than 165,000 words for the training corpus and more than 200,000 words for the test corpus and achieved high accuracy results (99.4%). The authors have also performed a test comparing their method to the traditional TBL, reporting a 3.7% increase in accuracy, in favor of their method.

Text chunking as an approach to data reduction in computational linguistics was the subject of the paper by Veenstra [25]. The author used Memory-Based Learning in the context of subject/object identification to perform efficient NP, VP and PP chunking. The method has been tested on an annotated and POS-tagged newspaper corpus of more than one million words and presented high precision and recall scores (94-95%).

5 Conclusion

It seems that the cooperation and exchange of ideas between the field of Machine Learning and the field of Human Language Technology provides many benefits for both fields. Some ML techniques have direct impact to HLT applications and HLT has some very interesting problems that require specially designed or augmented ML techniques. The papers presented in the ACAI '99 Workshop on Machine Learning in Human Language Technology give substantial evidence that research in this direction is promising and worthwhile.

Acknowledgments. We are grateful to all who have contributed to make this Workshop possible: the ACAI '99 Organisers, the Workshops Chair Dr. V. Karkaletsis, the Workshop Committee, the invited speaker Prof. Yves Kodratoff, the authors who submitted and presented their papers and all the participants.

References

1. Baker, J. (1979). "Trainable Grammars for Speech Recognition", in Speech Communication Papers for the 97th Meeting of the Acoustical Society of America, June 1979,Cambridge, MA, New York, pp.547-550.
2. Basili R., Pazienza M. T., Vindigni M. (1999). "Lexical Learning for Improving Syntactic Analysis". In [27].
3. Black E., Garside R., Leech G. (eds.) (1993). "Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach", Language and Computers: Studies in Practical Linguistics 8, Rodopi, Amsterdam.
4. Brent M. (1993). "From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax", Computational Linguistics: Special Report on Using Large Corpora: II, Vol.19, pp.243-262.
5. Brill E., (1993). "Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach", in Proc. of the 31st Annual Meeting of ACL, June 1993, Columbus, OH, pp.259-265.
6. Briscoe E., Carroll J. (1991). "Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars", Cambridge University, Technical Report 224.
7. Brown P., Della Pietra V., deSouza P., Lai J., Mercer R. (1992). "Class-Based n-Gram Models of Natural Language", Computational Linguistics, Vol.18, pp.467-479.
8. Cardie C., Mooney R.J. (1999). "Guest Editor's Introduction: Machine Learning and Natural Language", Machine Learning, Vol.34 (Special Issue on Natural Language Learning), pp.5-9, Kluwer Academic Publishers.
9. Charniak E. (1993). "Statistical Language Learning", Cambridge, MA, MIT Press.
10. Cole R. A., Mariani J., Uszkoreit H., Zaenen A., Zue V., Varile G. B., Zampolli A. (eds.). (1996). "Survey of the State of the Art in Human Language Technology", (http://cslu.cse.ogi.edu/HLTsurvey/).
11. Dermatas E., Kokkinakis G. (1995). "Automatic Stochastic Tagging of Natural Language Texts", Computational Linguistics, Vol.21, pp.137-164.

12. Joshi A. (1995). “Some Recent Trends in Natural Language Processing”, in Zampoli A., Calzolari N., Palmer M. (eds.) *Current Issues in Computational Linguistics: In Honor of Don Walker*, Kluwer Academic Publishers, pp.491-501.
13. Kameda H., Sakurai T., Kubomura C. (1999). “Unknown Word Acquisition System for Japanese Written-Language Document”. In [27].
14. Kodratoff Y. (1999). “About Knowledge Discovery in Texts: A Definition and an Example”. In [27].
15. Levinson D. (1999). “Corpus-Based Method for Unsupervised Word Sense Disambiguation”. In [27].
16. Megyesi B. (1999). “Brill’s PoS Tagger with Extended Lexical Templates for Hungarian”. In [27].
17. Nagao M., Nakamura J. (1982). “A Parser which Learns the Application Order of Rewriting Rules”, in Proc. of the 9th International Conference on Computational Linguistics (COLING), Prague, 1982, North-Holland Publishing Co., pp.253-258.
18. Orphanos G., Kalles D., Papagelis T., Christodoulakis D. (1999). “Decision Trees and NLP: A Case Study in POS Tagging”. In [27].
19. Petasis G., Paliouras G., Karkaletsis V., Spyropoulos C. D., Androutsopoulos I. (1999). “Resolving Part-of-Speech Ambiguity in the Greek Language Using Learning Techniques”. In [27].
20. Sakurai S., Endo T., Mukai T., Oka R. (1999). “Automatic Task Modeling for Realizing a Multi-modal Interface System”. In [27].
21. Sparck-Jones K. (1994). “Natural Language Processing: A Historical Review”, in Zampoli A., Calzolari N., Palmer M. (eds.) *Current Issues in Computational Linguistics: In Honor of Don Walker*, Kluwer Academic Publishers, pp.3-15.
22. Stamatatos E., Fakotakis N., Kokkinakis G. (1999). “Automatic Extraction of Rules for Sentence Boundary Disambiguation”. In [27].
23. Stolcke A. (1995). “Efficient Probabilistic Context-Free Parsing”, Computational Linguistics, Vol.21, pp.165-202.
24. Thomas B. (1999). “Learning T-Wrappers for Information Extraction”. In [27].
25. Veenstra J. (1999). “Memory-Based Text Chunking”. In [27].
26. Wilms G. J. (1995). “Automated Induction of a Lexical Sublanguage Grammar using a Hybrid System of Corpus- and Knowledge-Based Techniques”, Ph.D. Dissertation, Mississippi State University, Department of Computer Science.
27. Proceedings of the Workshop on Machine Learning in Human Language Technology, Advanced Course on Artificial Intelligence (ACAI ’99), Chania, Greece, 1999 (http://www.iit.demokritos.gr/skel/eetn/acai99/Workshops.htm).