EBL^2: AN APPROACH TO AUTOMATIC LEXICAL ACQUISITION

LARS ASKER* asker@dsv.su.se
BJÖRN GAMBÄCK† gam@sics.se
CHRISTER SAMUELSSON† christer@sics.se

Keywords: linguistic tools; lexical acquisition; explanation-based learning

Abstract

A method for automatic lexical acquisition is outlined. An existing lexicon that, in addition to ordinary lexical entries, contains prototypical entries for various non-exclusive *paradigms* of open-class words, is extended by inferring new lexical entries from texts containing unknown words. This is done by comparing the constraints placed on the unknown words by the natural language system's grammar with the prototypes and a number of hand-coded phrase *templates* specific for each paradigm. Once a sufficient number of observations of the word in different contexts have been made, a lexical entry is constructed for the word by assigning it to one or several paradigm(s).

Parsing sentences with unknown words is normally very time-consuming due to the large number of grammatically possible analyses. To circumvent this problem, other phrase templates are extracted automatically from the grammar and domain-specific texts using an explanation-based learning method. These templates represent grammatically correct sentence patterns. When a sentence matches a template, the original parsing component can be bypassed, reducing parsing times dramatically.

1 Introduction

A persisting trend in unification-based approaches to natural language processing is to incorporate large quantities of information in the lexicon, information that has traditionally resided in the grammar rules. Acquiring a lexicon has thus become a difficult and time consuming task, even for moderately sized lexica. In addition to this, any natural language processing system intended for serious applications must include a large lexicon – several thousands of words or more is commonly considered a minimum size – which adds even more to the complexity of the problem. In view of this, tools for lexical acqusition are not only desirable – they become a necessity.

Most approaches to this problem have been to construct a range of tools that require various degrees of *interactive* support when new lexical entries are created, either from raw text material (as in e.g., [Trost & Buchberger 86, Grosz *et al* 87, Wilensky 90] and the early work by Zernik [Zernik & Dyer 85, Zernik 87]), or from machine readable dictionaries (see e.g., [Boguraev *et al* 87, Calzolari & Bindi 90]). Although interactive tools for lexical acquisition greatly simplifies the task of constructing a lexicon, it is desirable to go one step further and fully remove the need for user interaction.

One of the first systems that aimed at constructing lexical entries automatically from raw text was Granger's FOUL-UP system [Granger 77]. FOUL-UP extended a lexicon by inferring restrictions placed on unknown words by instantiating scripts that matched the sentences containing the unknown words. This built on a number of assumptions which in general do not hold, in particular: that all the information needed to create an entry is contained in one text; that no morphological information is needed; that specific (hand-coded) scripts covering the domain can be made available in advance. In one of the later approaches to automatic lexical acquisition from raw text, [Jacobs & Zernik 88] have shown the need to consult a variety of knowledge sources such as morphological, syntactic, semantic, and contextual knowledge when determining a new lexical entry.

This paper describes an automatic method to acquire new lexical entries by using analytical learning in combination with strategies used in an existing interactive tool for lexical acquisition (VEX [Carter 89]). In the process of constructing a lexical entry, the system combines several different sources of information: the underlying NL system (CLE, [Alshawi (ed.) 92]) will contribute information on syntactically and semantically permissible phrases and on the rules for inflectional morphology. The corpus will contribute information on which of these constructions actually occur. This information is combined with the the linguistic knowledge encoded in the interactive lexical acquisition tool to infer lexical entries for unknown words in the text.

The rest of the paper is laid out as follows: Section 2 contains information about the various elements on which the method is based. Section 3 de-

*Department of Computer and Systems Sciences, Stockholm University, Electrum 230, S – 164 40 KISTA, Sweden.
†NLP-group, Swedish Institute of Computer Science, Box 1263, S – 164 28 KISTA, Stockholm, Sweden.

scribes the method itself and Section 4 reports on the current state of the implementation.

2 The elements of the scheme

2.1 The Core Language Engine, CLE

The Core Language Engine is a general purpose natural language processing system for English developed by SRI Cambridge. It is intended to be used as a building block in a broad range of applications, e.g. data-base query systems, machine translation systems, text-to-speech/speech-to-text systems, etc. The object of the CLE is to map certain natural language expressions into appropriate predicates in logical form (or Quasi-Logical Form [Alshawi & van Eijck 89]). The system is based completely on unification and facilitates a reversible phrase-structure type grammar.

The Swedish Institute of Computer Science has with support from SRI generalized the framework and developed an equivalent system for Swedish (the S-CLE, [Gambäck & Rayner 92]). The two copies of the CLE have been used together to form a machine translation system [Alshawi *et al* 91]. The S-CLE has a fairly large grammar covering most of the common constructions in Swedish. There is a good treatment of inflectional morphology, covering all main inflectional classes of nouns, verbs and adjectives.

The wide range of possible applications have put severe restrictions on the type of lexicon that can be used. The S-CLE has a function-word lexicon containing about 400 words, including most Swedish pronouns, conjunctions, prepositions, determiners, particles and "special" verbs. In addition, there is a "core" content-word lexicon (with common nouns, verbs and adjectives) and domain specific lexica. This part of the system is still under development and all these content-word lexica together have about 750 entries.

The lexical entries contain information about inflectional morphology, syntactic and semantic subcategorization, and sortal (selectional) restrictions. Information about the linguistic properties of an entry is represented by complex categories that include a principal category symbol and specifications of constraints on the values of syntactic/semantic features. Such categories also appear in the CLE's grammar and matching and merging of the information encoded in them is carried out by unification during parsing. Two categories can be unified if the constraints on their feature values are compatible.

In the actual "core" and domain lexica, this information is kept implicit and represented as pointers to entries in a "paradigm" lexicon with a number of words representing basic word usages and inflections. For these "paradigm words" only, the complete set of feature values is explicitly specified.

2.2 The Vocabulary EXpander, VEX

In the English CLE, new lexicon entries can be added by the users with a tool developed for the purpose. This lexicon acquisition tool, the Vocabulary EXpander, is fully described in [Carter 89]. In parallel with the development of the S-CLE, a Swedish version of the VEX system was designed [Gamback 92].

VEX allows for the creation of lexical entries by users with knowledge both of a natural language and of a specific application domain, but not of linguistic theory or of the way lexical entries are represented in the CLE. It presents example sentences to the user and asks for information on the grammaticality of the sentences, and for selectional restrictions on arguments of predicates. VEX adopts a *copy and edit* strategy in constructing lexical entries. It builds on the "paradigm" lexicon and *sentence patterns*, that is, declarative knowledge of the range of sentential contexts in which the word usages in that lexicon can occur.

In the present work we want to investigate to what extent such creation of lexicon entries can be performed with a minimum of user interaction. Instead of presenting example sentences to the user we are allowing the program to use a very large text where hopefully unknown words will occur in several different sentence patterns. This strategy will be further described in the following sections.

First, however, we will define what we mean by the notion of (subcategorization) "paradigm". The definition we adopt here is based on the one used in [Carter 89], namely that

Definition 1
a paradigm *is any minimal non-empty intersection of lexical entries. Every category in a paradigm will occur in exactly the same set of entries in the lexicon as every other category (if any) in that paradigm. Every entry consists of a disjoint union of paradigms.*

Here, we assume that a lexicon can be described in terms of (a small set of) such paradigms, relying on the fact that the open-class words exhibit at least approximate regularities.[1]

2.3 The Lexicon Learning system, L^2

Previous experiments in automatic lexical acquisition at SICS (L^2 – Lexicon Learning) used a set of

[1] The system does not attempt to cope with closed-category words. They have to be entered into a specific function-word lexicon by a skilled linguist.

sentences and a formal grammar to infer the lexical categories of the words in the sentences. The original idea was to start with an empty lexicon, assuming that the grammar would place restrictions on the words in the sentences sufficient to determine an assignment of lexical categories to them [Rayner *et al* 88]. This can be viewed as solving a set of equations where the words are variables that are to be assigned lexical categories and the constraints that all sentences parse with respect to the grammar are the equations.

Unfortunately, it proved almost impossible to parse sentences containing several unknown words. For this reason the scheme was revised in several ways [Hörmander 88]; instead of starting with an empty lexicon, the starting point became a lexicon containing closed-class words such as pronouns, prepositions and determiners. The system would then at each stage only process sentences that contained exactly one unknown word, the hope being that the words learned from these sentences would reduce the number of unknown words in the other ones. In addition to this, a morphological component was included to guide the assignments. Although the project proved the feasibility of the scheme, it also revealed some of its inherent problems, especially the need for faster parsing methods.

2.4 Explanation-based learning, EBL

A problem with all natural language grammars is that they allow a vast number of possible constructions that very rarely, if ever, occur in real sentences. The application of explanation-based learning[2] (EBL) to natural language processing allows us to reduce the set of possible analyses and provides a solution to the parsing inefficiency problem mentioned above (Subsection 2.3).

The original idea [Rayner 88] was to bypass normal processing and instead use a set of learned rules that performed the tasks of the normal parsing component. By indexing the learned rules efficiently, analysing an input sentence using the learned rules is very much faster than normal processing [Samuelsson & Rayner 91]. The learned rules can be viewed as *templates* for grammatically correct phrases which are extracted from the grammar and a set of training sentences using explanation-based learning. Here, we assume the following definition:

Definition 2
a template *is a generalization constructed from the parse tree for a successfully processed phrase. A template is a tree spanning the parse with a mother category as root and a collection of its ancestor nodes (at arbitrary, but pre-defined, deep levels of nesting) as leaves.*

The fact that the templates are derived from the original grammar guarantees that they represent correct phrases and the fact that they are extracted from real sentences ensures that they represent constructions that actually occur.

3 Explanation-based lexical learning, EBL[2]

The basic algorithm goes as follows:

1. Using a large corpus from the domain, extract templates from the sentences containing no unknown words.
2. Analyse the remaining sentences (the ones containing unknown words) using the templates, while maintaining an interim lexicon for the unknown words.
3. Compare the restrictions placed on the unknown words by the analyses obtained with other hand-coded phrase templates specific for the paradigms in the lexicon.
4. Create "real" lexical entries from the information in the interim lexicon when a full set of such templates (covering a paradigm) has been found.

In the following subsections, we will address these issues in turn.

3.1 Extracting templates from a domain-specific corpus

A typical situation where we think that this method is well suited is when a general purpose NL system with a core lexicon (such as the S-CLE) is to be customized to a specific application domain. The vocabulary used in the domain will include e.g. technical terms that are not present in the core lexicon. Also, the use of the words in the core lexicon may differ between domains. In addition to this, some types of grammatical constructs may be more common in one domain than in another. We will try to "get the flavour of the language" in a particular application environment from domain-specific texts.

The corpus is divided into two parts: one with sentences containing unknown words, and another where all the words are known. The latter group is used to extract phrase templates that capture the grammatical constructions occurring in the domain. The process of extracting phrase templates from training sentences is outlined in Subsection 2.4.

[2] Explanation-based learning is a machine learning technique closely related to macro-operator learning, chunking, and partial evaluation and is described in e.g., [DeJong & Mooney 86, Mitchell *et al* 86].

3.2 Analysing the remaining sentences

Assuming that a particular set of phrase templates is applicable to a sentence containing an unknown word will associate a set of constraints with the word. Naturally, the constraints on the *known* words of the sentence should be satisfied if this template is to be considered.[3] This will correspond to a particular parse or analysis of the sentence. Thus a set of constraints is associated with each different parse.

A number of entries in the prototype lexicon will match the set of constraints associated with a sentence. Each prototype is an incarnation of a paradigm. Thus we can associate a word with a set of paradigms. (Note that the paradigms may be non-exclusive.) All such associations (corresponding to different parses of the same sentence) are collected, and used to update the interim lexicon.

The most obvious constraints come from syntactic considerations. If, in the sentence *John loves a cat* the word *loves* were unknown, while the other words did indeed have the obvious lexical entries, the grammar will require *loves* to be a transitive verb of third person singular agreement. Since the prototypes of verbs are in the imperative form, we must associate a finite verb form with the imperative. This is done by applying a morphological rule that strips the '-s' from the word *loves*, reinforcing the hypothesis and gaining the tense information in the process.

Now, this morphological information may seem unimportant in English, but it definitely is not in Swedish: a word with more than one syllable ending with '-or' has to be an indefinite common gender noun. If it is not of latin origin it must be a plural form and thus its entire morphology is known. The odds that it is a countable noun (like *duck*), as opposed to a mass noun (such as *water*), are overwhelming.

3.3 Constructing lexical entries

During the analysis of the set of sentences containing unknown words, an interim lexicon for these unknown words is kept. The interim lexicon is indexed on word stems and updated each time a new sentence is processed. For each word stem, two pieces of information are retained in this lexicon: a hypothesis about which paradigm or set of paradigms the word is assumed to belong to, and a justification that encodes all evidence relevant to the word. The justification is used to make the hypothesis and is maintained so that the entry may be updated when new information about the word arrives. When all the phrase templates (sentence patterns) for fulfillment of a specific paradigm have been found, an entry for the word is made in the domain-specific lexicon that is being constructed. This is done while still keeping the justification information, since this might contain evidence indicating other word-senses or homographs.

[3] Unless they do in fact correspond to other non-lexicalized senses of the word, or to homographs.

4 Implementation status

A preliminary version of the lexical acquisition system has been implemented in Prolog. The module extracting templates from sentences with known words is fully operational. The parser for sentences with unknown words has also been tested, while the interim lexicon still is subject to experimentation. Presently, a very simple strategy for the interim lexicon has been tested. This version uses the set of all hypotheses as the justification and use their disjunction as the current hypothesis. We are currently working on extending this scheme to one incorporating the full algorithm described above.

Unknown words are matched with the subcategorization paradigms of the S-CLE. In total 62 different syntactic/semantic paradigms are known by the present system; 5 for Swedish nouns, 10 for adjectives, and all the others for verbs. The morphological inflections are subdivided into 14 different inflectional classes of nouns, 3 classes of adjectives, and 24 classes of verbs.

5 Conclusions

We have outlined a method for automatic lexical acquisition. An existing lexicon built on the usage of prototypical entries for paradigms of open-class words, is extended by inferring new lexical entries from texts containing unknown words. The constraints placed on these words by the grammar are compared with the prototypes and a hypothesis is made about what paradigm the word is most likely to belong to.

The hypotheses about the unknown words are kept in an interim lexicon until a sufficient level of confidence is reached. Phrase templates are both hand-coded and extracted from the grammar and domain-specific texts using an explanation-based learning method.

6 Acknowledgements

The work reported here was funded by the Foundation for the Swedish Institute of Computer Science and the Swedish National Board for Industrial and Technical Development (NUTEK).

We would like to thank Manny Rayner and David Carter (SRI Cambridge) and Seif Haridi (SICS) for helpful discussions and suggestions, and Pierre Gander (Stockholm University) for valuable support.

References

Alshawi, H. and J. van Eijck (1989). "Logical Forms in the Core Language Engine", *the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, British Columbia, pp. 25-32.

Alshawi, H., D. Carter, B. Gambäck and M. Rayner (1991). "Translation by Quasi Logical Form Transfer", *the 29th Annual Meeting of the Association for Computational Linguistics*, University of California, Berkeley, California, pp. 161–168.

Alshawi, H. (ed.) (1992). *The Core Language Engine*, Cambridge, Massachusetts: The MIT Press.

Boguraev, B., T. Briscoe, J. Carroll, D. Carter and C. Grover (1987). "The Derivation of a Grammatically Indexed Lexicon from the Longman Dictionary of Contemporary English", *the 25th Annual Meeting of the Association for Computational Linguistics*, Stanford, California, pp. 193–200.

Carter, D. (1989). "Lexical Acquisition in the Core Language Engine", *the 4th Conference of the European Chapter of the Association for Computational Linguistics*, Manchester, England, pp. 137–144. Also available as SRI International Report CCSRC-012, Cambridge, England.

Calzolari, N. and R. Bindi (1990). "Acquisition of Lexical Information from a Large Textual Italian Corpus", *the 13th International Conference on Computational Linguistics*, Helsinki, Finland, Vol. 3, pp. 54–59.

DeJong, G. and R. Mooney (1986). "Explanation Based Learning: An Alternative View", *Machine Learning*, **1**:145–176.

Gambäck B. and M. Rayner (1992). "The Swedish Core Language Engine", *the 3rd Nordic Conference of Text Comprehension in Man and Machine*, Linköping, Sweden.

Gambäck B. (1992). "Lexical Acquisition: The Swedish VEX System", *the 3rd Nordic Conference of Text Comprehension in Man and Machine*, Linköping, Sweden.

Granger, R.H. (1977). "FOUL-UP: A program that figures out meanings of words from context", *the 5th International Joint Conference on Artificial Intelligence*, Cambridge, Massachusetts, pp. 172–178.

Grosz, B.J., D.E. Appelt, P. Martin, and F.C.N. Pereira (1987). "TEAM: An Experiment in the Design of Transportable Natural-Language Interfaces", *Artificial Intelligence*, **32**:173–243.

Hörmander, S. (1988). "The Problems of Learning a Lexicon with a Formal Grammar", SICS Research Report – R88019, Stockholm, Sweden.

Jacobs, P. and U. Zernik (1988). "Acquiring Lexical Knowledge from Text: A Case Study", *the 7th National Conference on Artificial Intelligence*, Saint Paul, Minnesota, pp. 739–744.

Mitchell, T.M., R.M. Keller and S.T. Kedar-Cabelli (1986). "Explanation-Based Generalization: A Unifying View", *Machine Learning*, **1**:47–80.

Rayner, M. (1988). "Applying Explanation-Based Generalization to Natural-Language Processing", *the International Conference on Fifth Generation Computer Systems*, Tokyo, Japan, pp. 1267–1274.

Rayner, M., Å. Hugosson and G. Hagert (1988). "Using a Logic Grammar to Learn a Lexicon", *the 12th International Conference on Computational Linguistics* Budapest, Hungary, pp. 524–529. Also available as SICS Research Report – R88001, Stockholm, Sweden.

Samuelsson, C. and M. Rayner (1991). "Quantitative Evaluation of Explanation-Based Learning as an Optimization Tool for a Large-Scale Natural Language System", *the 12th International Joint Conference on Artificial Intelligence*, Sydney, Australia.

Trost, H. and E. Buchberger (1986). "Towards the Automatic Acquisition of Lexical Data", *the 11th International Conference on Computational Linguistics*, Bonn, Germany, pp. 387–389.

Wilensky, R. (1990). "Extending the Lexicon by Exploiting Subregularities", *the DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania, pp. 365–370.

Zernik, U. and M. Dyer (1985). "Towards a Self-Extending Lexicon", *the 23rd Annual Meeting of the Association for Computational Linguistics*, University of Chicago, Chicago, Illinois, pp. 284–292.

Zernik, U. (1987). "Language Acquisition: Learning a Hierarchy of Phrases", *the 10th International Joint Conference on Artificial Intelligence*, Milan, Italy, pp. 125–131.