Abstract

I propose research in automatic grammar induction and discovery of semantic relationships. The former has applications in the study of language similarity, language acquisition, and numerous other fields within linguistics and language technology; the latter can improve information retrieval, information extraction, dialog systems, and automatic spelling and grammar checkers. Both are active research topics in the cross-section between linguistics and computer science, and I feel that the two can be combined into something that is better than the sum of its parts.

To this end, I will construct and evaluate a system capable of extracting grammatical and semantic information from unannotated corpora (and thereby for any language). Work on the grammar portion of this system has already begun, the results so far are encouraging, and it appears that, to a certain extent, I have independently arrived at some of the same ideas as those behind the ADIOS system, which is developed jointly by the Cornell and Tel Aviv universities and which I will study further. In my work, I will consider not only phrase grammars, but also rules for phonology, morphology, and discourse.

I will then extend this system to also extract semantic information from the corpora presented to it, allowing a deeper and more thorough analysis. Here, I will explore to what extent this can be done with little or no additional data outside the corpora themselves.

1 Introduction

Significant progress has been made in the area of grammar induction over the last few years, both for dependency grammars (Nivre and Scholz, 2004) and phrase structure grammars (Collins, 1999), (Megyesi, 2002). New methods and, to a certain extent, more powerful computers have made unsupervised algorithms, with their enticing prospects of language independence and superior versatility compared to supervised ones, possible. Highly accurate unsupervised grammar induction algorithms could help many research disciplines by automating otherwise extremely time-consuming tasks; more on this later.

It should be noted that, for convenience, throughout this document we will use the "grammar" concept in the broadest possible sense. We will think of it not only as rules that describe how to combine words into phrases and phrases into sentences, but also ones that tell us how to combine phonemes into morphemes, morphemes into words, and sentences into texts or discourses. We will further consider rules that relate how musical notes are combined into music, how gene sequences are constructed, etc., as "grammatical."

Once a grammar (of any type) has thus been gleaned from some kind of corpus, it is not difficult to write a computer program that generates new morphemes, words, phrases, etc. that are described by said grammar and that therefore, theoretically, could have been taken from that corpus. But, of course, these generated entities will be random and devoid of any meaning, unless the system (and the grammar) has some notion of semantics — some way of expressing relationships between some kind of data model on the one hand and entities described by the grammar on the other. If semantic relationships, too, can be automatically extracted from the corpus and described in or associated with the grammar, the overall system's usefulness will be vastly augmented. Some work has, of course, been done in this border zone between grammar and semantics; for instance, Briscoe and Carroll (1997).

Thus, I propose research in the area of automatic induction of grammars and discovery of semantic relationships.
2 Motivation

The ability to automatically construct grammars and other formalized sets of rules for human language alone can provide many useful things:

- In various fields within linguistics, valuable time-saving tools can be built to assist the analysis and comparison of different languages, dialects, or sociolects, or to help deepen the understanding of grammatical structures. Above all, such automatic tools can be used to make analyses that have heretofore been practically infeasible, and for small, endangered or extinct languages for which no annotated corpora are available.
- The understanding of human language acquisition can be helped by, for instance, observing the emergence of more complex syntactic constructs with increasing proficiency or grammatical traits in individual speakers’ or writers’ idiolects. Again, such analyses may be impractical to make manually.
- Many commercial applications, in areas such as grammar checking and readability analysis, can be made more accurate and reliable.
- In seemingly unrelated and sometimes unexpected areas, “grammar-like” rules can be used to facilitate the understanding of underlying structures or phenomena. For instance, this applies to music and biochemistry.

Furthermore, if this is combined with an ability to also extract semantic relationships, several additional benefits present themselves:

- Information retrieval can be made more accurate through systems that are better able to disambiguate homonyms and equate synonyms based on semantic, not just syntactic, context.
- Information extraction can be made more reliable by making systems behave more like humans when “reading” texts.
- Dialog systems, whether text- or speech-based, can be improved since their grammars, with semantic information, can be made more complete and more complex with less effort.
- Machine translation applications can infer semantic-vs.-grammatical patterns from the source and target languages in order to improve translations of ambiguous or otherwise “tricky” passages.

3 Goals

The concrete goal of the proposed research project is to build, refine and evaluate a set of tools for automatic induction of grammar and discovery of semantics:

- A tool to induce grammars from completely unannotated corpora. (And, again, we use “grammar” to mean sets of rules on all levels — from phonology to discourse.)
- A tool to infer semantic relationships from those same corpora, if necessary in conjunction with other resources such as dictionaries, sense inventories, or “word nets.”

By so doing, I would also achieve the more general goal of gaining an increased formalized understanding of properties of grammars and the mechanisms of how grammatical and semantic entities are related.

4 Approach

Some time ago, while experimenting with a grammar induction system based on observed frequencies of part-of-speech bigrams (Wastholm, 2005), I observed that that system’s main flaw was its fixed two-token window. Its inability to discover patterns beyond this window — and thereby its inability to consider entire phrases or sentences — meant that it would never be able to arrive at a grammar that would come particularly close to describing the language at hand. It seemed to me that a system that instead looked at entire phrases or sentences would be much more effective. Furthermore, I felt that such a system would be able to induce grammars from (sufficiently large) unannotated corpora, since the extra information gained by examining and generalizing from entire sentences would make up for the loss of part-of-speech data.

I started constructing a new system, called Grind, based on a modified and extended version of the “longest common subsequences” algorithm (Hunt and Szymanski, 1977) and was recently able to begin making some informal and preliminary experiments, with encouraging results.
Then, in August of this year, I learned of the existence of the Adios system (Solan et al., 2005). Intriguingly, it uses an algorithm quite similar to the one I devised, and also works on unannotated corpora. I would therefore very much like to study it further, compare it to mine (which, admittedly, is not yet as complete and not as well-tested), and attempt to synthesize the strong points of both into an improved system.

I would then like to move further by adding semantics to the mix. Cilibrasi and Vitanyi (2004) has demonstrated how to automatically discover the meanings of words using an unannotated corpus, namely Google’s index of the World-Wide Web. Barker (1998) has showed how to semi-automatically, with some user assistance, extract semantic relationships from unannotated text without the need for external data sources. I would like to explore these and similar techniques for automatic discovery of semantics, with or without such additional data sources, and also examine how grammar induction and semantic discovery each can improve the other’s results.

5 Time Plan
I propose a preliminary time plan as follows:

- **2006 — Phrase Grammar:**
  - Collection of material.
  - Further study of Adios and other similar systems.
  - Completion and formal evaluation of first version of the Grind grammar induction system.
  - Report.

- **2007 — Generalized Grammar:**
  - Collection of material.
  - Study of strategies for automatic phonology, morphology, and discourse analysis.
  - Construction and evaluation of phonology, morphology, and/or discourse grammar induction system.
  - Survey of other areas where similar grammar-like rules can be discovered and applied — sheet music, gene sequences, amino acids, etc.
  - Report.

- **2008 — Semantics:**
  - Collection of material.
  - Study of modeling techniques for semantic relationships.
  - Study and comparison of semi-automatic and automatic semantic discovery systems.
  - Construction and evaluation of semantic data model.
  - Report.

- **2009 — Synthesis:**
  - Development of algorithms for improving grammatical analysis though semantics, and vice versa.
  - Construction and evaluation of complete system for grammar induction and semantic discovery.
  - Report.

6 Closing Remarks
I am very excited about the momentum that grammar induction has gained in recent years and would very much like to make my contribution to this field. I feel that, with the grammar induction itself as good as it has now become, the next logical step is to include a more elaborate semantic element. What I propose, therefore, is to try to find the best way of accomplishing this.

References


