ML techniques and text analysis

Pieter Adriaans

Syllogic B.V.
Houten

Abstract

In this paper text analysis is presented as a special subdiscipline of automated language learning, which in itself is a subdiscipline of machine learning. A formal classification scheme for analysis of language learning algorithms in terms of abstract learners and speaker/authors is introduced. The inductive inference approach of Gold and successors is rejected as being of little practical value. The perspectives of this newly emerging field are discussed in the light of a number of exemplifying research projects.

1 Introduction

Although the ideal of a completely transparant natural language interface to a computer is still way out of reach, there is an abundance of interesting applications of ML techniques to text analysis. People are producing more and more texts at increasing speed. It is impossible to read everything. Therefore the need for automatic text analysis is growing rapidly.

The field of machine learning of language has witnessed substantial growth in interest and results in the past few years. Machine Learning techniques are in principle very useful in the context of language learning. Yet language learning has special problems of its own, that are not in the focus of interest of most researchers in the ML community, e.g.:

- The special algebraic structure of linguistic samples
- The highly structured and complex nature of language, and in particular the supposed irregularities, synonyms, metaphors etc.
- The complex interplay between the partial information about syntax and the lack of definition in the semantics of the samples.
- Special biases concerning the 'cooperativeness' of the author or speaker

These aspects call for another approach incorporating different algorithms, different complexity measures and different sampling techniques. At the moment contributions to this field tend to be scattered over various subfields (ML, AI, linguistics, psychology etc.). In this paper we will try to give an impression of this newly emerging field.

Before we proceed we wish to make two disclaimers. In the first place it is not possible to give a complete picture of the field within the limited space available. We will only touch upon developments that we think are important from our own limited point of view. Given the lack of a clearly defined research community it is possible that we overlook developments that are interesting. The fact that some research is not mentioned does not mean that we do not consider it to be important since the projects we discuss are only exemplifying. In the second place the explicit focus on text analysis does not imply that we are not interested in speech analysis. By excluding speech we hope to make life easier. There is an abundance of interesting applications of ML techniques in Text analysis. Incorporating speech seems to be much harder, although there are indications that interesting forms of cross fertilization between these two fields are possible.

The field of language learning has long been neglected in traditional linguistics. Learnability is an important criterion to judge the validity of any grammatical formalism aiming at explaining structures in natural language. One of the reasons interesting NL applications are still out of reach is the knowledge acquisition bottleneck in the definition of grammars and lexicons. ML techniques are beginning to be used to alleviate this problem.

1.1 A formal model

In [3] a formal model to study various forms of automated language learning is proposed. Language learning is seen as a process that involves at least two agents: a learner and a speaker (or author). Their interaction can be described in an operational setting in terms of rights and obligations in a language game. An abstract speaker/author is a mathematical formalisation of a certain kind of linguistic behaviour. It is a model of performance. In the general we can say that we can learn a language if we can deduce competence from performance. Consequently it is not possible to say something about the learnability of a language per se. We can only investigate learnability in relation to a certain kind of behaviour of a dialogue partner: i.e. a 'teacher' or an abstract opponent. Learnability therefore is not an abstract quality of a language. Our ability to learn a language may vary with ability or willingness of our teacher to adapt his behaviour. Learning a language is equal to getting to understand a speaker. We have learned (c.q. we understand) a language when we can predict (or imitate) the behaviour of a speaker based on information about his performance so far. The problem of learning a language is reduced to that of learning behaviour by observing behaviour.

The interactions in a pure linguistic rational dialogue presuppose that we are able to analyze the meaning of new concepts introduced by our opponent in pure operational terms. To do this we introduce the concept of an abstract speaker/author. An abstract speaker/author is a formal model of a speaker or author who is defending certain views. Formally an abstract speaker is a set $A = < D, L, S, E_P, O >$ where:

- D is a lexicon,
- L is a language which is a subset of the Kleene closure $D*$ of D,
- S is the set of true sentences of L,
- E_P is an examples routine that produces an element of S according to a mode of presentation P when called and
- O is an oracle routine that tells us whether or not a certain element of $D*$ is member of L and S.

The mode of presentation P is a formal notion corresponding to teacher behaviour. Various approaches to automated language learning can be modeled in relation to various definitions of abstract speakers. Different levels of supervision can be interpreted as different possibilites to influence the examples routine. An extreme form of unsupervised learning just discards the oracle routine. Different modes of presentation can be analysed to estimate their effect on the learnability of a language. This brings us to the following definition:

Learnable Abstract Linguistic Behaviour (LALB). Let $A = < D, L, S, E_P, O >$ be an abstract speaker of a language with a finite syntax and semantics. The behaviour of A is effectively learnable if there exists an algorithm that in time polynomial to a description of the language L and its semantics S constructs a function F that (with high probability) correctly predicts the behaviour of the oracle O.

It is clear that the learnability of behaviour in this sense depends on the mathematical properties of the lexicon D, the syntactic structure of the language L, the regularity of the semantics S and the mode of presentation P of the examples routine E. There is a whole four dimensional spectrum of possible learning situations. The definition of LALB can be relaxed to syntactic learning when we consider examples routines that produce wellformed sentences instead of only true sentences.

The formal concept of an abstract speaker/author can be used to model various interesting real life projects. It is not possible to go into formal details but examples of interesting applications are:

- Semantic and syntactic disambiguation of texts

- Text search algorithms for free text databases
- Automated document classification
- Automatic information/data extraction from text
- Adaptive NLP systems
- Automatic creation of dictionaries
- Automatic analysis of bi-lingual corpora
- Automatic creation of indexes
- Automatic acquisition of grammar rules

In the third part of this paper we will see that a number of these projects are currently being investigated.

1.2 Early attempts: Distributional analysis

The fact that it is fairly easy to recognize interesting syntactical patterns in texts using statistical techniques has been rediscovered over and over again ever since people began to analyse text with computers. Attempts have been made to construct algorithms that generate phrase structure grammars from the analysis of plain text. One early approach, advocated by Lamb, uses the distributional analysis of Harris and Hockett [15,16,17,19]. One defines phrase categories by associating phrase structures that are found in the same context. Contexts that share the same categories are considered to be equivalent. Complex phrase structure categories can be constructed by concatenation of simple categories. This approach has been criticised by Gold, who showed that even for the class of context free languages it is impossible to construct a learning algorithm that can learn a grammar from arbitrary free text. This means that we cannot learn a context free grammar by just listening to a speaker. The crux of the argument lies in the possibility to construct a teacher that for any learning algorithm constructs an example set that forces the learner to make an infinite number of wrong guesses. For a long time the criticism of Gold paralyzed serious research efforts in the field of automated text analysis so it is necessary to take a closer look at his results.

1.3 Gold's critique on unsupervised text analysis

In his landmark paper in 1967 Gold introduced the concept of identification in the limit [14]. The paper is important because it gives a conceptualization of the language learning problem that is mathematically feasible. First we specify a set of languages. Each language is taken to be a set of strings on the same finite alphabet. A teacher chooses a language from this set and a method of presentation. Every string in the language may only be presented once. The learner gets complete information if he gets positive as well as negative examples. Positive information means that he only gets positive examples. This equals the presentation of the language as plain text. The learning session starts at a certain moment of time t_0 and continues for ever. At each time t_n the learner receives a unit of information and is to make a guess as to the identity of the unknown language on the basis of the information received from the teacher sofar. The class of languages will be considered learnable with respect to the specified method of information presentation if there is a learning algorithm the learner can use to make his guesses and that has the following property: Given any language of the class there is some finite time after which the guesses will all be the same and they will be correct. One of the striking conclusions of this research was that context-free languages can not be identified on the basis of positive examples alone.

1.4 An evaluation of Gold's critique

What is the value of Gold's theories? We believe that the importance of Gold's work lies in the fact that he gave the theory of language learning a firm basis in recursion theory. His results however have a very limited practical value. They indicate extreme borderlines that only exist if we have infinite learning time. The concept is too abstract to give us any indication for the development of effective learning algorithms. The whole dimension of a systematic analysis of teacher behaviour and teacher-student interaction simply is not caught by the model of learning by enumeration. Therefore the most important question of finding effective practical heuristic strategies for language learning is not covered by Gold's research. It is exactly this dimension that we are interested in.

2 New developments from ML perspective

Although Gold's approach led to a number of interesting developments in the theory of inductive inference (See [6]) there has been very little progress in terms of practical applications of automated language learning. In the past few years however we have seen a number of developments that create a more promising perspective for language learning:

- Hardware with more power, e.g. cheap workstations and (massively) parallel architectures

- Better software tools, e.g. the general recognition of Prolog as implementation environment

- New ML paradigms, such as connectionist approaches, ILP and EBL

- New theoretical developments, such as Kolmogorov complexity and related concepts like the a priori probability of binary strings
- The availability of large text corpora on electronic media

Partly these developments are the same as those the ML community in general has benefited from, but there also has been a special impact on language learning research. In the following we will give an overview of some exemplary results.

2.1 Some current research

2.1.1 Inductive Inference, Complexity Theory and Information Compression

Some authors have deepened Gold's results in a more practical direction by formulating natural constraints that make certain types of grammar effectively learnable in certain circumstances. Theoretical results concerning constraints that make context-free languages effectively learnable are reported by Abe and Yokomori [33,1,2], although Abe's locality constraint seems rather artificial. Adriaans [5] has formulated 'Naturalness Constraints' that make context-free languages effectively learnable when sampling under the universal Solomonoff-Levin distribution [22].

In general minimum description length theory seems to be a promising approach to language learning. Powers for instance suggests that an unsupervised black-box approach based on information compression, could be most suitable for learning many structural properties exhibited by natural language [27].

In linguistic circles ideas concerning lexical cohesion of texts are beginning to emerge which suggest interesting approaches to automated language learning [25]. Also statistical approaches to parsing using socalled 'stochastic Context-Free Grammars' are promising in this respect [18].

2.2 Explanation-Based Learning (EBL)

Stan Matwin and Stan Szpakowicz (University of Ottawa) investigate methods to extract knowledge from expository texts. In such texts, examples are often introduced to show how to assemble rules acquired from the text into an operational concept or procedure. They apply EBL to accomplish this automatically [23]. Rey-Long Liu and Von-Wun Soo present a new language acquisition model to acquire parsing related knowledge via an EBL approach [21]. The domain theory in the model consists of two parts: a static part and a dynamic part. The static part consists of the universal linguistic principles proposed in the Generalized Phrase Structure Grammar (GPSG) formalism, while the dynamic part contains the context-free grammar rules as well as syntactic and thematic features of lexicons. In parsing both parts work together. Asker et al. describes a method for automatic lexical acquisition that extends an existing lexicon that, in addition to ordinary lexical entries, contains prototypical entries for various non-exclusive *paradigms* of open-class words [7]. This is done by reasoning about the constraints places on the unknown words in a text by *phrase templates* that have been abstracted from the grammar and domain specific texts using an explanation based learning method [29].

2.3 Connectionist approaches

Wellknown is the early research of Rumelhart and McClelland on learning the past tenses of english verbs [28]. St. John and McClelland present a parallel distributed processing model that learns to comprehend single clause sentences. The learning procedure allows the model to take a statistical approach to solving the bootstrapping problem of learning the syntax and semantics of a language from the same data [30]. In 1991 'Machine Learning' devoted a special issue to connectionist approaches to language learning [31], with special contributions on learning from ordered examples, inferring graded state machines and grammatical structure. The practical value of the proposals seems however limited.

2.4 Linguistic approaches

Valardi et al. observe that a poor encoding of the semantic lexicon is the bottleneck of many existing systems. To overcome these problems they propose an algorithm to learn syncategoremattical concepts from text examplars. Their knowledge acquisition method is based on learning by observations from examples of word co-occurrences (collocations) in a large corpus, detected by a morphosyntactical analyzer. Interactive human intervention is required in the training phase [32].

3 Conclusions: Directions for the future

These examples make clear that in different areas people from different disciplines are working on problems that have a close connection. It is the purpose of the workshop to bring these people together.

We conclude with a list of open problems that have to be solved before practical applications of ML in text analysis can be realized:

- Multi-layer learning, the combination of information on letter-, word-, sentence- or paragraph level in the learning process
- Ergonomic aspects of Text Analysis applications, i.e. batch analysis vs. on-line support
- Tractable practical complexity measures for various types of texts, i.e. a 'learnability' taxonomy

References

1. Abe, N., *Polynomial Learnability and Locality of Formal Grammars*, in Proceedings of the *26th Annual Meeting of the Association for Computational Linguistics*, ACL, 7-10 June 1988.

2. Abe, N. *Polynomial Learnability of Semilinear Sets*, Unpublished manuscript, 1988.

3. Adriaans, P.W., *A Domain Theory for Categorial Language Learning Algorithms*, in Proceedings of the Eighth Amsterdam Colloquium, P. Dekker and M. Stokhof (eds.), University of Amsterdam, 1992.

4. Adriaans, P.W., *Bias in Inductive Language Learning*, Proceedings of the ML92 Workshop on Biases in Inductive Language Learning, Abderdeen 1992.

5. Adriaans, P.W., *Language Learning from a categorial point of view*, Diss. Universiteit van Amsterdam, 1992.

6. Angluin, D., and C.H. Smith, *Inductive inference: Theory and Methods*, Computing Surveys, Vol 15, No. 3, pg. 237-269, 1983.

7. Asker, L., C. Samuelsson and B. Gambäck, *EBL^2: An approach to automatic Lexical Acquisition* in Proceedings of the *14th International Conference on Computational Linguistics (COLING '92)*, Nantes, France.

8. Brent, M.R., *Automatic Acquisition of Subcategorization Frames from Untagged Text*, in Proceedings of the *29th Annual Meeting of the Assocation for Computational Linguistics*, pg. 209-214, ACL, 18-21 June 1991.

9. Brill, E., *Discovering the Lexical Features of a Language*, in Proceedings of the *29th Annual Meeting of the Association for Computational Linguistics*, pg. 339-340, ACL, 18-21 June 1991.

10. Cottrell, G.W., *A Connectionists Approach to Word Sense Disambiguation*, Morgan Kaufmann, 1989.

11. Daelemans, W. and D. Powers (eds.), *Background and Experiments in Machine Learning of Natural Language*, in *Proceedings First SHOE Workshop*, ITK, Tilburg University, 1992.

12. Dagan, I., *Lexical Disambiguation: Sources of Information and their Statistical Realization*, in Proceedings of the *29th Annual Meeting of the Association for Computational Linguistics*, pg. 341-342, ACL. 18-21 June 1991.

13. Finch, S. and N. Chater, *A Hybrid Approach to the Automatic Learning of Linguistic Categories*, University of Edinburgh, 1991.

14. Gold, E. M., *Language Identification in the Limit*, in *Information and Control*, Vol. 10, pg. 447-474. Academic Press, Inc., 1967.

15. Harris, Z.S., *Methods in Structural Linguistics*, Univ. of Chicago Press, Chicago, 1951.

16. Harris, Z.S., *Distributional Structure*, in *The Structure of Language*, J.A. Fodor and J.J. Katz (eds.), Prentice Hall, New York, 1964.

17. Hocket, C.F., *A Course in Modern Linguistics*, Macmillan, New York, 1958.

18. Jelinek, F. and J.D. Lafferty, *Computation of the Probability of Initial Substring Generation by Stochastic Context-Free Grammars*, in *Computational Linguistics*, Vol. 17, No. 3, pg. 315-324, MIT Press, 1991.

19. Lamb, S.M., *On the mechanization of syntactic analysis*, in *1961 Conference on Machine Translation of Languages and Applied Language Analysis*, (National Physical Laboratory Symposium No. 13), Vol. II, pg. 674-685, Her Majesty's Stationery Office, London, 1961.

20. Last, R.W., *Artificial Intelligence Techniques in Language Learning*, Ellis Horwood, 1989.

21. Liu, R.-L. and V.-W. Soo, *Augmenting and Efficiently Utilizing Domain Theory in Explanation-Based Natural Language Acquisition*, in *Machine Learning*, Proceedings of the Ninth International workshop (ML92), D. Sleeman and P. Edwards (Eds.), Morgan Kaufmann Publishers, San Mateo, 1992.

22. Li, M. and P.M.B. Vitányi, *Learning simple concepts under simple distributions*, SIAM Journal of Computing, pg. 911-935, 1991.

23. Matwin, S., and S. Szpakowicz, *Machine Learning Techniques in Knowledge Acquisition from Text*, in *Think*, Vol. 1, No. 2, pg. 37-50, ITK, 1992.

24. McClelland, J.L. and A.H. Kawamoto, *Mechanisms of Sentence processing; Assigning roles to Constituents of Sentences*, in *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, Volume 2: Psychological and Biological Models, J.L. McClelland , D.E. Rumelhart, and the PDP Research Group, MIT Press, Massachusetts, 1988.

25. Morris, J. and G. Hirst, *Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text*, in *Computational Linguistics*, Vol. 17, No. 1, pg. 21-48, MIT Press, 1991.

26. Powers, D.M.W. and W. Daelemans, *The extraction of hierachical structure for Machine Learing of natural language*, in Daelemans, W. and D. Powers (eds.), *Background and Experiments in Machine Learning of Natural Language*, in *Proceedings First SHOE Workshop*, ITK, Tilburg University, 1992.

27. Powers, D.M.W., *A Basis for Compact Distributional Extraction*, in *Think*, Vol. 1, No. 2, pg. 51-63, ITK, 1992.

28. Rumelhart, D.E. and J.L. McClelland, *On Learning the Past Tenses of English Verbs*, in *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, Volume 2: Psychological and Biological Models, J.L. McClelland, D.E. Rumelhart and the PDP Research Group, MIT Press, Massachusetts, 1988.

29. Samuelsson, C. and M. Rayner, *Quantitative evaluation f Explanation Based Learning as an Optimization Tool for a Large-scale Natural Language System*, in Proc. of the *12th International Joint Conference on Artificial Intelligence*, Sydney, Australia, pp 609-615.

30. St. John, M.F. and J.L. McClelland, *Learning and Applying Contextual Constraints in Sentence Comprehension*, in *Artificial Intelligence*, Vol. 46, No. 1-2, pg. 217-257, North-Holland, 1990.

31. Touretzky, D.S. (ed.), *Machine Learning, Special Issue on Connectionist Approaches to Language Learning*, Kluwer Academic Publishers, Vol. 7, no. 2/3, 1991.

32. Velardi, P., et al (eds.), *How to Encode Semantic Knowledge: A Method for Meaning Representation and Computer-Aided Acquisition*, in *Computational Linguistics*, Vol. 17, No. 2, pg. 153-170, MIT Press, 1991.

33. Yokomori, T., *Learning Context-Free Languages Efficiently, A Report on Recent Results in Japan*, in K.P. Jantke (ed.) *Proceedings Int. Workshop Analogical and Inductive Inference*, J. Siekmann (ed.), Lecture Notes in Artificial Intelligence, no. 397, pg. 104-123, October 1-6, 1989.