PORTING TO NEW DOMAINS USING THE LEARNER^{tm}

Robert J. P. Ingria
Lance Ramshaw

BBN Systems and Technologies Corporation
Cambridge, MA 02138

ABSTRACT

Acquiring syntactic and semantic information about a new application domain for a natural language processing system is often a time-consuming task. To address this problem, various researchers have developed acquisition tools to speed the process. While such tools are very useful, they are typically tied to particular systems and so their benefits cannot be shared by other researchers.

In this paper, we discuss an experiment using the Learner—a software tool for acquiring information about a new task domain for Parlance,[1] an ATN-based natural language system—to configure a quite different natural language system, the BBN ACFG, a unification-based system.

We have used the Learner to produce information in three major areas: syntactic and semantic information about the lexical items used in the new domain; translation rules from the parser output to the application system; and a class grammar for use in the speech recognition component of HARC, the BBN spoken language system.

Initial results are encouraging: 1499 lexical items have been acquired, of which 91% were directly usable, without any manual editing; all of the translation rules are usable; and a speech vocabulary of 2170 items, with an associated class grammar with a perplexity of 89, has been acquired with a small amount of manual editing.

INTRODUCTION

A major problem that restricts the usefulness of natural language processing systems is the cost, in time and effort, of porting a system to a new domain. Typically, the system designers are not experts in the application domain in which their system will be used, although they have the knowledge of syntax, semantics, and knowledge representation to configure the system. The end users, on the other hand, are experts in the application domain, but are almost always unversed in computational linguistics. Thus, the people who can most efficiently modify the system do not know what is necessary for a given domain, while the experts in that domain do not have the linguistic knowledge to perform the modifications themselves. Typically, then, the system designers spend long periods of time, on the order of months or even years, trying to acquire expertise in a domain and "tweaking" their systems to perform well in it, or else they try to impart just enough knowledge of syntax and semantics to the end users to allow them to do the configuration themselves. At best, these approaches result in a working version of the system, but at enormous cost; at worst, the end product is a not very usable system.

This problem is sometimes referred to as the "knowledge acquisition bottle neck": how can the knowledge of an application area be combined with the knowledge of a working system and its component technologies to produce a useful system without placing heavy burdens on either the end users or the system designers?

THE LEARNER

Some researchers have addressed this problem by developing acquisition tools targeted for end users, to allow them to provide the syntactic and semantic information necessary for the NL system, but without requiring them to become experts in these areas. Such tools usually take the burden of providing detailed syntactic and semantic analyses off the user through a guided acquisition procedure (either with menus or questions) and through the use of queries couched in terms of actual examples of language usage, rather than in terms from syntactic or semantic analysis (e.g. by asking "Can you say 'Someone

[1] Learner and Parlance are trademarks of Bolt Beranek and Newman Inc.

deployed something?'" rather than "Is 'deploy' transitive?") Such acquisition tools include the acquisition component of TEAM (Grosz, 1983), the LapItUp lexical acquisition package for the JANUS system (Cumming and Albano, 1986), and the TELI system's semantic acquisiton facility (Ballard and Stumberger, 1986).

BBN has developed a software package, the Learner, as a porting tool for non-expert users (Bates, 1989; BBN Parlance Learner Manual, 1989). The Learner creates a number of files that are used to configure the Parlance natural language processing system for a new application domain in a short time. Like the tools already mentioned, it uses an interactive, guided procedure to acquire syntactic and semantic information from the user. It also acquires information about the structure and content of the database directly from the database itself. Previous work with the Learner (Bates, 1989) has demonstrated a speed-up of ten times or more, compared to manual acquisition of the same information.

Recently, we have begun porting our ACFG natural language system (Boisen, et al, 1989b) to a personnel database domain, a domain for which Parlance had been configured using the Learner. This raised the possibility of using the output files created by the Learner as knowledge sources for components of the ACFG system. This is a particularly interesting test, since the Learner is not designed to be a general acquisition tool for arbitrary natural language processing systems, but is optimized to produce information necessary for the Parlance system. While Parlance is an ATN-based system, the BBN ACFG utilizes a unification grammar formalism, similar to a Definite-Clause Grammar; the two systems, then, are quite different. If, despite this difference, the same Learner output could be used to configure the ACFG system, this would suggest that the speed-up of using the Learner to port to a new domain already demonstrated for the Parlance system, could be extended to the ACFG system and, perhaps, to similar unification based grammars.

USING LEARNER OUTPUT

We have taken the output files produced by the Learner and developed software tools to convert them into forms usable by the ACFG system. We have experimented with acquiring three kinds of knowledge:

- Syntactic and Semantic Information
- Database Information
- Words and Word Classes

In the rest of this section, we discuss in somewhat more detail the information provided by the Learner in each of these areas. In the next section, we discuss the results of our efforts.

SYNTACTIC AND SEMANTIC INFORMATION

We have used Learner output to acquire lexical entries for the open class categories: nouns (both common and proper), adjectives, and verbs, that include both syntactic and semantic information. For each of these categories, the Learner provides any necessary morphological information, such as inflectional paradigm (e.g. that *city* forms its plural by affixing *-es*), associated irregular forms (e.g. that *got* is the past tense of *get*), etc. In addition, the Learner outputs information specific to each category:

Nouns: For ordinary entity type nouns, such as *programmer*, the Learner provides information about the semantic type of the noun and of the underlying concept with which it is associated (e.g. that a *programmer* is one of type *person* whose area of expertise is *programming*). For relational nouns, such as *salary*, the Learner provides information about the underlying relation associated with the noun, the domain of the relation, and the range of the relation (e.g. that the underlying relation of *salary* is *salary-of*, which applies to a *person* and produces a *monetary-amount*).

Adjectives: For adjectives, the Learner provides information about the type of noun to which the adjective can be applied and about the underlying function with which the adjective is associated (e.g. the adjective *asian american* is applied to nouns of type *person* and is true of those whose ethnic group is *asian-american*).

Verbs: For verbs, the Learner provides information about the underlying relation associated with the verb, the type restrictions associated with its noun phrase arguments, as well as any prepositions that the verb may select; for example, for the verb *graduate*, a noun of type *person* does the graduating,

the preposition *from* is used to mark the place from which the graduation took place, which is a noun of type *school.*

In addition to information about specific lexical items, the Learner produces information about the underlying concepts of the domain; for example, that there is a relation *gender-of* that applies to *persons* and produces a result of type *genders.* We have also used the Learner output to acquire this information.

DATABASE INFORMATION

As part of its output, the Learner produces a file of pattern transformation rules, which map from concepts in the semantic domain model—and, so, ultimately, from words associated with those concepts—to fields in the data base. This file contains the information that allows the associated natural language system to actually obtain an answer from the database. Since the database system used in the ACFG system—described in (Boisen, 1989)—is essentially a modified version of that used by Parlance, these rules are straightforwardly usable.

WORDS AND WORD CLASSES

We have also used the Learner to acquire a set of vocabulary items and word classes for a class grammar (Derr and Schwartz, 1989) for use in HARC, the BBN Spoken Language System (Boisen, et al, 1989a), which incorporates the ACFG system as its natural language component. Though the Learner does not directly produce a class grammar, we used the syntactic categories, semantic classes, and inflectional paradigms and forms which it provides to produce a class grammar. To create a complete set of words for our speech recognition system, we did not use the Learner output directly, since it only contains the base and irregularly inflected forms of words, but a lexicon that was created on the basis of the Learner output and which included inflected forms, as well.

RESULTS

Since our experiment in obtaining information from existing Learner output was performed at the same time that we were writing the code to perform the necessary translations, we cannot measure the efficiency of using the Learner output in terms of elapsed time or person weeks. Therefore, as a rough measure of the benefits of using Learner output as a source, we propose to compare the number of items obtained from the translation program which were usable without further manual modification with the number that required some hand editing.

SYNTACTIC AND SEMANTIC INFORMATION

A total of 1499 lexical items were acquired from the Learner output; of these 1379 (91%) were directly usable, without any human intervention. Since there were many proper nouns in the lexicon obtained, and since proper nouns typically did not need to be edited, we also present the percentage of lexical items that were immediately usable excluding the proper nouns, so that their presence does not bias the result; in this case, 74% of the derived lexical items were directly usable. These results are shown in the following table:

	Total	# "As Is"	% "As Is"
All Words	1499	1379	91%
– Proper Nouns	504	375	74%

Some comments are in order about the manual editting required. In the case of nouns and adjectives, some editting was required for syntactic and morphological features, owing to differences between the grammars of the Parlance and ACFG systems. All of the semantic information for nouns and adjectives, however, was left untouched. In the case of verbs, on the other hand, the difference between the Learner and ACFG semantic representations was too great to allow automatic acquisition of semantic information; for verb entries, the semantic portion was written by hand. However, even in the case of verbs, the semantic information was not written from scratch; rather, the semantic entry for each verb was a manual translation of the information in the Learner output.

We also obtained a total of 109 semantic concepts from the Parlance output, which were used in the semantic entries of lexical items. These concepts required no manual editting at all.

DATABASE INFORMATION

For this domain, the Learner produced 65 translation rules, all of which were usable unedited.

WORDS AND WORD CLASSES

The lexicon derived from the Learner was used to create a speech vocabulary of 2170 items, with an associated class grammar of 637 classes with a perplexity of 89, with a small amount of manual editing.

CONCLUSION

We have demonstrated that it is possible to use Learner output to produce information that is usable in the BBN ACFG system. This result is remarkable for a number of reasons. First, it suggests that the great boost in productivity reported in (Bates, 1989) by using the Learner to port the Parlance system may be extended to the ACFG system. Since the ACFG system is based on a unification grammar and parser, this, in turn, suggests that the Learner might be useful for other unification-base systems. Since the Learner was designed with the Parlance system in mind, the fact that its output is usable by a system based on a radically different grammar formalism and parsing algorithm may indicate that there has been a sufficient convergence in syntactic and analytic techniques, at least in the database retrieval area, to allow tools developed for one framework or system to be useful to others, as well.

Acknowledgments

The work reported here was supported by the Advanced Research Projects Agency and was monitored by the Office of Naval Research under Contract No. N00014-89-C-0008. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Government.

References

BBN Parlance Learner Manual (1989) BBN Systems and Technologies Corporation, Cambridge, Massachusetts.

B. Ballard and D. Stumberger (1986) "Semantic Acquisition in TELI: A Transportable, User-Customized Natural Language Processor" In *24th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, Association for Computational Linguistics, Morristown, NJ, pages 20–29.

Madeleine Bates (1989) "Rapid Porting of the Parlance Natural Language Interface" In *Proceedings of the Speech and Natural Language Workshop February 1989*, Morgan Kaufmann Publishers, Inc., San Mateo, California, pages 83–88.

Sean Boisen (1989) *The SLS Personnel Database (Release 1)*, SLS Notes No 2, BBN Systems and Technologies Corporation, Cambridge, Massachusetts.

S. Boisen, Y. Chow, A. Haas, R. Ingria, S. Roucos, R. Scha, D. Stallard and M. Vilain (1989a) *Integration of Speech and Natural Language: Final Report*, Report No. 6991, BBN Systems and Technologies Corporation, Cambridge, Massachusetts.

Sean Boisen, Yen-Lu Chow, Andrew Haas, Robert Ingria, Salim Roukos, and David Stallard (1989b) "The BBN Spoken Language System" In *Proceedings of the Speech and Natural Language Workshop February 1989*, Morgan Kaufmann Publishers, Inc., San Mateo, California, pages 106–111.

Susanna Cumming and Robert Albano (1986) *A Guide to Lexical Acquisition in the JANUS System* ISI Research Report ISI/RR-85-162, ISI, Marina del Rey/California.

Alan Derr and Richard Schwartz (1989) "A Simple Statistical Class Grammar for Measuring Speech Recognition Performance" In *Proceedings of the Speech and Natural Language Workshop October 1989*, Morgan Kaufmann Publishers, Inc., San Mateo, California.

Barbara J. Grosz (1983) "TEAM: A Transportable Natural-language Interface System" In *Conference on Applied Natural Language Processing: Proceedings of the Conference*, Association for Computational Linguistics, Morristown, NJ, pages 39–45.