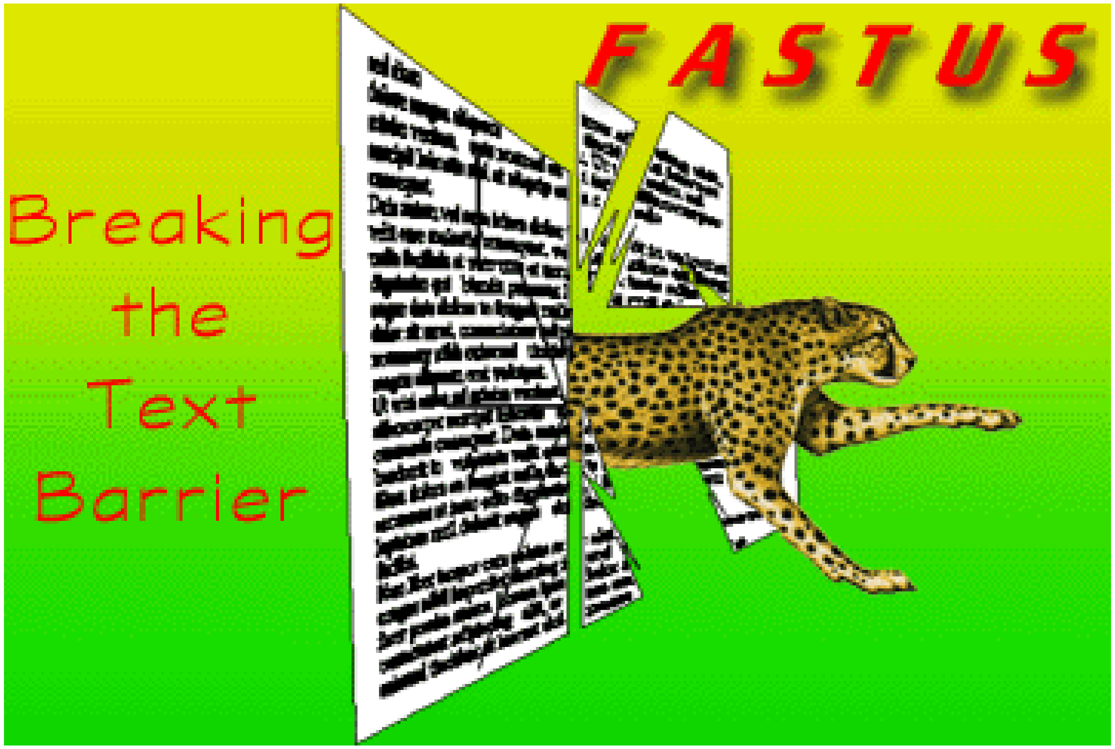
FASTUS



Jerry Hobbs, Doug Appelt, John Bear, David Israel,
Andy Kehler, David Martin, Karen Meyers,
Megumi Kameyama, Mark Stickel, Mabry Tyson

Information Extraction

- Simple, fixed definition of the information to be sought.
- Much, or even most, of the text is irrelevant to the information extraction goal.
- Large volumes of text need to be searched.

FASTUS Applications

- Business News: Joint Ventures (English and Japanese), Labor Negotiations, Management Succession
- Geopolitical News: Terrorist Incidents
- Military Messages: DARPA Message Handler
- Legal English: Document Analysis Tool
- Integration with OCR

FASTUS Sponsors

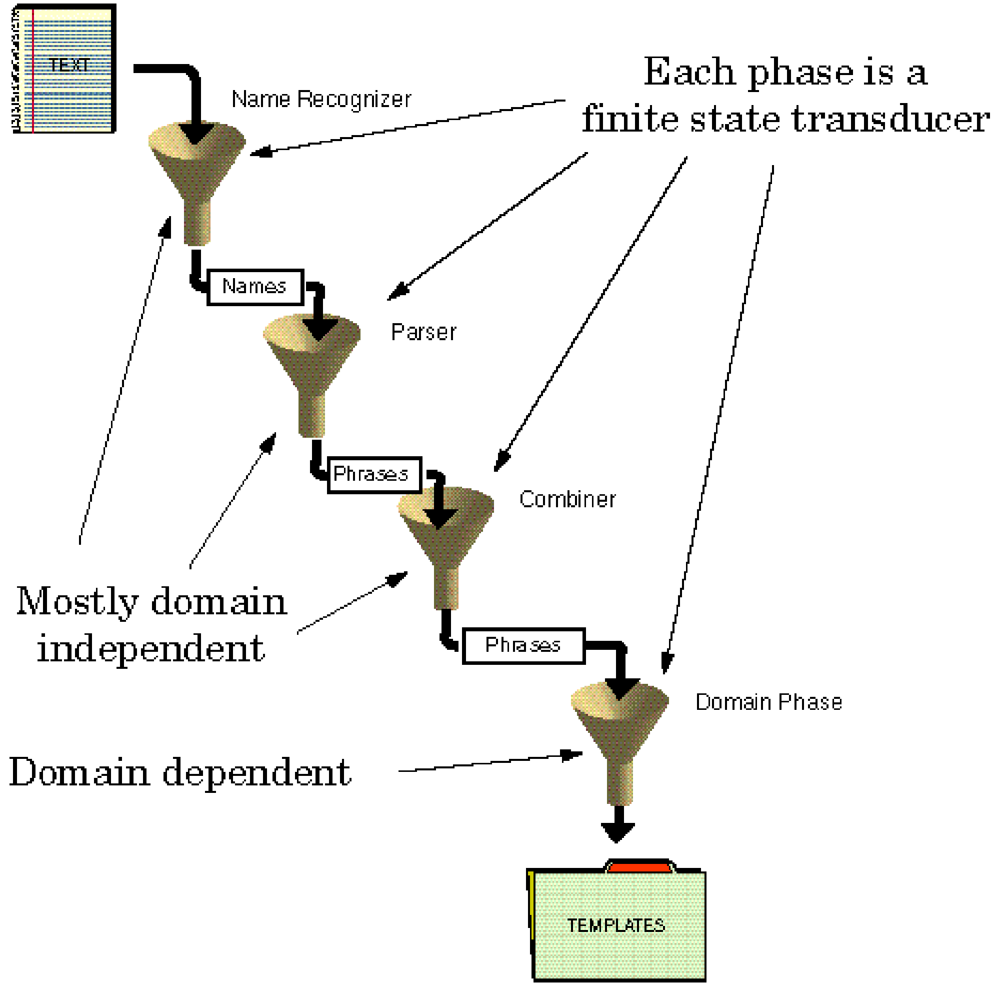
Human Language Technology
ISO Battlefield Awareness

TIPSTER IE Program

Office of Research and Development

FASTUS System

From Raw Text to the Analyst's File



Each phase is a finite state transducer

Mostly domain independent

Domain dependent

Name Recognition

Bridgestone Sports Co. C said **Friday** D it had set up a joint venture in **Taiwan** L with a local concern and a Japanese trading house to produce golf clubs to be shipped to **Japan**.L

Simple Phrasal Parsing

Bridgestone Sports Co.	said	Friday	it	has set up	a joint venture
Company-Name	VG	NG	NG	VG	NG

in	Taiwan	with	a local concern	and	a Japanese trading house
NG	Loc	P	NG	Conj	NG

to produce	golf clubs	to be shipped	to	Japan.
VG(Inf)	NG	VG(Inf,Pass)	P	Loc

Complex Phrases

Bridgestone Sports Co.	said	Friday	it	has set up	a joint venture
Company-Name	VG	NG	NG	VG	NG
	Complex VG				

in	Taiwan	with	a local concern	and	a Japanese trading house
NG	Loc	P	NG	Conj	NG
			Complex NG		

to produce	golf clubs	to be shipped	to	Japan.
VG(Inf)	NG	VG(Inf,Pass)	P	Loc

Recognizing Domain Patterns

- Match domain patterns against complex phrase heads
  - <company> <form><joint venture> with <company>
  - <company> capitalized at <currency>

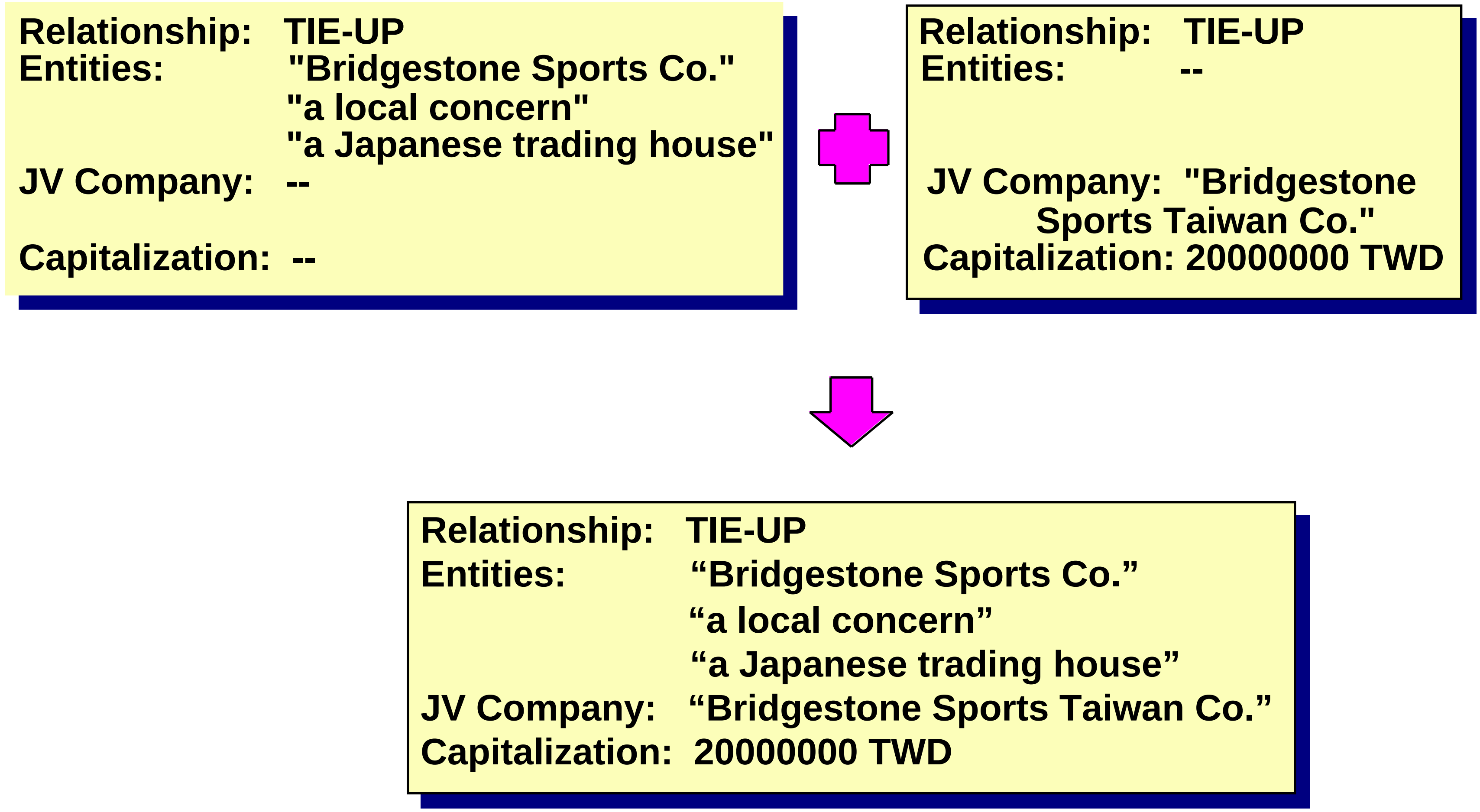
"Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

"The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990."

Relationship:	**TIE-UP**
Entities:	**"Bridgestone Sports Co."** **"a local concern"** **"a Japanese trading house"**
JV Company:	**--**
Capitalization:	**--**

Relationship:	**TIE-UP**
Entities:	**--**
JV Company:	**"Bridgestone Sports Taiwan Co."**
Capitalization:	**20000000 TWD**

Merging



Relationship: TIE-UP
Entities: "Bridgestone Sports Co."
"a local concern"
"a Japanese trading house"
JV Company: --
Capitalization: --

Relationship: TIE-UP
Entities: --
JV Company: "Bridgestone Sports Taiwan Co."
Capitalization: 20000000 TWD

Relationship: TIE-UP
Entities: “Bridgestone Sports Co.”
“a local concern”
“a Japanese trading house”
JV Company: “Bridgestone Sports Taiwan Co.”
Capitalization: 20000000 TWD

FASTUS Accuracy

- MUC-6 Evaluation
  - Name Recognition:
    - Recall 92% Precision 96% *(near human performance)*
  - Information Extraction:
    - Recall 44% Precision 61%
- Consistently among the best performing systems.

FASTUS Speed

- In 1994 MUC-4 evaluation -- extraction of information about terrorist events from newspaper articles
  - Processing time for 100 texts (SPARC-2)
    - 15.9 minutes of real time
  - Rate of text processing:
    - 2,375 words per minute
    - 1 text every 9.6 seconds
    - 9,000 texts per day

Making FASTUS more accessible to the analyst

- Declarative language for the specification of grammar rules.
- Meta rules to capture linguistic variants of subject-verb-object patterns.
- Automatic learning of patterns from examples.

Improving FASTUS Accuracy

- Using corpus statistics to evaluate competing hypotheses.
- Maintaining a lattice of hypotheses to avoid premature commitment to an analysis.
- Basic toolkit system for domain portability
  - Large Lexicon
  - Name Recognition
  - Basic English Grammar

Conclusions

- Fastus is a mature, robust, effective, efficient information extraction system.
- We now need
  - handling of broader domains
  - easier adaptation to user's needs
  - use in a greater variety of new applications