Knowledge and Learning in Natural Language (Charles D. Yang)

Walter Daelemans (University of Antwerp, Belgium)

Introduction.
This review was written from the perspective of someone working in the field of Machine Learning of Natural Language: the application of techniques from statistical pattern recognition and Machine Learning to problems in Natural Language Processing (NLP). As in any subfield of Artificial Intelligence, the goals of this type of research can be either practical or theoretical. From the point of view of practice (language engineering), Machine Learning methods have been shown to allow the construction of more efficient, more robust, and more accurate NLP systems in a faster way than by handcrafting rules. The explanation for this is that these inductive systems are trained on large corpora of real language use, rather than being based on linguistic intuitions of one or a few (computational) linguists. On the other hand, the induced theories lack the linguistic sophistication of hand-made models. Theoretical results include an increased understanding which linguistic knowledge can be learned from primary linguistic data, which sources of knowledge are necessary for learning a particular NLP task, what distribution of effort and importance should be assigned to programmed knowledge and induced knowledge, and which learning algorithm has the right 'bias' to learn linguistic knowledge.

The research on the grammar competition model proposed by Charles Yang, while developed with different goals and from a different perspective, fits the theoretical goals of machine learning of natural language very well. His dissertation is an important contribution to work in the P&P framework, written in a clear and attractive style. In this review, I will focus on what results in Machine Learning and Computational Linguistics can contribute to a discussion of the issues he raised and the model he developed.

Framework and methodology.
The dissertation was written within the framework of the innateness of linguistic knowledge (more specifically the P&P approach), and spends some time in Chapter 1 arguing this position, especially by referring to the APS (Argument from the Poverty of Stimulus). There are several reasons, logical and empirical, why the APS is a weak argument for innateness of linguistic knowledge, an excellent overview and synthesis of these can be found in a recent dissertation by Alexander Clark (Clark, 2001), but it doesn't seem useful to go into this discussion in detail here as it is not the subject of the dissertation, and there are other, perhaps more convincing, arguments in favour of UG. It does seem useful, however, to point out that at least in computational linguistics, the APS is far from uncontroversial, and several systems have succeeded in learning linguistic structure from primary linguistic data. There is a lot of work in computational linguistics on syntactic category induction from distributional information with statistical pattern matching techniques (e.g. Brown et al. 1992; Schütze, 1997) and similar techniques for syntax acquisition (e.g. van Zaanen, 2000), making the argument that structure-dependence can only be learned with a priori knowledge doubtful. It is regrettable that

Yang, in his criticism on empiricist models (p. 18; p. 39-41) focuses only on neural network models and limited corpus-based argumentation, ignoring Computational Linguistics work using unsupervised machine learning techniques. The latter research clearly shows that syntactic structure can be induced from data.

As far as methodology is concerned, the success criterion Charles Yang sets himself is to measure up the hypothesis that "Child language in development reflects a statistical combination of possible grammars allowed by UG, only some of which are eventually retained when language acquisition ends. (p. 15)" against especially the criteria of formal sufficiency and developmental compatibility.

Variational Learning.

In Chapter 2, Yang makes a convincing case that the triggering model for parameter setting in a P&P framework (the most influential approach) does not succeed for the formal criterion because of problems with local maxima and noise sensitivity (lack of robustness). The approach also fails because of developmental incompatibility because it incorrectly predicts abrupt changes in child language as the learner moves from grammar tot grammar, and as it assumes consistency of child language with the current grammar.

Inspired by Darwinian evolutionary biology, Yang proposes an approach to language acquisition that is based on (the metaphor of) competition between grammars. Given a set of grammars from the space of possible grammars allowed by UG, a weight (probability) is associated with each grammar. Given an input sentence, a grammar is selected according to this probability distribution. If the grammar can be used to parse the sentence, its probability is increased at the cost of the probability of the other grammars, if the grammar can't parse the sentence, its weight is decreased at the advantage of all the other grammars. This simple reinforcement learning scheme can be shown to converge to a situation where grammars more compatible with the input data are better represented in the population of grammars. The extent to which a grammar is incompatible with the input data can be estimated by computing the relative frequency of sentences in a sample of the input data (e.g. the CHILDES child-directed language corpus) that cannot be parsed with that grammar. This error measure is called here the penalty probability.

But where do these grammars come from? Given realistic numbers of parameters, the search space of possible grammars is huge, and the approach becomes computationally intractable. The proposed solution to this problem is to represent grammars as bit strings of parameters. Parameters can be on or off, and a configuration of parameter settings can be translated into a grammar. Instead of applying the reinforcement learning to weights associated with grammars, it will be applied to weights associated with the parameters directly. If a sentence can be parsed by a grammar, all parameter settings get rewarded, if not they all get punished. This approach of course leads to the well-known problem of credit assignment in reinforcement learning: which parameter settings were really instrumental in the grammar succeeding in parsing the sentence, and which parameters simply "took a hitchhike" and got rewarded incorrectly?

Although a computer simulation of a very simple example shows that this credit assignment problem need not get the learner into problems, it seems to me that more advanced credit assignment algorithms are necessary to avoid the reinforcement learning getting trapped in local maxima. However, the author mentions work in preparation in which a formal proof (of convergence?) is promised. The huge literature on reinforcement learning and credit assignment, not mentioned in this dissertation, might help finding a solution. See e.g. Sutton (1998).

I think Yang has developed an interesting 'weak' (domain-independent) learning algorithm that improves tremendously upon the cue-based learning methods of the type of Dresher and Kaye (1990) where innate knowledge about cues for each parameter is hypothesised to make the learning work. At first sight, the algorithm also improves on the more traditional genetic algorithm (GA) approach of Robin Clark (1992) because it works incrementally instead of needing to parse a sample of input data for computing the fitness of each individual (a particular series of parameter settings) in the population. However, the advantage of the GA approach is that because of the way fitness-based selection, mutation and crossover operations cooperate, the credit assignment problem is solved. I therefore think Clark's approach is dismissed too easily and a thorough empirical and formal comparison should be made. For one thing, it isn't at all clear that the GA approach would need more parse actions than the reinforcement learning approach to achieve convergence given a reasonable number of parameters (rather than the limited numbers used in this dissertation). Particularly strange is the footnote on page 33 that the crossover operator used in Clark (1992) needs empirical justification, as crossover is a cornerstone of Darwinian evolution. Again, the author could have acknowledged the relevance of some of the enormous literature on GAs and Genetic Programming, some of which was applied to language learning as well (though not necessarily as a model of language acquisition), e.g. Smith & Witten (1996).

Developmental compatibility of Variational Learning.
The really groundbreaking contribution of this dissertation is not so much the originality of the developed learning algorithm (reinforcement learning on parameter vectors) and acquisition model, but the way in which it was used to make quantitative predictions about child language development. In Chapter 3, a detailed analysis of Null Subjects (NS) in English children is used to demonstrate how the variational learning method explains the empirical development data. At least for this example, the results of the analysis are convincing, and it is clear that this approach has the potential to add corpus-based quantitative reasoning and prediction to the tools of child language acquisition research. Somewhat less developed, yet still convincing is the discussion in Chapter 5 of how grammar competition can be used as an acquisition-based model of language change. Again, an illustration is given how the approach taken allows researchers to investigate language change in a quantitative, corpus-based way.

I will go into his analysis of irregular verb morphology proposed in Chapter 4 in somewhat more detail. Incidentally, this chapter takes up almost a third of the dissertation, which is rather inconsistent with his claim that the interest which the English

past tense has generated is unfortunate as it is a fairly marginal problem in linguistics (p. 47).

The proposed approach is to have different rules not only for the regular cases in English verb morphology (the default rule), but also for the different types of irregular verb morphology. E.g., a rule is proposed that associates words like feed and shoot with Vowel Shortening. These rules are in competition with each other. Where do these rules come from? They are somehow induced from the primary linguistic data with the help of UG constraints. Where does the default rule come from? It comes from the sensitivity of children to type frequency. How does the child assign verbs to rules? By considering the semantic relatedness of e.g. shoot and shot, the verb can be assigned to in this case the Vowel Shortening rule. The rarity of irregularization as opposed to regularization suggests an active role of the default rule as rule used in cases of lacking evidence.

Each of the rules has a weight (probability), and each assignment of a verb to an irregular class of verbs (associated with the same rule) has a probability. These probabilities are updated during learning with a method similar to the one described above.

The approach is explicitly targeting Pinker's Words and Rule (WR) approach, and succeeds, with its probabilistic approach to blocking, in improving upon it. Moreover, the model seems to provide a good fit with the available developmental data. It also predicts effects of systematic regularities in irregular verbs, which are hard to explain in the WR approach.

However, I strongly disagree with Yang's caricature of learning by analogy. Whereas Yang claims (p. 63) that there are no clear models of how analogy would be used in language learning, the Machine Learning literature is replete with exemplar-based, instance-based, and other analogical learning algorithms (See Aha, 1997 for an overview) which could be and have been used for modelling language acquisition and processing. Some of these algorithms (e.g. Skousen, 1989) have even been developed within linguistics, have been around for a very long time, and are actively being used to model (among many other phenomena) irregular morphology (e.g. Eddington, 2000). These exemplar-based models don't use any explicit rules, but nevertheless show ``rule-based'' behaviour because of the way the similarity metric interacts with the contents of memory (the density and homogeneity of regions in the mental lexicon where the similarity metric operates). Similarity metrics can be adapted using statistical and information-theoretic methods to the problem to be solved by means of feature weights and value difference metrics. This solves e.g. the problem that analogical methods would, according to Yang, not be able to prevent (frequent) irregularization. See Daelemans (1999) for pointers to the computational (psycho-) linguistics literature applying implemented analogical methods to language processing and acquisition problems.

Conclusion.

In conclusion, I think Charles Yang's dissertation is an important milestone in P&P-based theories of language learning, and Variational Learning deserves to be widely studied. For the first time, a general learning method is combined with a UG-based hypothesis

space into an acquisition model that seems to have largely the right formal characteristics and that improves upon earlier proposals. Especially interesting is the fact that the approach provides a new tool for the study of both child language development and language change in an objective, corpus-based, and quantitative way. In due time, I hope the relevance of related approaches in (statistical) machine learning, computational linguistics and genetic algorithms will be acknowledged also in this field, and used in the development of even more sophisticated and accurate models of language acquisition.

Aha, D. (editor) (1997). *Lazy Learning*. Dordrecht: Kluwer Academic Publishers.

Brown P., V. Della Pietra, P. deSouza, J. Lai, and R. Mercer (1992). Class-based n-gram models of natural language. *Computational Linguistics***18**, 467-479.

Clark, A. (2001). Unsupervised Language Acquisition: Theory and Practice. D.Phil. dissertation, University of Sussex.

Clark, R. (1992). The selection of syntactic knowledge. *Language Acquisition* **2**, 83-149.

Daelemans, W. (1999). Memory-Based Language Processing. *Journal for Experimental and Theoretical Artificial Intelligence* **11** (3), 287-296.

Dresher, E. & J. Kaye (1990). A computational learning model for metrical phonology. *Cognition* **34**, 137-195.

Eddington, D. (2000). Analogy and the dual-route model of morphology. *Lingua* **110**, 281-298.

Schütze, H. (1997). *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. Stanford: CSLI Publications.

Skousen, R. (1989). *Analogical modeling of language*. Kluwer: Dordrecht.

Smith, T. & I. Witten. (1996). Learning language using genetic algorithms. *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, edited by S. Wermter, E. Riloff, and G, Scheler, 132-146. Berlin: Springer Verlag.

Sutton, R. & A. Barto (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Van Zaanen, M. (2000). ABL: Alignment-based learning. *Proceedings of the 18th International Conference on Computational Linguistics*, 961-967.