Which Came First, the Grammar or the Lexicon?

Tom Armstrong[1] and Tim Oates[2]

[1] Wheaton College
Norton, MA 02766 USA
armstrong_tom@wheatoncollege.edu
[2] University of Maryland Baltimore County
Baltimore, MD 21250 USA
oates@cs.umbc.edu

Abstract. Computational approaches to learning aspects of language typically reduce the problem to learning syntax alone, or learning a lexicon alone. These simplifications have led to disconnected solutions and some unreasonable assumptions about inputs to their algorithms. In this paper, we present an approach that exploits a grammar learning algorithm to learn its own alphabet, or lexicon. We present empirical results and categorize the successes and types of errors lexical acquisition approaches encounter.

Keywords: lexical learning, applications of grammars, natural language learning.

1 Introduction

Many grammar learning algorithms derive inspiration from the distinctly human ability to learn language. While children are facile at learning an alphabet (i.e., words in the language composed on phonemes) and constructing a generative grammar using that alphabet, our computational approaches are often brittle and prone to make unrecoverable errors. Learning grammars is an intractable problem unless, for one, concessions are made regarding the input, and having complete knowledge of the language's alphabet is a common assumption. For example, most algorithms expect an input like *the cat hates the dog*, and not an input like *thecathatesthedog*.

This paper explores the utility of including higher-level structural information (in the form of a learned grammar) in the unsupervised learning of a lexicon. We remove the assumption that the grammar learning algorithms have perfectly segmented input data. We discuss this learning task in terms of the lexical-syntactic interface [1] where two learning tasks (i.e., lexical acquisition and grammar learning) are bootstrapped together. Here we extend our approach through experimentation with additional lexical data.

A. Clark, F. Coste, and L. Miclet (Eds.): ICGI 2008, LNAI 5278, pp. 283–285, 2008.

2 Lexical-Syntactic Interface

The lexical-syntactic interface is the interplay between the learning tasks of lexical acquisition and grammar induction. A typical lexicon learning algorithm begins with a stream of categorical data or a set of strings, and its goal is to induce an inventory of lexical items. A typical grammar induction algorithm begins with a set of strings, and its goal is to learn a generative structural model like the RPNI example above. While lexical learning is done without any regard for structural information, grammar induction assumes a known lexicon and correctly segmented input strings. In the lexical-syntactic interface, we exploit the structure inherent in the sequences of words and inside of words.

GramLex is the algorithmic instantiation of the lexical-syntactic interface in the form of a bootstrap algorithm [1]. We detailed the specific algorithmic components of the interface and presented experimental results on a variety of benchmark languages. The grammar learning community has a series of benchmark languages for comparing learning algorithms: $\mathcal{L}_1$ through $\mathcal{L}_{15}$ (Canonical deterministic finite automata and data are available from `http://www.irisa.fr/symbiose/people/coste/gi_benchs.html`) [2,3].

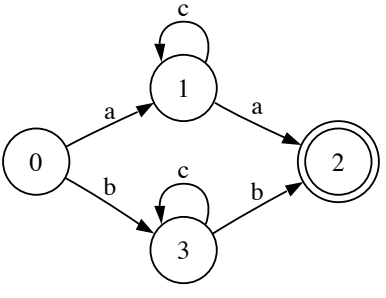

Fig. 1. Target Machine Topology

Let us look at an example of $\mathcal{L}_{15}$ with a random lexicon where GramLex learns a lexicon using a grammar learning algorithm. Using the three words a = *ow r ih n jh* (orange), b = *p er p ax l* (purple), and c = *r ey z ih n z* (raisins) in $\mathcal{L}_{15}$, we begin with Σ (the alphabet) and Γ (the initial lexicon) = {λ, *ow*, *r*, *ih*, *n*, *jh*, *p*, *er*, *ax*, *l*, *ey*, *z*}. Given a characteristic sample for this language, GramLex returns the automaton in Figure 1.

3 Experiments

While the grammar learning community has made an effort to evaluate algorithms empirically, it is less obvious that the lexicon learning community has done the same. Proper evaluation of our lexical and grammatical bootstrap is a challenge with respect to the lexicon we select. To begin, we choose a random collection of words found in the SWITCHBOARD corpus. Using the human-annotated phonetic transcriptions for each word, we provide the sequence of

phonemes for each word (an ARPAbet symbol for each phoneme) to the bootstrap in place of the standard alphabet in the characteristic sample. We run our bootstrap algorithm and analyze the resulting learned machine. The words vary in length from a maximum length of 15 phonemes (e.g., *eh k s t r ow r d ih n eh r ax l iy* or extraordinarily) to a minimum length of 2 phonemes (e.g., *m ey* or my) and collectively had a mean length of about 7 phonemes. Here we report the results on a trial of 50 randomly selected lexicons and the language $\mathcal{L}_{15}$. Of the 50 trials, 39 learned the correct lexicon and the correct grammar.

The remaining 11 trials that did not completely learn the lexicon or the grammar are categorized into four distinct classes of errors. Class 1 errors (2/11) are trials where we correctly learn the lexicon, but not the correct grammar. Class 2 errors (2/11) are trials where we correctly learn the automaton (with one caveat), and we correctly learn the lexicon (with one caveat). That is, the grammar accept all of the strings in the language, but also accepts a special overgeneralize-type string. Class 3 errors (5/11) are trials that fail to learn the entire lexicon and overgeneralize beyond the surface equivalence found in class 2 errors. While the machine contains some correct structure, the resulting machines further and further fail the *looks good* test. Class 4 errors (2/11) are the most egregious in terms of incorrectly learned structure and incorrect lexical items.

4 Conclusion and Future Work

In this paper, we presented an extension to our novel framework for bootstrapping the acquisition of a lexicon and learning a grammar. Prior work on lexical acquisition has ignored how those terms are used from a syntactic point of view, and grammar learning approaches typically require perfectly formed inputs to guarantee any learning result. This work demonstrates the viability of learning both tasks in tandem for a rich diversity of languages and lexicons.

Future work will proceed in two directions. First, we will focus on futher defining the boundaries between the learnable and the pathological for certain lexicons. Second, we will expand the class of languages we are interested in learning. Context-free grammars and natural languages may provide even more structure to guide lexicon learning and, in fact, make learning easier.

References

1. Armstrong, T., Oates, T.: Learning in the lexical-grammatical interface. In: FLAIRS Conference. AAAI Press, Menlo Park (2008)
2. Tomita, M.: Dynamic construction of finite automata from examples using hill climbing. In: Proceedings of the 4th Annual Cognitive Science Conference, pp. 105–108 (1982)
3. Dupont, P.: Regular grammatical inference from positive and negative samples by genetic search: the gig method. In: Carrasco, R.C., Oncina, J. (eds.) ICGI 1994. LNCS, vol. 862, pp. 236–245. Springer, Heidelberg (1994)