FASTUS: An Information Extraction System

Jerry R. Hobbs
and
David Israel
Artificial Intelligence Center
SRI International
333 Ravenswood Ave.
Menlo Park, California 94025
Tel: (650) 859-4254

1 Introduction

FASTUS is a (slightly permuted) acronym for Finite State Automaton Text Understanding System. It is a system for extracting information from free text in English, and other languages as well, for entry into a database, and potentially for other applications, such as Information Retrieval. It works essentially as a set of cascaded, nondeterministic finite state automata.

FASTUS is most appropriate for *information extraction* tasks, rather than full text understanding. That is, it is most effective for tasks where

- Not all of the text in the text stream are relevant, and not all of a relevant text is itself relevant.
- There is a pre-defined target representation that the information is to be mapped into.
- The subtle nuances of meaning and the writer's goals in writing the text are not of paramount interest.

2 The Structure of the FASTUS System

In simple terms, the operation of **FASTUS** is comprised of three steps.

1. Recognizing Phrases: Sentences are segmented into noun groups, verb groups, and other phrases.

2. Recognizing Patterns: The sequence of phrases produced in Step 2 is scanned for patterns of interest, and when they are found, corresponding "incident structures" are built.

3. Merging Incidents: Incident structures from different parts of the text are merged if they provide information about the same incident.

Many systems have been built to do pattern matching on strings of words. The crucial innovation in the **FASTUS** system has been separating that process into the two steps of recognizing phrases and recognizing patterns. Phrases can be recognized reliably with purely syntactic information, and they provide precisely the elements that are required for stating the patterns of interest.

The system is implemented in Lisp and runs on Unix and Linux platforms, on Macs, and on PCs running Windows or NT.

3 An Example

One recent application was to scan news reports and extract information about terrorist incidents. The following sentences occurred in one report:

1. Recognizing Phrases: Step 1 segments the sentences into the following phrases:

Noun Group:	Salvadoran President-elect
Name:	Alfredo Cristiani
Verb Group:	condemned
Noun Group:	the terrorist killing
Preposition:	of
Noun Group:	Attorney General
Name:	Roberto Garcia Alvarado
Conjunction:	and
Verb Group:	accused
Noun Group:	the Farabundo Marti National Liberation Front (FMLN)
Preposition:	of
Noun Group:	the crime

2. Recognizing Patterns: Two patterns are recognized in this sequence of phrases:

<Perpetrator> <Killing> of <HumanTarget>
<GovtOfficial> accused <PerpOrg> of <Incident>

Two corresponding incident structures are constructed:

Incident:	KILLING
Perpetrator:	"terrorist"
Confidence:	–
Human Target:	"Roberto Garcia Alvarado"

and

Incident:	INCIDENT
Perpetrator:	FMLN
Confidence:	Suspected or Accused by Authorities
Human Target:	–

3. Merging Incidents: These two incident structures are merged into a single incident structure, containing the most specific information from each.

Incident:	KILLING
Perpetrator:	FMLN
Confidence:	Suspected or Accused by Authorities
Human Target:	"Roberto Garcia Alvarado"

4 The Performance of FASTUS

FASTUS has always performed at or near the top on the evaluations of Information Extraction technology run by the government. In terms of speed, **FASTUS** can analyze an average length news article in under 10 seconds. This translates into 10,000 articles per day, on a single processor.

This fast run time translates directly into fast development time; we have also developed tools to make grammar writing easier.

5 Summary

The advantages of the **FASTUS** system are as follows:

- It is conceptually simple. **FASTUS** is a set of cascaded finite-state automata.
- The basic system has a relatively small footprint.
- It is effective.
- It has very fast run time. The average time for analyzing a message is less than 10 seconds.
- In part because of the fast run time, it has a very fast development time. This is also true because the system provides a very direct link between the texts being analyzed and the data being extracted.

The **FASTUS** system has been extended in the following ways:

- We have developed a convenient interface that will allow grammar writers to define patterns more easily.
- We have implemented a Japanese language version of **FASTUS**.
- We have adapted **FASTUS** for use as a high-precision Information/Document Retrieval engine.

We have applied the system in a number of domains, beyond that of terrorist activities: for example, joint-ventures, mergers and acquisitions, IPOs and other financing events, high-level corporate management changes,

For further details about the **FASTUS** system, contact David Israel at (650) 859-4254.