The Challenge of Discovering Meta-Data

Katharina Morik and **Stefan Haustein**
Univ. Dortmund, Computer Science VIII,
D-44221 Dortmund, Germany
{morik, haustein}@ls8.cs.uni-dortmund.de

Introduction

Machine learning research has always been driven by scenarios. The scenarios were put together from ideas of anthropological disciplines (e.g., linguistics, cognitive and social science, biology) with reference to different theoretical frameworks (e.g., logic, recursion or complexity theory) and were illustrated by challenging applications. The stimulating triple of anthropological, theoretical, and application-oriented aspects led to the development of new algorithms. New applications were investigated. Then, what appeared to be a large set of diverse algorithms with applications finally turned out to be one learning task – with several parameters or choices – in the light of theoretical analysis. This happened to language learning or function identification in the framework of identification in the limit and it happened to concept learning in the framework of probably approximately correct (PAC) learning.[1]

Now, we might be satisfied and just exploit and scale-up the results of a mature field. However, if we compare the stimulating origin of language learning with the theoretically well-based approaches to learning pattern languages (e.g., (Angluin 1980; Angluin & Smith 1983), (Lange & Wiehagen 1991)) or context-free grammars (e.g., (Berwick & Pilato 1987)), we might feel disappointed. First, the approaches cannot be fed back to insights or observations of linguistics as could well be the approaches towards concept learning to cognitive science (e.g., (Mechelen *et al.* 1993), (Mühlenbrock & Morik 1998), (Neri, Saitta, & Tiberghien 1997)). Second, the challenging applications are not solved by today's sophisticated methods. Hence, language learning keeps its high position on our research agenda. New scenarios may build upon some common ground. The capability of learning a language – be it a "linguistic organ" or a function of (inter-)actions – combines

- the recognition of syntactic structures and
- the formation of semantic categories with
- social actions.

[1]It has not yet happened to reinforcement learning.

With respect to natural language, new scenarios are currently under development that refer to other theoretical frameworks, namely logic ((Cussens *et al.* 1997), (Cussens 1999)) or Kolmogorov complexity (Adriaans 1992). An evolutionary approach is followed by (Steels 1997; Steels & Kaplan 1999). With respect to applications, the daily expanding set of WWW pages as well as the increasing demands of human-computer interfaces demand a variety of learning tasks to be solved. Text classification that recognizes a document as member of a category of topics can be handled as a particular concept learning task (e.g., (Joachims 1998; 1999), (M.Craven, Slattery, & Nigam 1998; Freitag 1998)). This task abstracts away the syntactic structures by the bag of words representation as well as the structure of categories by a set of given, alternative categories. Relational approaches do not change the learning task although they cover some aspects of the structure of linked hypertexts (Cohen 1995; M.Craven, Slattery, & Nigam 1998)). Text classification is connected with social actions if software agents are to use the WWW as their source of information. Until now, however, the activities of the agents are not the driving force of learning how to classify texts nor is their ativity used by the learning algorithm. Hence, text classification does not belong to the new tasks that share charactistic features with the task of language learning.

In the following, we propose a scenario of new learning tasks, which take into account syntactic structures, semantic categories, and social actions. These are not the ones of natural language and human interaction, though. The scenario restricts the investigation of language learning capabilities to formal languages, the interactions to those of agents or software systems. The scenario is motivated by applications. It does not yet refer to a particular theoretical framework, although contributions of the field of parallel computation (p-complete theory) can be expected (e.g., (Greenlaw, Hoover, & Ruzzo 1995; Leighton 1992)) or communication complexity (Kushilevitz & Nisan 1997). We consider the scenario a step towards one of the most challenging long-term goals of research, language learning.

The Scenario of Learning How to Communicate

Modern communication facilities allow access to diverse services, from the housing server (controling, e.g., heating, the refrigerator check list, and the coffee machine), the personal time scheduler, the electronic mail server to databases to the WWW as the distributed document server. A person interacts with these different servers using the mobile phone or any computer at any place. Although the protocols for short and long distance (blue tooth, WAP) are not yet established, we may already investigate a scenario in which a user communicates with several very different servers.

In a restricted sense, softbots or other agent systems already face the task to assist a user in accessing distributed information sources (e.g., (Perkowitz & Etzioni 1995), (Maes 1994; Sheth & Maes 1993)). The user hands over a task to the personal agent which itself may send out several working agents to perform subtasks. For instance, such a task may be to find the cheapest offer for holidays on a farm that is still available at school holidays. The personal agent structures this task into the subtask for looking up school holidays (calendar agent), finding relevant WWW-sites (information gathering agent), finding price information within them (information extraction agent), recognizing the form sheets for booking questions (a specialised agent), filling in the form sheet and sending it to the respective adress (form-filling agent), and looking in the inbox for answers concerning the availability (a specialised filtering agent and a message understanding agent). The working agents deliver the results to the personal agent which combines them and puts together a set of cheapest and available holiday offers. The agents can fulfil their tasks (more) easily, if the know the representation formats of the systems which they access. They have to parse the structure of a document or table in order to get at the needed information.

In this example, a calendar system, an electronic mail system, and the WWW considered as one large and distributed server are used. The different systems offer different formats. For looking up the school holidays, the agent should know the meaning of columns and rows – which indicates the relevant federal state and which the season[2]? For filtering the inbox, it is useful to know about "To:", "From:", MIME formats and path information. Looking more closer into the WWW, we still see the very restricted HTML format where annotations are mostly presentation-oriented instead of application-oriented. We could want to learn a mapping from the presentation-oriented tags (e.g., bold, paragraph, empty line) to more structural annotations (e.g., title, section heading). Approaches to learning the document structure with respect to the presentation have been put forward by (Esposito, Malerba, & Semeraro 1994). Currently, with XML there is the chance to come back to the advantages of SGML. A document type definition (DTD) indicates the categories to be used in a document of a certain class. Although the same mark-up language (XML) is used, we then find a large variety of formats corresponding to diverse document classes. It would be extremely useful to know the DTD for each document. If we know the DTD and it corresponds to the one used by the agents, they could look up the desired information on the basis of the meta-information. This is the starting point of the ONTOBROKER, a system that extracts ontologies from the web (Decker *et al.* 1999). There, the issue of seeking information in distributed and semi-structured document repositories is investigated, but the DTD is known and corresponds to the DTD expressing the ontology. Machine learning is not yet applied. In contrast, we assume (marked-up) texts where we don't know the underlying DTD. We want to solve the access problem by a learning approach. This is a *syntax learning task*, namely to learn a DTD from documents that use it.

Even if we know the DTD of a document, the real problem is still unsolved, namely, the meaning of the tags is unclear. The personal agent of a user may have a DTD which reflects the user's view and is the basis of the commmunication between the personal and the working agents. This sets up a *translation learning task*: learning the translation between the DTD of the user's multiple agent system and the DTDs of external documents. By mapping tags of external pages to the "own" tags, the meaning becomes clearly established. Note, that it is a many-to-many mapping, where possibly several tags of one DTD correspond to a combination of tags of the other DTD. The task of learning a mapping of two different formats can be illustrated by an instance outside the WWW. We have developed a system that translates bibliography styles on the basis of examples: the same reference to a paper is given in two styles and the contential fields together with their different orderings in the styles is learned (Mansfeld 2000). Learning the contential fields such as author, title,... is the task of *forming semantic categories*. The translation of bibliographic styles cannot be learned by gap-pattern approaches (Nix 1985), because these require each example to have the same structure. References to literature vary a lot (number of authors, structure of the title, editors, journals or books, etc.). The task is complex, because no negative examples are available (Gold 1967), the structure differs from example to example but cannot be generalised away, and the semantic categories (fields) are to be formed. Hence, we must ease the format translation task. One option is to use an oracle (i.e., to ask the user). We have chosen this option for the bibliography style translation. Another option is to know one of the formats. An example is then a format, a text with annotations of that format, and the same text with the annotations of the unknown format. This eases the task of category formation within the translation task. We choose this option for the translation of an external unknown DTD into

[2] In Germany, each of the four seasons has a holiday period, but the particular dates vary from state to state.

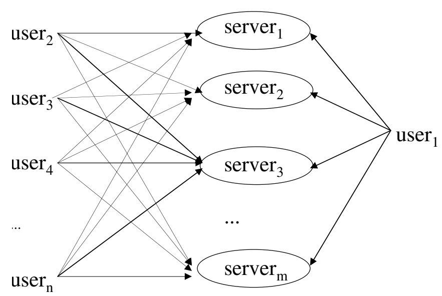

Figure 1: The Communication Environment

the "own" DTD.

The translation learning task also occurs in the context of knowledge discovery in databases. The mapping from one database schema to another one is important if we want to re-use successful chains of preprocessing or data mining steps. If, for instance, we have successfully found patterns of warranty cases in the database of one large car producer, we might want to follow the same procedure for another car producer. Actually, the two databases store the same kind of information. However, the database schemata are completely different. In contrast to the bibliographic translation, the particular entries in the databases are not the same: different cars are sold to different customers. A translation from one database schema to the other allows to adapt the successful case to the new database. The crucial point is the formation of semantic categories that correspond to the use of attributes for learning. A seasonal attribute, a sequence that is given by several attributes (e.g., sales date, date of warranty claim,...), a product characterization, and a customer characterization are all semantic categories that determine how to form examples for further learning steps. Whereas the recognition of syntactic structures is obsolete – the database schemata are given – the category formation is decisive.

We have now considered a user (his or her multi-agent system) acessing diverse servers. Turning it the other way around, a server has to fulfill the demands of many users (their agents). Again, there is the need of learning how to communicate: the server must learn how to answer requests of different users or agents. Information requests can be complex and answering them can require a series of learning steps. Think, for instance, of a server storing insurance data. The users can be sales agents who want to look for potential customers of a certain product in their district. Their query could be: "What is the characteristic of customers of a life risk insurance? Return the addresses of people in my district with this characteristic!" Other users are the potential customers themselves who want to be informed about prices and conditions of various insurances. Other users could be the insurance managers who want reports on various aspects of the sales development. The users input their questions by filling in forms. The server generates these interactive forms on the basis of the meta-data about the database. SQL queries or data mining runs are generated for answering the user queries. In other words, we consider knowledge discovery in databases a sophisticated query language.

Another topic of answering queries of diverse users is the presentation of the results. This becomes particularly important, if the mobile phone or another very limited display is used. The menues of items must be organised such that for each user the most frequent choices are seen without scrolling. Adapting the presentation of information to individual users is a learning task in its own right. Jeff Schlimmer and colleagues have, for instance, investigated such a task (Schlimmer & Hermens 1993). For a hospital information system, (Bullerdick & Ritthoff 1999) have learned a grammar that is used by a form sheet generator for entering diagnoses into a medical database. We may want this adaptation be performed at the client (the agent system of the user) or at the server. No matter whether the intelligence is at the server or at the accessing agent, the problem remains the same: the agent or the server must learn how to communicate.

Learning tasks in our scenario are all based on the communication needs of different agents. Syntactic structures and semantic categories are anchored in a kind of "social actions", i.e. in collaboratively achieving a goal of the user and interacting with distributed information sources such as the WWW. This embedding into actions of agents and servers allows to validate learning results with respect to their effectiveness. Accuracy is no longer the ultimate criterion. For instance, it may well be that the learned syntactic structure differs from the target syntax, but this does not matter as long as the task can be solved using the inaccurate grammar. This is similar to natural language, where we never know whether all native speakers do in fact use the same grammar of the language. Bertrand Russell has finished the philosophical debate about this by the simple statement: "I can order a meal using the language and I get it!"

Conclusion

The scenario of accessing systems, each with its own representation, offers learning tasks that are linguistical by nature while restricting them to formal languages. The recognition of syntactic structures and the formation of semantic categories are placed in the context of tasks (goals) of various agents. The need for translations is given by the distributed computing of very many different users. Hence, the three characteristics of language learning are given, although it is not natural language which is investigated.

References

Adriaans, P. 1992. *Language Learning from a Categorical Perspective.* Ph.D. Dissertation, Univ. Amsterdam.

Angluin, D., and Smith, C. 1983. Inductive inference: Theory and methods. *Computing Surveys* 15:237–269.

Angluin, D. 1980. Inductive inference of formal languages from positive data. *Information and Control* 45:117–135.

Berwick, R. C., and Pilato, S. 1987. Learning syntax by automata induction. *Machine Learning Journal* 2:9–38.

Bullerdick, K., and Ritthoff, O. 1999. mAGENTa – Ein Internet-Agent zur automatischen Generierung von graphischen Benutzeroberflächen mittels maschineller Lernverfahren am Beispiel radiologischer Befunde. Diplomarbeit, Fachbereich Informatik, Universität Dortmund.

Cohen, W. W. 1995. Text categorization and relational learning. In *Macbine Learning: Proceedings of the Twelfth International Conference (ICML '95)*. Lake Taho, CA: Morgan Kaufmann.

Cussens, J.; Page, D.; Muggleton, S.; and Srinivasan, A. 1997. Using inductive programming for natural language processing. In Daelmaens, W.; Weijters, T.; and van der Bosch, A., eds., *Workshop Notes on Empirical Language Processing Tasks*, 25–34.

Cussens, J., ed. 1999. *Learning Language in Logic*, Workshop Notes.

Decker, S.; Erdmann, M.; Fensel, D.; and Studer, R. 1999. Ontobroker: Ontology based access to distributed and semi-structured information. In Meersman, R., and other., eds., *Semantic Issues in Multimedia Systems, Kluwer Academic Publisher, Boston, 1999*. Boston: Kluwer Academic Publisher.

Esposito, F.; Malerba, D.; and Semeraro, G. 1994. Multistrategy Learning for Document Recognition. *Applied Artificial Intelligence: An International Journal* 8:33 – 84.

Freitag, D. 1998. Information extraction from html: Application of a general learning approach. In *Procs. 15th National Conference on Artificial Intelligence AAAI-98*.

Gold, E. M. 1967. Language identification in the limit. *Information and Control* 14:447–474.

Greenlaw, R.; Hoover, H.; and Ruzzo, W. L. 1995. *Limits to Parallel Computation: P-Completeness Theory*. Oxford University Press.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, 137 – 142. Berlin: Springer.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning (ICML)*.

Kushilevitz, E., and Nisan, N. 1997. *Communication Complexity*. Cambridge University Press. ISBN 0-521-56067-5.

Lange, S., and Wiehagen, R. 1991. Polynomial-time inference of arbitrary pattern languages. *New Generation Computing* 8:361–370.

Leighton, F. 1992. *Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes*. Morgan Kaufmann.

Maes, P. 1994. Modeling adaptive autonomous agents. *Artificial Life Journal* 1(1 & 2):135–162. MIT Press.

Mansfeld, E. 2000. Entwicklung eines Agenten zur Unterstützung der Umformatierung von Literaturhinweisen. Master's thesis, Fachbereich Informatik, Universität Dortmund. In German.

M.Craven; Slattery, S.; and Nigam, K. 1998. First-order learning for web-mining. In *Procs. 10th European Conference on Machine Learning*.

Mechelen, I. V.; Hapton, J.; Michalski, R. S.; and Theuns, P. 1993. *Categories and Concepts. Theoretical Views and Inductive Data Analysis*. Cognitive Science Series. London: Academic Press.

Mühlenbrock, M., and Morik, K. 1998. Rekonstruktion des Erwerbs einer Begriffsstruktur zur Erlärung des Tag/Nacht-Zyklus. *Kognitionswissenschaft* 2.

Neri, F.; Saitta, L.; and Tiberghien, A. 1997. Modelling physical knowledge aquisition in children with machine learning. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, 566–571. Stanford (CA): Morgan Kaufmann.

Nix, R. P. 1985. Editing by example. *ACM Transactions on Programming Languages and Systems* 7(4):600–621.

Perkowitz, M., and Etzioni, O. 1995. Category translation: Learning to understand information on the Internet. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI '95)*, 930–936. Montreal, Canada: Morgan Kaufmann.

Schlimmer, J. C., and Hermens, L. A. 1993. Software agents: Completing patterns and constructing user interfaces. *Journal of Artificial Intelligence Research (JAIR)* 1:61–89.

Sheth, B., and Maes, P. 1993. Evolving agents for personalized information filtering. In *Proceedings of the Ninth IEEE Conference on Artificial Intelligence for Applications*. IEEE Press.

Steels, L., and Kaplan, F. 1999. Bootstrapping Grounded Word Semantics. In Briscoe, T., ed., *Linguistic evolution through language acquisition: formal and computational models*. Cambridge University Press.

Steels, L. 1997. The origins of syntax in visually grounded robotic agents. In Pollack, M., ed., *Proceedings of IJCAI 97*. Morgan Kaufmann Publishers, Los Angeles.