

CS544: Information Extraction

April 8, 2008

Jerry R. Hobbs

**USC/ISI
Marina del Rey, CA**

Three Tasks

1. Document Retrieval

Find probable relevant documents,
using keywords.



2. Information Extraction

Find relevant information: who did what
to whom, when and where.
But only the relevant information.

3. Text Understanding

Extract all information, relevant or not.
Understand writer's intention and
nuances of meaning.

Prehistory

Summer 1984: DARPA's Strategic Computing Program: SRI inference-based system; NYU NL-directed simulation; Unisys parsing of fragmentary texts

1985: Focus on 35 short CASREPs

May 1987: MUC-1: ~100 short naval msgs, demos of what systems could do (organized by Beth Sundheim)

May 1989: MUC-2: naval msgs, build system for training set, get 3 hrs to modify for 5 new texts

1990: Redirected by DARPA to 1700 long terrorist reports

Jun 1991: MUC-3, terrorist reports

Situation before MUC-4

UMass system (Lehnert et al.): Best performer in MUC-3; astonishingly simple technology

Pereira's implementation of finite-state approximations to context-free grammars; very fast

Inference-based systems took forever

Statistical systems didn't exist

Decision: Don't try to solve the NLU problem; solve the MUC problem in the shortest way possible

Examination of data --> cascaded finite-state transducers (SRI's FASTUS)

Later History

Dec 91 - mid 98: Tipster I, II and III programs

Jul 93: MUC-5: joint ventures, microelectronics

Mar 95: MUC-6 dryrun: labor negotiations

Sep 95: MUC-6: management succession

Nov 97: MUC-7 dryrun: plane crashes

Mar 98: MUC-7: rocket payloads

M1990s-E2000s: various commercialization efforts, including Inquest (SRA spinoff), Discern (SRI spinoff)

Evaluation

Sites given training corpus with keys, and task definition.

One month development time.

Sites run system on blind test set.

Measures:

Recall: Completeness, percent of answers the system gets right.

Precision: Correctness, percent of the system's answers that are right.

F-score: weighted mean of recall and precision

$$F = \frac{(b^2 + 1) P R}{b^2 P + R}$$

Recall, Precision, and Perjury

I swear to tell the truth,

the whole truth,

and nothing but the truth.

100%
Recall



100%
Precision



Uses of Information Extraction

Extraction of time-critical information from large volumes of text

- * for rapid searching of large amount of text, such as WWW, for more specific information than keywords can discriminate.**
- * for commercial, government, and military intelligence.**
- * for scientific literature searches.**
- * for building databases from large textual corpora, e.g., in biomedicine**

The Need for Information Extraction

Two factoids:

The major use for computers in business is probably for business intelligence.

90% of the data out there is in the form of natural language text.

WWW Searches: The average user does not scroll.

Organic chemistry factoid: If it costs less than \$100K to invent it, don't search the literature.

An Intelligence Analyst:

1990: "like reading all of *War and Peace* every day."

1995: "way beyond that"

Gulf War: "10,000 messages a day"

Pre-9/11: > 100,000 relevant messages

Biomedicine: 500,000 articles a year; large amounts of money spent on curatorial activities.

Example

"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.

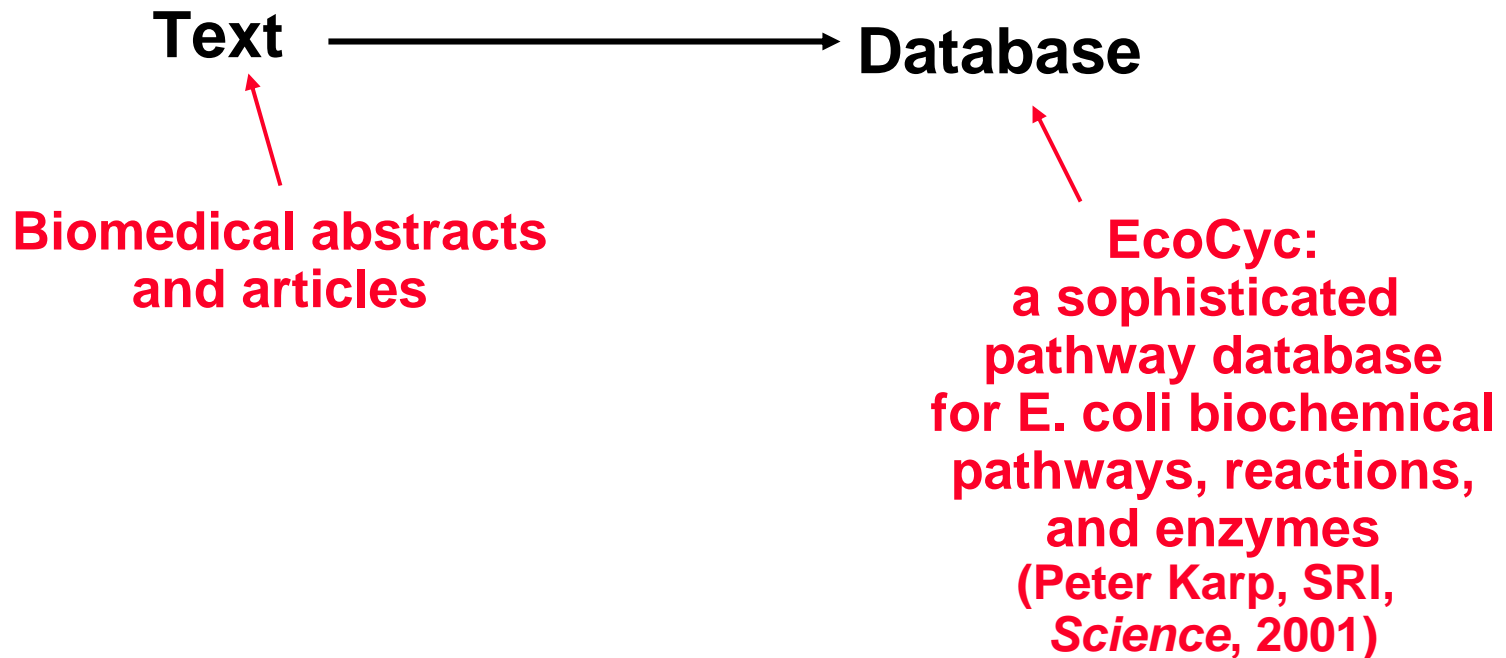
"The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990."

==>

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

Biomedical Domain: EcoCyc

Information Extraction:



(Simplified) Structure of EcoCyc

Reaction:
Pathway:
Reactant1:
Reactant2:
Enzyme Reaction:
Product1:
Product2:
.....

Enzymatic Reaction:
Enzyme:
Inhibitor:
Activator:
Cofactors:
.....

Enzyme (Protein):
Name:
Molecular Weight:
Subunit Composition: : #
Gene:
.....

.....

Example

“gamma-Glutamyl kinase, the first enzyme of the proline biosynthetic pathway, was purified to a homogeneity from an Escherichia coli strain resistant to the proline analog 3,4-dehydropoline. The enzyme had a native molecular weight of 236,000 and was apparently comprised of six identical 40,000-dalton subunits.”

==>

Reaction:
Pathway: proline
Enzyme:

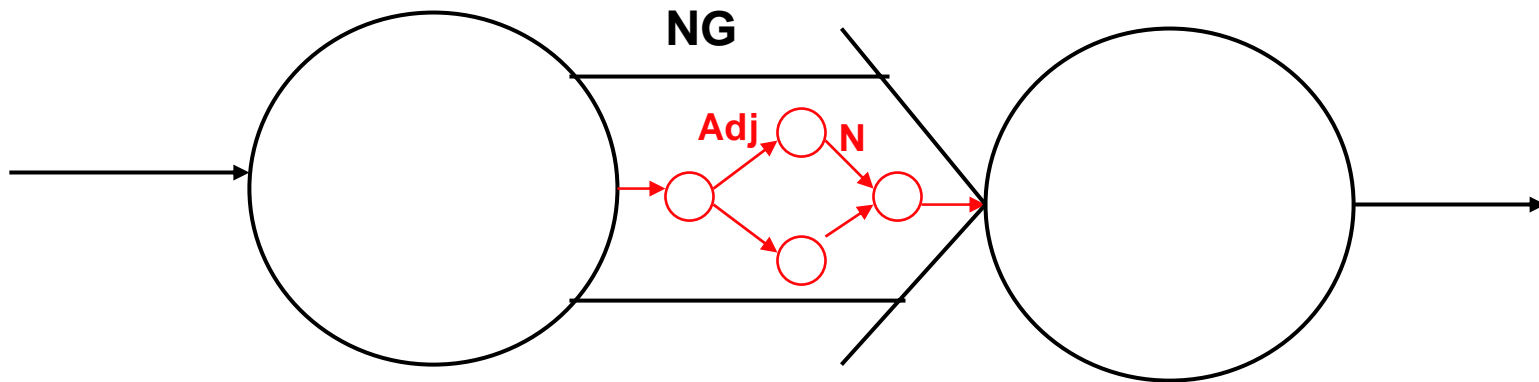
Enzyme:
Name: gamma-Glutamyl kinase
Molecular-Wt: 236,000
Subunit-Comp:
Subunit-Num: 6

Enzyme:
Name:
Molecular-Wt: 40,000
Subunit-Comp:
Subunit-Num:

Cascaded Finite-State Transducers

**Finite-state transducer = finite-state automaton
with output in final states**

**Cascaded: each transition of higher-level fsa is
realized by lower-level fsa**



Phases or Levels in Information Extraction

1. Name Recognition
2. Basic Phrase Recognition
3. Complex Phrase Recognition
4. Domain Event Recognition
5. Merging

not FST



1. Name Recognition

Bridgestone Sports Co._C said Friday_T 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

2. Recognizing Basic Phrases

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
P Loc P NG Conj NG

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

Noun Group: NP up through head noun

Verb Group: verb + auxiliaries, trapped adverbs, and particle

3. Recognizing 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

Complex NG: NG + possessive, appositive, of-PP, ...
Conjoined NGs

Complex VG: content VGs + “empty” verbs
Conjoined VGs

4. Recognizing Domain Patterns

As Patterns are recognized, Templates are built up.

"**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."

Company **Form** **Joint-Venture** with **Company**

Company Capitalized at Money

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:	2000000 TWD

5. Merging

"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

=

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

1. Named Entity Recognition

gamma-Glutamyl kinase_C, the first enzyme of the proline_C biosynthetic pathway, was purified to a homogeneity from an Escherichia coli_O strain resistant to the proline_C analog 3,4-dehydroproline_C.

The enzyme had a native molecular weight of 236,000 and was apparently comprised of six identical 40,000-dalton subunits.

Terminology is a huge challenge.

2. Recognizing Basic Phrases

gamma-Glutamyl kinase_C, the first enzyme of

NG

NG

P

the proline_C biosynthetic pathway, was purified to a homogeneity

NG

VG

P

NG

from an Escherichia coli_O strain resistant to the proline_C analog

P

NG

AdjG

P

NG

3,4-dehydroproline_C.

NG

The enzyme had a native molecular weight of 236,000

NG

VG

NG

P

NG

and was apparently comprised of

Conj

VG

six identical 40,000-dalton subunits.

NG

2. Entities from Basic Phrases

gamma-Glutamyl kinase, the first enzyme of
the **proline biosynthetic pathway**

Reaction:

Pathway: proline

Enzyme:

Enzyme:

Name: gamma-Glutamyl kinase

Molecular-Wt: --

Subunit-Comp: --

Subunit-Num: --

3. Recognizing Complex Phrases

gamma-Glutamyl kinase_C, the first enzyme of
NG NG P
the proline_C biosynthetic pathway,
NG
ComplexNG

was purified to a homogeneity
VG P NG

from an Escherichia coli_O strain resistant to
P NG AdjG P

the proline_C analog 3,4-dehydroproline_C.
NG NG
ComplexNG

The enzyme had a native molecular weight of 236,000
NG VG NG P NG

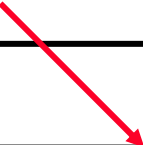
and was apparently comprised of
Conj VG

six identical 40,000-dalton subunits.
NG

3. Relations and Events from Complex Phrases

gamma-Glutamyl kinase, the first **enzyme of**
the proline biosynthetic **pathway**

Reaction:
Pathway: proline
Enzyme:



Enzyme:
Name: gamma-Glutamyl kinase
Molecular-Wt: --
Subunit-Comp: --
Subunit-Num: --

4. Recognizing Clause-Level Domain Patterns

The enzyme had a native molecular weight of 236,000 and was apparently comprised of six identical 40,000-dalton subunits.

Compound have Measure of Value
Compound comprised-of Compound

Enzyme:
Name: --
Molecular-Wt: 236,000
Subunit-Comp: --
Subunit-Num: --

Enzyme:
Name: --
Molecular-Wt: --
Subunit-Comp: --
Subunit-Num: 6

Enzyme:
Name: --
Molecular-Wt: 40,000
Subunit-Comp: --
Subunit-Num: --

4. Recognizing Clause-Level Domain Patterns

The enzyme had a native molecular weight of 236,000 and was apparently comprised of six identical 40,000-dalton subunits.”

Compound have Measure of Value
Compound comprised-of Compound

Enzyme:
Name: --
Molecular-Wt: 236,000
Subunit-Comp: --
Subunit-Num: --

Enzyme:
Name: --
Molecular-Wt: --
Subunit-Comp: --
Subunit-Num: 6

Enzyme:
Name: --
Molecular-Wt: 40,000
Subunit-Comp: --
Subunit-Num: --

5. Merging

“**gamma-Glutamyl kinase**, the first enzyme of the proline biosynthetic pathway, was purified to a homogeneity from an Escherichia coli strain resistant to the proline analog 3,4-dehydroproline. **The enzyme had a native molecular weight of 236,000** and was apparently **comprised of six identical 40,000-dalton subunits.**”

Enzyme:
Name: gamma-Glutamyl kinase
Molecular-Wt: --
Subunit-Comp: --
Subunit-Num: --

+

Enzyme:
Name: --
Molecular-Wt: 236,000
Subunit-Comp: --
Subunit-Num: --

+ Enzyme:
Name: --
Molecular-Wt: --
Subunit-Comp: --
Subunit-Num: 6

=

Enzyme:
Name: --
Molecular-Wt: 40,000
Subunit-Comp: --
Subunit-Num: --

Enzyme:
Name: gamma-Glutamyl kinase
Molecular-Wt: 236,000
Subunit-Comp: --
Subunit-Num: 6

Enzyme:
Name: --
Molecular-Wt: 40,000
Subunit-Comp: --
Subunit-Num: --

Syntactic Levels

Clause-level Event Recognition

Event

Complex Phrase Recognition

VG

Basic Phrase Recog

NG

VG

NG

NG

VG

NG

P

NG

P

NG

Name/ POS Recog

Name

V

NTime

Pro

Aux

V

P

Det

Adj

N

P

NLoc

P

Det

Adj

N

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern.

Information Extraction Systems: The Basic Idea

**A set of cascaded finite-state transducers,
recognizing successively larger structures.**

**Each level of analysis corresponds to a linguistic
“natural kind”.**

**Primarily syntactic knowledge is used at lower levels
where it can decide upon unambiguous readings.**

**Domain knowledge used at higher levels, where
syntactic ambiguity is rife.**

**Lower levels of analysis produce just the right objects
needed for stating higher-level patterns.**

Name Recognition

List of names: **IBM vs DNA**

Internal structure: **XYZ Corp., Universal Widgets**

Local context (when recognizing complex phrases):

XYZ's sales

Vaclav Havel, 53, president of the Czech Republic

Acronym Recognition

In Name Recognition phase:

Associates unknown words or words in parentheses with nearby phrases (not just names, Warbreaker)

Matching criteria:

One or more letters from matched words:

“Aluminum Company of America (ALCOA)”

“chemical exercise (CHEMEX)”

Allows skipping of some words.

Allows equivalences:

“field training exercises (FTX)”

“Iraqi Air Force (IZAF)”

Part of Speech Tagging?

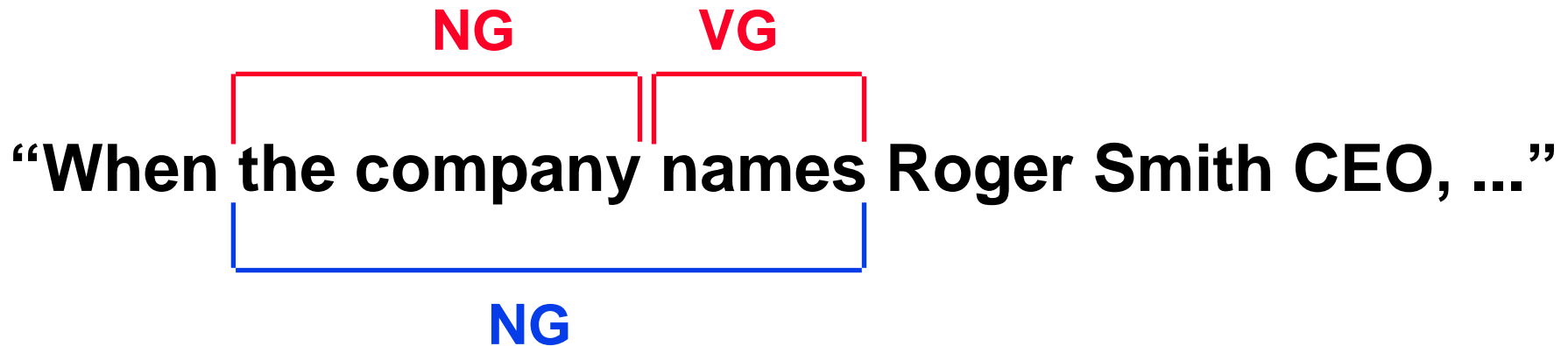
Without POS tagging, FASTUS got 96% correct.

POS tagger got 96% correct.

Both made same mistakes.

POS tagger doubled time.

Lattices for Ambiguity



Retain multiple parses at each phase, and let subsequent phases resolve ambiguity.

Complex Verb Groups

GM **formed** a joint venture with Toyota.

GM **announced it was forming** a joint venture with Toyota.

GM **signed an agreement forming** a joint venture with Toyota.

GM **announced it was signing an agreement to form** a joint venture with Toyota.

Terrorists **kidnapped and killed** three people.

Complex NG Coreference

Definite NGs:

Look for a previous word with the same head noun:

“the agreement” --> “an agreement”

Look for a previous object of the right domain-specific type:

“the Detroit automaker” --> “General Motors”, “a company”

No use of synonymy or sort hierarchy otherwise.

* “the agreement” --> “a contract”

Window of 10 sentences.

Complex NG Coreference

3rd Person Pronouns:

1. Scan current sentence L --> R to current word.
2. Scan previous sentence L--> R.
3. Scan remainder of window (2 more sentences)
R --> L.

“they” can be identified with Plural NG or Organization.

1st Person Pronoun:

- “I”, “me” --> nearest person (should check for speaking)
“we” --> plural person NG or organization.
Allow all of current sentence.

MUC-6 Results: R = 59%, P = 72%

Pseudo-Syntax at Clause Level

Subject {Preposition NounGroup}* **VerbGroup**

Subject Relpro {NounGroup | Particle}* VerbGroup
{NounGroup | Particle}* **VerbGroup**

Subject Relpro VerbGroup {NounGroup | Particle}* **VerbGroup**

The mayor, who was kidnapped yesterday, was found dead today

Subject VerbGroup {NounGroup | Particle}* **Conj** **VerbGroup**

The mayor was kidnapped yesterday and was found dead today

..... **Company-Name** “capitalized” **Money**

Merging in MUC-6

Method 1 (used in evaluation):

Merge if no slots are inconsistent.

R = 44, P = 61

Method 2 (tested afterwards):

Require exact match on at least one slot:

R = 36, P = 69

No constraints on distance.

But

“Roger Smith, president of GM, appointed president of Microsoft.”

Person: Roger Smith
Position: president
Org: GM

Person: ?X1
Position: ?
Org: ?

==>

Person: ?X1
Position: president
Org: Microsoft

Want to merge these
but no slots match

Interpreting Tables

- 1. Recognize existence of tables by alignment, spaces.**
- 2. Distribute subheads through subsumed items.**
- 3. Recognize types of entries, with same type throughout column.**
- 4. Limited use of headers to disambiguate types of entries.**
- 5. Hypothesize most plausible relation among items in the rows (shortest path in graph).**
- 6. Interpret pretabular sentence with parameters for table entries.**

Interpreting Tables: Example

FIELD EXERCISES WERE CONDUCTED BY THE FOLLOWING UNITS:

UNIT	HOME BASE	LOCATION
21 MAY 94:		
1ST MECH INF BN	FT SAM HOUSTON	LAFAYETTE
2ND MECH INF BN	FT LEWIS	BATON ROUGE
22 MAY 94:		
3RD MECH INF BN	MONTEREY	LAFAYETTE

Rules for Phrases

**VG --> VG2 Adv* V-en:1;
head = (obj 1);
active = T;
aspect = perf;;**

**“... could not really
have left.”**

VG2 --> VG1 “have”;;

**VG2 --> V[have]:1 (Not);
tense = (tense 1);;**

**VG1 --> Modal:1 (Not) Adv*;
tense = (tense 1);;**

Not --> “not”; negative = T;;

Rule for Event

Resume1:

**Event --> Event-Adj* NG[Org]:1 (Compl)
VG[Active, Resume-word]:2
NG[Talk-word]
{ “with” NG[Org]:3 | Event-Adj}*;
type = Talk;
parties = (List (obj 1) (obj 3));
talk-status = Bargaining;;**

“The pilots’ union resumed talks with American Airlines yesterday.”

Compile-Time Transformations

GM manufactures cars.

==>

Cars are manufactured by GM.

... GM, which manufactures cars. . . .

... cars, which are manufactured by GM. . . .

... cars manufactured by GM. . . .

GM is to manufacture cars.

Cars are to be manufactured by GM.

GM is a car manufacturer.

+ Insertion of optional time and place adverbials.

Compile-Time Transformations (Metarules)

- 1. Parameterized metarules specify the possible linguistic variations, expressed in terms of subject, verb, object, and semantics.**

e.g., active, passive, relative clauses, nominalizations

- 2. Domain-specific patterns that provide particular instantiations of the metarules.**

**e.g. Company Promotes Person to Position;
<semantics: what template to build>**

Basic Transformations

Event-Adj* NG[??subj]:1 VG[Active,??head]:2
NG[??obj]:3 {P[??prep] NG[??pobj] | Event-Adj}*;
semantics;; Active “The company resumed talks.”

==> Event-Adj* NG[??obj]:3 VG[Passive,??head]:2
{P[??prep] NG[??pobj] | Event-Adj}*;
semantics;; Passive “Talks were resumed.”

==> NG[??subj]:1 P{Relpro} VG[Active,??head]:2
NG[??obj]:3 {P[??prep] NG[??pobj] | Event-Adj}*;
semantics;; Relative Clause “The company which resumed talks ...”

==> (NG[??subj]:1 P[Gen]) NG[??head]:2
 (“of” NG[??obj]:3)
{P[??prep] NG[??pobj] | Event-Adj}*;
semantics;; Nominalization
“the company’s resumption of talks”

Middle Verbs

NG[??subj]:1 VG[Active,Middle,??head]:2
NG[??obj]:3 {P[??prep] NG[??pobj] | Event-Adj}*;
semantics;;

“They resumed the talks”

==> NG[??obj]:3 VG[Active,??head]:2
{P[??prep] NG[??pobj] | Event-Adj}*;
semantics;;

“The talks resumed”

Symmetric Verbs

Event-Adj* NG[??subj]:1 VG[Active,Symmetric,??head]:2
(NG[??obj]:3) {P[“with”] NG[??pobj]:4 | Event-Adj}*;
semantics;;

“The union met with the company.”

==> Event-Adj* NG[??subj]:1 “and” NG[??pobj]:4
VG[Active,??head]:2 (NG[??obj]:3) Event-Adj*;
semantics;;

“The union and the company met.”

==> NG[??head]:2 (“of” NG[??obj]:3)
{P[“between”] NG[??subj]:1 “and” NG[??pobj]:4
| Event-Adj}*;
semantics;;

“the meeting between the union and
the company”

Domain Event Specification

Transformations: Middle, Basic;

1: Subj = Org;

2: Head = Resume-word;

3: Obj = Talk-word;

Prep = “with”;

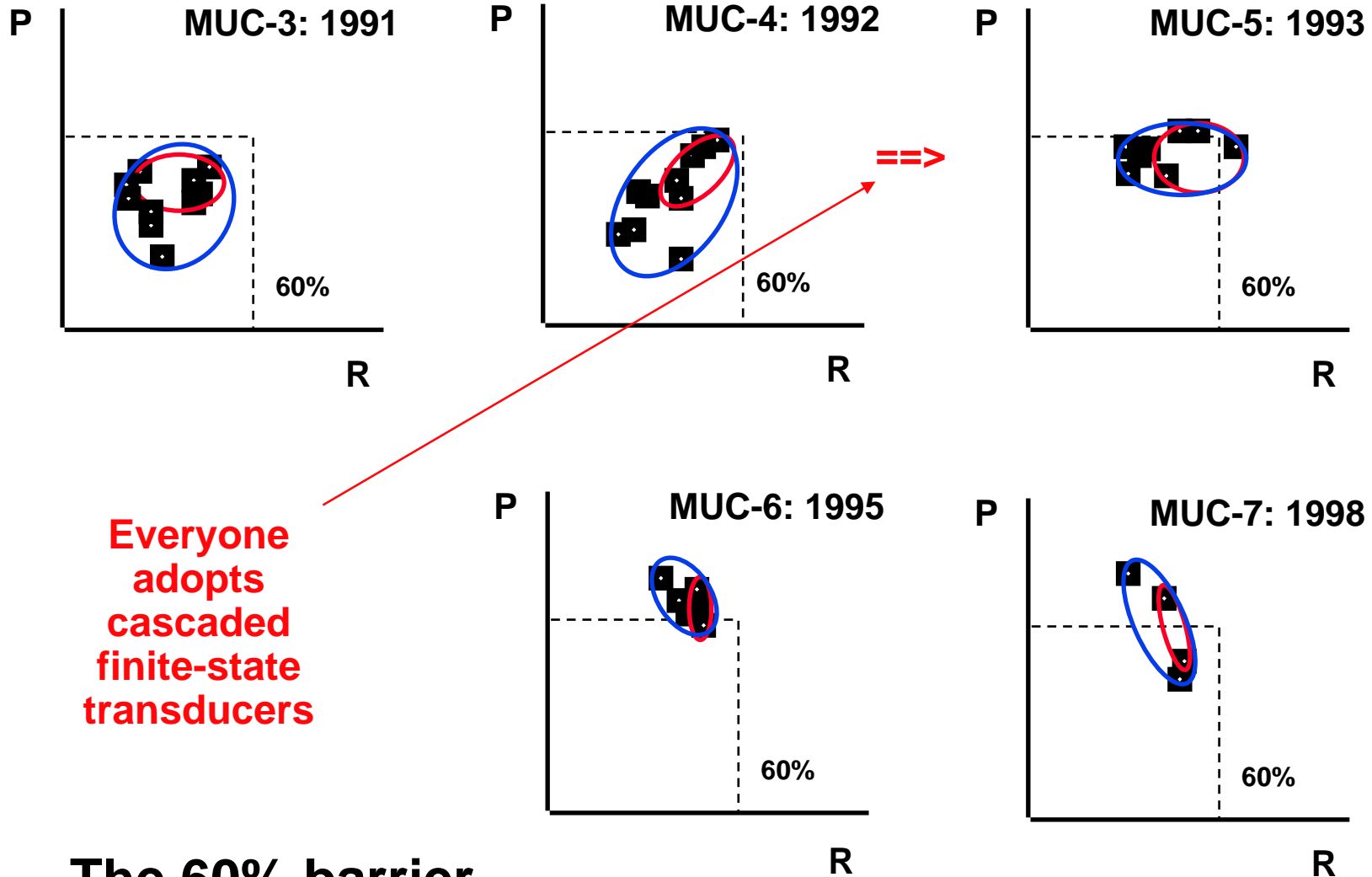
4: PObj = Org;

Semantics = < type = Talk;

parties = (List (obj 1) (obj 4));

talk-status = Bargaining;; >

How Did the Field Progress?



The 60% barrier

Open Domain

Open Domain Library of Patterns:

150 most common verbs/events in business news.

**<Legal-Person-1> sells <Financial-Instrument>
to <Legal-Person-2> for <Money>**

**User can access this pattern and further specify
the arguments:**

<Legal-Person-2>: Japanese corporation

**<Financial-Instrument>: American entertainment
stock**

Clusters of event types:

“sell” ==> “buy”, “trade”, “acquire”, ...

Open Domain: Methodology

“Health costs add **\$700** to the price of each of its cars, about \$300 to \$500 more per car than foreign **competitors pay** for **health**.”

“A **client pays** a **fee** to a **bank** for custom-tailored **protection** against adverse interest-rate swings.”

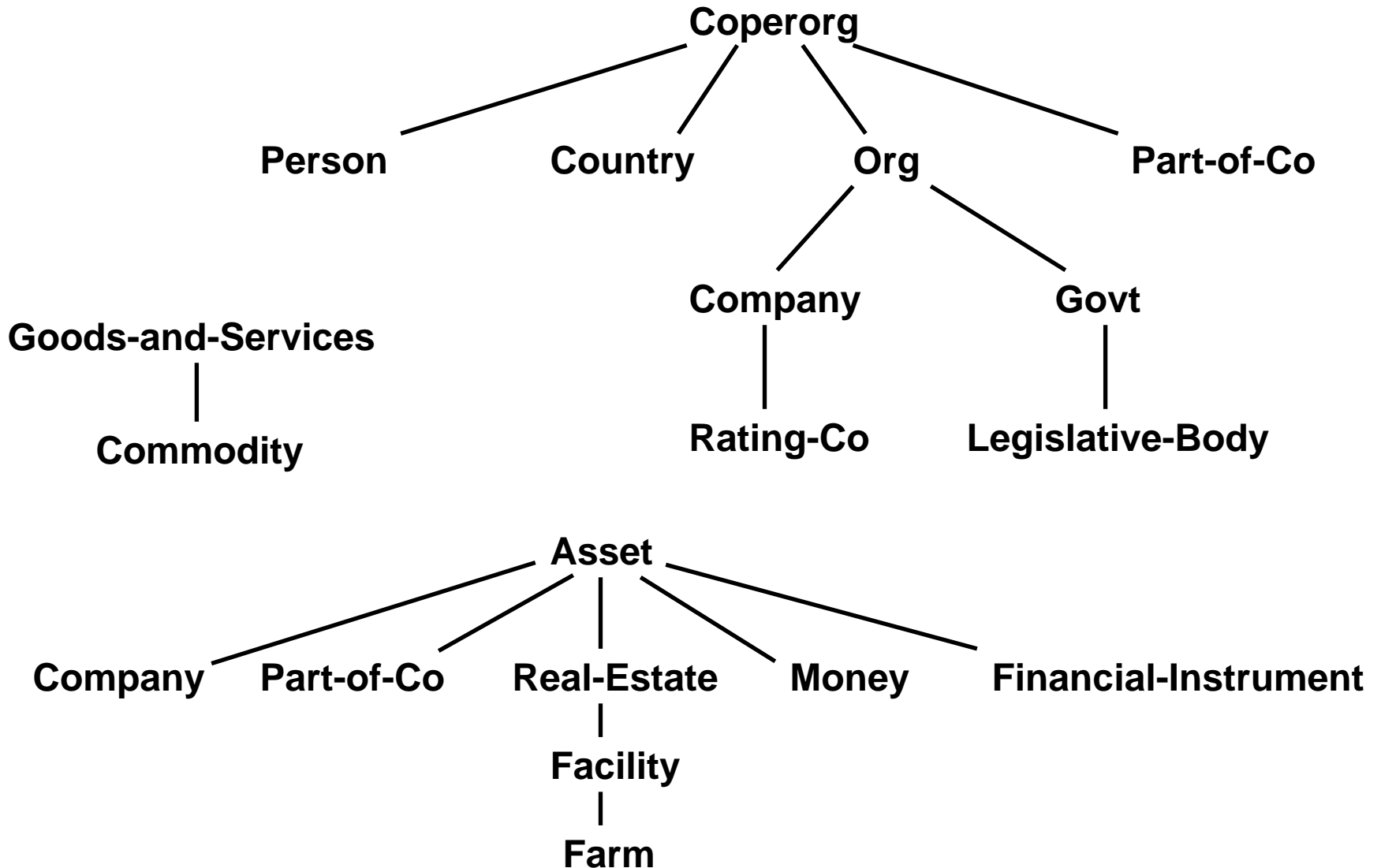
“**DuPont** agreed to **pay \$4.5 million** for **rights** to superconductor work.”

“**Manville** has offered to **pay** the **trust \$500 million** for a majority of the convertible preferred **stock**.”

“**Manville** ... its cash flow has lagged behind **its payments** to **victims**.”

Legal-Person pays Money to Legal-Person for Any

Open Domain: Basic Ontology



Open Domain: Basic Patterns

Person **analyzes** { Industry | Commodity | Financial-Instrument }

Coperorg **buys** { Company | Financial-Instrument | Goods }
from Coperorg for Money

{ Company | Person } **controls** Company

{ Company | Person } **earns** Money { for | from } Goods-or-Services

{ Company | Country } **exports** Goods to Country

Coperorg **invests** Money in { Financial-Instrument | Market | Country |
Company }

{ Company | Person } **manages** { Money | Financial-Instrument |
Company | Part-of-Co }

Unsupervised Learning of Patterns

(Grishman & Yangarber)

Seed system with two or three patterns

**Process parsed corpus for subject-verb-object patterns
that co-occur preferentially with existing patterns**

Add those patterns

Cycle until no new patterns

**Results: Can do slightly better in 12 hours than person
can do in one month**

Why the 60% Barrier

1. Merging problems accounted for 60% of our errors.
2. Entity recognition performance is 90%; event recognition requires recognizing ~4 entities; $.9^4 = .6$
3. The distribution of problems has a very long tail.
4. 60% is what the text wears on its sleeve; the rest is implicit and requires inference and world knowledge.