UCSG Shallow Parser

G. Bharadwaja Kumar, Kavi Narayana Murthy

Department of Computer and Information Sciences
University of Hyderabad, India
email: knmuh@yahoo.com,g_vijayabharadwaj@yahoo.com

Abstract. Recently, there is an increasing interest in integrating rule based methods with statistical techniques for developing robust, wide coverage, high performance parsing systems. In this paper[1], we describe an architecture, called UCSG shallow parser architecture, which combines linguistic constraints expressed in the form of finite state grammars with statistical rating using HMMs built from a POS-tagged corpus and an A* search for global optimization for determining the best shallow parse for a given sentence. The primary aim of the design of the UCSG parsing architecture is developing a judicious combination of linguistic and statistical methods to develop wide coverage robust shallow parsing systems, without the need for large scale manually parsed training corpora. The UCSG architecture uses a grammar to specify all valid structures and a statistical component to rate and rank the possible alternatives, so as to produce the best parse first without compromising on the ability to produce all possible parses. The architecture supports bootstrapping with an aim to reduce the need for parsed training corpora. The complete system has been implemented in Perl under Linux. In this paper we first describe the UCSG shallow parsing architecture and then focus on the evaluation of the UCSG finite state grammar for the chunking task for English. Recall of 91.16% and 93.73% have been obtained on the Susanne parsed corpus and CoNLL 2000 chunking task test data set respectively. Extensive experimentation is under way to evaluate the other modules.
Key Words:- Chunking, Shallow Parsing, Finite State Grammar, HMM, A* search, UCSG Architecture

1 Introduction

Although a lot of work has gone into developing full syntactic parsers, high performance, wide coverage syntactic parsing has remained a difficult challenge [1]. In recent times, there has been an increasing interest in wide coverage and robust but partial or shallow parsing systems. Shallow parsing is the task of recovering only a limited amount of syntactic information from natural language sentences. Often shallow parsing is restricted to finding phrases in sentences, in which case it is also called chunking. Steve Abney[2], has described chunking as

[1] The research work reported here was supported in part by the University Grants Commission under the UPE scheme