PARTIAL PARSING WITH GRAMMATICAL FEATURES

A. G. Manousopoulou, G. Papakonstantinou and P. Tsanakas
Computer Systems Laboratory, National Technical University of Athens
Iroon Polytechneioy 9, 15773, Athens, Greece
e-mail: natasha@cslab.ece.ntua.gr

Abstract

This paper describes a rule based method for partial parsing, particularly for noun phrase recognition, which has been used in the development of a noun phrase recognizer for Modern Greek. This technique is based on a cascade of finite state machines, adding to them a characteristic very crucial in the parsing of words with free word order: the simultaneous examination of part of speech and grammatical feature information, which are deemed equally important during the parsing procedure, in contrast with other methodologies.

1 Introduction

Partial parsing is an area of natural language processing that has been widely explored in the past years with rule based and statistical methodologies. The rule based techniques usually employ pattern matching techniques [1], special grammars such as Constraint Grammars [2] or finite state cascades [3, 4]. The statistical techniques, on the other hand, mainly use Hidden Markov Models, such as the one introduced by Church [5], which is an extension of his well known statistical part of speech tagging methodology. There are also hybrid techniques that combine rules and statistics [6].

In this paper we describe a rule based method for partial parsing, particularly for noun phrase recognition. This technique is based on a cascade of finite state machines, adding to them a characteristic very crucial in the parsing of words with free word order: the simultaneous examination of part of speech and grammatical feature information, which are deemed equally important during the parsing procedure, in contrast with previously employed methodologies.

2 Grammatical Features in Free Word Order Languages

The noun phrase in Modern Greek consists of various parts of speech; it can be elementary or composite, containing recursively other noun phrases or secondary clauses. In this paper we will concentrate on simple, non recursive noun phrases. Modern Greek is a language with free order of the sentence constituents; this also holds partially for the words inside the noun phrase. The possible part of speech combinations are numerous, although there are some general rules that apply to any noun phrase [7].

Due to the freedom in the order of words inside a noun phrase the patterns that describe it are very flexible. An important parameter in the correct determination of noun phrase boundaries are thus the grammatical feature values. A change in feature values from one word to its next usually denotes a phrase boundary, which cannot be generally detected in another manner merely from the POS information. A noun recognizer - or a partial parser in general - for such a language should examine part of speech information together with grammatical features and not just use the features for later verification.

3 Finite State Transducers with Constraints

A very common way of constructing multi-level rule based partial parsers is that of combining finite state transducers (FSTs), which generally operate on the POS tag of the word and cannot take grammatical features into account in a straightforward manner. To solve this problem, we can use parameterization of FSTs, creating finite state machines that manage a limited amount of internal data. These data represent grammatical features

and will be used to verify constraints during the transitions of the machine.

Features are added to an FST as a finite list of feature sets A = (a_1, a_2, …, a_m) attached to symbols of the input and output alphabet of the trabsducer. Each set a_i contains the possible values of the i-th feature. The input of an FST with constraints is a string of the form:

$$x_1 \left\{T_1^1, T_1^2, \ldots, T_1^{k_1}\right\} x_2 \left\{T_2^1, T_2^2, \ldots, T_2^{k_2}\right\} \ldots x_n \left\{T_n^1, T_n^2, \ldots, T_n^{k_n}\right\}$$

Each tuple T_i^j is a member of the cartesian product $a_1 \times a_2 \times \ldots \times a_m$ and represents a possible value combination for the features of the symbol x_i.

Transitions in FSTs with constraints accumulate intersections of tuple sets $\left\{T_i^1, T_i^2, \ldots, T_i^{k_i}\right\}$. If the accumulated intersection is empty, the transition fails. In this way, we can ensure feature agreement in noun phrases.

4 The Partial Parser for Greek

The model of FSTs with constraints has been employed in the creation of a partial parser for Modern Greek as a three level FST cascade. The first level locates and groups consecutive adjectives, as well as other auxiliary word groups, in order to make the task of subsequent levels easier. The second level performs the major part of the noun phrase recognition task. It resolves all types of noun phrases, without incorporating clitic forms of pronouns. This is due to the curious role of clitics in the noun phrase of Modern Greek. This level incorporates "stray" clitics (i.e., ones that are after a noun phrase and do not constitute a noun phrase by themselves) into the preceding noun phrases.

In the input text stream there are two types of ambiguities to be resolved: at the tag level and at the feature level. Resolution of a tag ambiguity takes place when a pattern matches only one of the tags assigned to the word. Ambiguities at the feature level happen when a symbol is accompanied by more than one feature tuples. These ambiguities are resolved during the transitions, since the FST actually calculates the intersection of tuple sets for consecutive symbols inside the sentence. The result of the intersection contains the disambiguated values of the features for both the word group recognized by the FST and the words themselves.

Each transducer of the parser has been implemented as a lex [8] program, and translated with flex, the implementation of the tool available for many platforms. The partial parser has been incorporated and used in a machine learning application for the identification of named entities in Greek business texts [9].

References

[1] Bourigault, D. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proc. of COLING-'92*, pages 977-981.

[2] Voutilainen, A. 1993. NPTool, a detector of English noun phrases. In *Proc. of the Workshop on Very Large Corpora*, pages 48-57.

[3] Abney, S. 1996. Part-of-speech tagging and partial parsing. In Ken Church, Steve Young and Gerrit Bloothooft, eds. *Corpus-based methods in Language and Speech*, Kluwer Academic Publishers, Dordrecht.

[4] Abney, S. 1996. Partial parsing via finite-state cascades. In *Proc. of the ESSLLI '96 Robust Parsing Workshop.*

[5] Church, K. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proc. of the 2nd Conference on Applied Natural Language Processing*, Austin, Texas.

[6] Chen, Kuang-hua and Hsin-His Chen. 1994. Extracting noun phrases from large-scale texts: A hybrid approach and its automatic evaluation. In *Proc. of the ACL94.*

[7] Mackridge, P. 1990. The Modern Greek Language (Greek translation). Patakis editions.

[8] Lesk, M. E. 1975. Lex - A Lexical Analyzer Generator. *Comp.Sci. Tech.Rep. No.39*, Bell Laboratories, New Jersey.

[9] Karkaletsis, V., et al. 1999. Named-Entity Recognition from Greek and English Texts. *Journal of Intelligent and Robotic Systems*, 23(5):123-135.