Workshop on Language Industries

Seoul, October 1997

The Construction of Local Grammars and the Automatic Analysis of Texts
Applications to Information Retrieval and Translation

Maurice Gross
Laboratoire d'Automatique Documentaire et Linguistique
Université Paris 7

Local grammars are finite-state automata, mostly without cycles applied to specific areas of language (M. Gross 1989, 1997):

- sets of semantically related lexical items (e.g. economic indices used on the *Stock Exchange*,
- sets of phrases, such as time adverbials and, more specifically, dates,
- sets of whole sentences that describe particular situations or events.

The example we are going to present corresponds to elementary sentences used to describe the movements of an index (i.e. the *Dow Jones industrial average*) on the *New York Stock Exchange*. Many other such sets can be built and sentences that express any kind of measurement (data from physics, economy, demography, etc.) can be described by the same methods.

Such grammars can then cover a large variety of texts. We have constructed detailed samples, for French and English, and we have used them to analyze automatically various corpora (daily newspapers in the example we are going to present). The grammars localize complex expressions in large corpora (E. Laporte 1994), they express semantic equivalences, hence they can be used to retrieve information in full texts data banks without previous indexing. They can also be completed by translations and the finite-state transducers (D. Perrin 1989) that are associated in this manner to the parsing grammars provide a tool for computer-aided translation.

The following documents are graphs that must be read from left to right, that is, from an initial state to a final state. Each path correspond to an utterance belonging to the automaton. In the parsing procedure, each path is going be matched with the text to be analyzed.

Boxes contains:
- sets of words explicily spelled out,

- sets of entries of various dictionaries: dictionaries of simple words or of compound words. In this case, a word between angles stands for its full morphological paradigm, for example <*go*> stands for the conjugated forms: *go, goes, went, going, gone* and their corresponding grammatical attributes (**Verb**, **Tense**, **Person**, **Mood**). Between angles one can also introduce grammatical constraints, for example <*go*:3s> signifies the forms of *go* in the third person of singular (i.e. *goes* and *went*). Dictionaries are automatically looked up in the parssing process M. Silberztein 1993, 1997),
- local grammars. The name of the corresponding automaton is in a shaded box.
An automaton with such shaded boxes can be viewed as an abbreviated form of an automaton where the shaded boxes have been replaced by the corresponding sub-automata. The sub-automata are called automatically in the parsing process.

Various problems arise in the construction process of large-scale grammars. We signal:

- linguistic problems: to what extend linguistic transformations can be applied automatically to a graph in order to generate the graph of transformed utterances? For example, we present various graphs containing only active sentences, can we generate automatically the graph of the corresponding passive sentences? In particular, transformations that permute items are not represented in a compact way by finite-state automata. How can we deal with a potential explosion of the number of automata? For examples, adverbial phrases can be moved to various locations of a sentence, do we build all the corresponding graphs?

- computational problems: large corpora and large automata require sophisticated algorithms in order to keep the parsing time within practical limits. Indexing of texts and the use of word frequency have allowed an important increase of the performances of parsing algorthms (J. Senellart 1996).

We show, on a sample of text, the forms which are recognized by the parsing procedure.

References

Gross, Maurice 1989a. La construction de dictionnaires électroniques, *Annales des Télécommunications*, Tome 44, No 1-2, pp. 4-16.

Gross, Maurice 1989b The use of finite automata in the lexical representation of natural language, *Electronic Dictionaries and Automata in Computational Linguistics*, Berlin-New York : Springer Verlag, pp.18-34.

Gross, Maurice 1997. The Construction of Local Grammars, In *Finite-State Language Processing*, E. Roche & Y. Schabes eds., Cambridge, Mass. : M.I.T. Press.

Laporte, Eric 1994. Experiments in Lexical Disambiguation Using Local Grammars, *Papers in Computational Lexicography* (COMPLEX), Budapest : Research Institute for Linguistics, Academie des Sciences de Hongrie, pp.163-172.

Perrin, Dominique 1989. Automates et algorithmes sur les mots, *Annales des Télécommunications*, Tome 44, No 1-2, pp. 20-33.

Senellart, Jean 1997. Indexing Texts for Fast Processing, To appear.

Silberztein, Max 1993. *Dictionnaires électroniques et analyse automatique de textes: le systÜme INTEX*. Paris: Masson, 233 p.

Silberztein, Max 1997. The Lexical Analysis of Natural Languages, In *Finite-State Language Processing*, E. Roche & Y. Schabes eds., Cambridge, Mass. : M.I.T. Press.