Chapter 76
Dependency Parsing in Bangla

Utpal Garain
Indian Statistical Institute, India

Sankar De
Gupta College of Technological Sciences, India

ABSTRACT

A grammar-driven dependency parsing has been attempted for Bangla (Bengali). The free-word order nature of the language makes the development of an accurate parser very difficult. The Paninian grammatical model has been used to tackle the free-word order problem. The approach is to simplify complex and compound sentences and then to parse simple sentences by satisfying the Karaka demands of the Demand Groups (Verb Groups). Finally, parsed structures are rejoined with appropriate links and Karaka labels. The parser has been trained with a Treebank of 1000 annotated sentences and then evaluated with un-annotated test data of 150 sentences. The evaluation shows that the proposed approach achieves 90.32% and 79.81% accuracies for unlabeled and labeled attachments, respectively.

1. INTRODUCTION

Dependency parsing is a method of analyzing natural language sentences and outputs a tree of word-on-word dependencies (as opposed to constituent trees of context-free derivations) in a sentence. Dependency is defined as a binary asymmetric relation between two words. Dependency relations are close to semantic relations, which facilitate semantic interpretation of the sentence. This is why dependency parsing gained a lot of attention and popularity for natural language analysis and understanding in recent years.

Dependency parsing can be broadly divided into grammar-driven and data-driven parsing. Most of the modern grammar-driven dependency parsers parse by eliminating the parses which do not satisfy the given set of constraints. They view parsing as a constraint-satisfaction problem. Some of the constraint based parsers known in the literature can be found in (Karlsson, 1995; Maruyama, 1990; Bharati, 1993; Bharati, 2002; Tapanainen, 1998; Schröder, 2002; Debusmann, 2004). Multi-dimensional paradigm proposed in these studies attempted to capture various aspect of a language. Data-driven parsers, on the other hand, use a corpus to induce a probabilistic model for disambiguation (Nivre, 2005).

Constraint based parsing has been successfully tried for Indian languages (Bharati, 1993; Bharati,

DOI: 10.4018/978-1-4666-6042-7.ch076

2002). Under this scheme the parser exploits the syntactic cues present in a sentence and forms Constraint Graphs (CG). It then translates the CG into an Integer Programming problem. The solutions to the problem provide the possible parses for the sentence. Recent works of Bharati (2008b, 2009a, 2009b) show a substantial improvement in grammar driven IL parsing. But most of the works are confined to Hindi only.

Bangla, like other ILs, is a morphologically rich free-word order language. Parsing such types of languages is very challenging (Saha, 2006; Sarkar, 2006; & Hasan, 2011). In this paper we have used dependency based framework and Paninian model (Bharati, 1993; Bharati, 1995) for parsing Bangla sentences. Complex and compound sentences have been simplified to derive two or more simple sentences. A constraint-based approach has then been applied to parse the simple sentential structures. The method also includes the demand-source concept of Paninian grammar described later in this paper.

The very next section of the paper describes the demand-source approach of Paninian grammar. After that, the overall parsing approach is presented. The approach makes use of demand frames and therefore, the important concepts of demand frames are presented in the next section and it is shown that how frames are changed depending on the TAM of the verbs. The subsequent sections describe the constraints applied and it is discussed that how such constraints can be applied to reduce the problem to a bipartite graph matching problem. The algorithm for parsing a simple sentence is then explained. Evaluation results and error analysis has been discussed in the sections thereafter.

2. PANINIAN GRAMMAR

The Paninian framework was originally designed more than two millennia ago for writing a grammar for Sanskrit. This framework is now being adapted for analyzing modern Indian Languages (ILs) which are actually the derivatives of Sanskrit. Paninian grammar is particularly suited for morphologically rich free word ordered languages like most ILs including Bangla.

As conceived by the syntactico-semantic model of Paninian Grammar, every verbal root (*dhaatu*) denotes an action consisting of: (1) an activity and (2) a result. Result is the state which when reached the action is complete. Activity consists of actions carried out by different *participants* or *Karakas* (mostly noun groups) involved in the action. The Karakas have direct relation to the verb. The Paninian model used only six such Karakas such as K1, K2, K3, K4, K5, K7. Some additional relations have been described in (Bharti, 2009c) and the complete tag set has been given in Appendix A. In this approach the verb demands some karakas carryout the activity. Thus verb groups are known as Demand Groups and Karakas as the Source Groups or arguments. So for a very simple sentence (single Demand Group) like S1, the verb group is the root of the dependency tree connecting some noun groups with appropriate Karaka labels (Bharati, 1993). Consider the sentence in Box 1. The parsed output will be shown in Figure 1.

In sentence S1, Ram is performing the act of eating. So Ram is marked as K1 (karta/doer/subject). The activity eating directly affects *bhat* (rice). It is thus marked as K2 (karma/object).

Simple sentences are parsed using the demand frames and transformation rules (Cormen, 2009)

Figure 1. Parsed output of S1