**AI approaches to statistical language models**

**Eugene Charniak**

**Brown University**
Providence, RI 02912

**Abstract**

We have made a number of advances under this grant. We have created what is currently the most accurate parser for parsing into Penn-treebank-style trees; a program that identifies the antecedents of pronouns with 85% accuracy; a program that assigns function tags to parse text with 85% accuracy; a program that assigns referents to full noun phrases with 65% accuracy; very efficient parsers — parsers that explore very few constituents that do not appear in the final parse; and programs that discover semantic information about words from unlabeled text. Furthermore, all these programs work by statistical means.
1 Administrative Information

Title: AI Approaches to Statistical Language Models
ONR Grant Number: N00014-96-1-0549
Organization: Brown University
Effective Date: 3/1/96
Closing Date: 8/31/99

Principal Investigator: Eugene Charniak
Professor of Computer Science
(401) 863-7636
ec@cs.brown.edu

2 Summary of Results

In the proposal that resulted in this research grant we said, "We propose to create better statistical language models by bringing together statistical methodology and traditional AI approaches." It is fair to say that we have been successful in this goal. In this period we have created:

- what is currently the most accurate parser for parsing into Penn-treebank style trees. This parser has a per-constituent precision and recall of 89.5% (91.1% for sentences of length less than or equal to 40 words and punctuation) [Charniak 1999]

- a program that identifies the antecedents of pronouns with a 85% accuracy[Ge, Hale, Charniak 1998]
• a program that assigns function-tags to parse text with 85% accuracy. (E.g., the noun phrase “yesterday” would be given the function tag “temporal” in “He ate yesterday” to distinguish it from the role played by the noun phrase “pizza” in “He ate pizza”) [Blaheta, Charniak 2000]

• a program that assigns referents to full noun phrases with a 65% accuracy [Hall, Charniak 2000]

• very efficient parsers — parsers that explore very few constituents that do not appear in the final parse [Caraballo, Charniak 1998] [Charniak, Goldwater, Johnson 1998] [Blaheta, Charniak 1999]

• programs that discover semantic information about words from unlabeled text [Roark, Charniak 1998] [Caraballo 1999] [Caraballo Charniak 1999] [Berland Charniak 1999]

Furthermore, as stated in our research proposal, all of these programs work by statistical means.

We have also been able to combine many of these programs in order to parse, find noun-phrase coreference, and function-tag, large quantities of text. For example, in the last month we have delivered to the LDC (Linguistic Data Consortium, the major organization for the distribution of large text and speech corpora) 35 million words of Wall Street Journal newspaper articles that have been machine annotated in the aforementioned fashion. The LDC will be distributing this new corpus.

We expect that in the years that follow we will be able to increase the accuracy and depth of this annotation. In particular we expect in the next few months to be able to provide not just a parse tree, but the predicate-argument structure of the sentences. More generally we hope to provide deeper semantic annotations such as case (or thematic) roles.

3 Publications


A statistical syntactic disambiguation program and what it learns, Eugene Charniak and Murat Ersan, TR CS-95-29 Brown University, Department of


Automatic construction of a hypernym-labeled noun hierarchy from text Sharon A. Caraballo, (1999)
