Example of IE from MUC/FASTUS (1993): it isn't easy!

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

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Example of IE: FASTUS(1993)

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

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Example of IE: FASTUS(1993)

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

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Example of IE: FASTUS(1993)

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

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Example of IE: FASTUS(1993)

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

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Examples of Existing IE Systems

- Systems to summarize medical patient records by extracting diagnoses, symptoms, physical findings, test results, and therapeutic treatments.
- Gathering earnings, profits, board members, etc. from company reports
- Verification of construction industry specifications documents (are the quantities correct/reasonable?)
- Classified advertisements: jobs, real estate, etc.
- Building museum database of artefacts from descriptions
- Extraction of company take-over events
- Extracting gene drug interactions from biomed texts

Three generations of IE systems

- Hand-Built Systems – Knowledge Engineering [1980s–]
  - Rules written by hand
  - Require experts who understand both the systems and the domain
  - Iterative guess-test-tweak-repeat cycle
- Automatic, Trainable Rule-Extraction Systems [1990s–]
  - Rules discovered automatically using predefined templates, using methods like ILP
  - Require huge, labeled corpora (effort is just moved!)
- Statistical Generative Models [1997 –]
  - One decodes the statistical model to find which bits of the text were relevant, using HMMs or statistical parsers
  - Learning usually supervised; may be partially unsupervised

FASTUS

Based on finite state automata (FSA) transductions

set up
new Taiwan dollars

1.Complex Words:
Recognition of multi-words and proper names

a Japanese trading house had set up

2.Basic Phrases:
Simple noun groups, verb groups and particles

production of 20, 000 iron and metal wood clubs

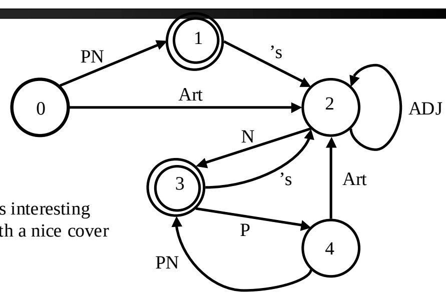
3.Complex phrases:
Complex noun groups and verb groups

[company]
[set up]
[Joint-Venture]
with
[company]

4.Domain Events:
Patterns for events of interest to the application
Basic templates are to be built.

5. Merging Structures:
Templates from different parts of the texts are merged if they provide information about the same entity or event.

Grep++

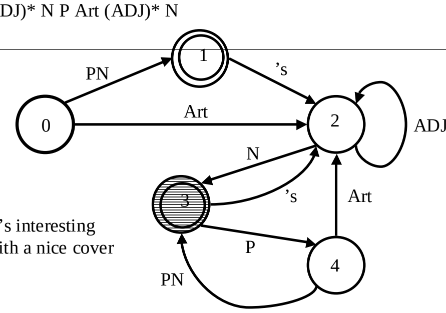


John’s interesting book with a nice cover

Pattern-maching

{PN ’s | Art}(ADJ)* N (P Art (ADJ)* N)*

PN ’s (ADJ)* N P Art (ADJ)* N



John’s interesting book with a nice cover

Example of IE: FASTUS(1993)

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

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 “metal wood” clubs a month.

1.Complex words

Attachment Ambiguities are not made explicit

2.Basic Phrases:

Bridgestone Sports Co.	: Company name
said	: Verb Group
Friday	: Noun Group
it	: Noun Group
had set up	: Verb Group
a joint venture	: Noun Group
in	: Preposition
Taiwan	: Location

Example of IE: FASTUS(1993)

[COMPANY] said Friday it [SET-UP] [JOINT-VENTURE] in [LOCATION] with [COMPANY] and [COMPANY] to produce [PRODUCT] to be supplied to [LOCATION].

[JOINT-VENTURE], [COMPANY] capitalized at 20 million [CURRENCY-UNIT] [START] production in [TIME] with production of 20,000 [PRODUCT] a month.

2.Basic Phrases:

Bridgestone Sports Co.	: Company name
said	: Verb Group
Friday	: Noun Group
it	: Noun Group
had set up	: Verb Group
a joint venture	: Noun Group
in	: Preposition
Taiwan	: Location

3.Complex Phrases

Some syntactic structures like …

Rule-based Extraction Examples

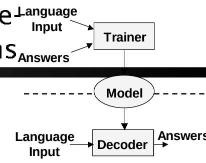
Determining which person holds what office in what organization

- [person] , [office] *of* [org]
  - Vuk Draskovic, leader of the Serbian Renewal Movement
- [org] (named, appointed, etc.) [person] P [office]
  - NATO appointed Wesley Clark as Commander in Chief

Determining where an organization is located

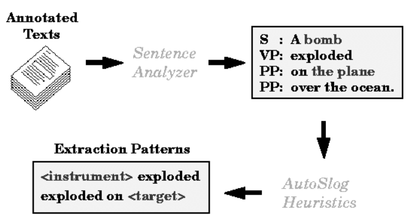
- [org] *in* [loc]
  - NATO headquarters in Brussels
- [org] [loc] (*division*, *branch*, *headquarters*, etc.)
  - KFOR Kosovo headquarters

Automatic Frame Learning Systems



- Pros:
  - Portable across domains
  - Tend to have broad coverage
  - Robust in the face of degraded input.
  - Automatically finds appropriate statistical patterns
  - System knowledge not needed by those who supply the domain knowledge.
- Cons:
  - Annotated training data, and lots of it, is needed.
  - Isn't necessarily better or cheaper than hand-built sol'n
- Classic e.g.: Riloff et al., AutoSlog (UMass) learns lexico-syntactic extraction patterns from templates

Autoslog (Riloff)



Rapier [Califf & Mooney, AAAI-99]

- Rapier learns three regex-style patterns for each slot:
  ▲Pre-filler pattern ▲ Filler pattern ▲ Post-filler pattern
- One of several recent trainable IE systems that incorporate linguistic constraints. (See also: **SIFT** [Miller *et al*, MUC-7]; **SRV** [Freitag, AAAI-98]; **Whisk** [Soderland, MLJ-99].)

"…paid $11M for the company…"
"…sold to the bank for an undisclosed amount…"
"…paid Honeywell an undisclosed price…"

Pre-filler:	Filler:	Post-filler:
1) tag: {nn,nnp}	1) word: undisclosed	1) sem: price
2) list: length 2	tag: jj	

RAPIER rules for extracting "**transaction price**"

Part-of-speech tags & Semantic classes

- Part of speech: syntactic role of a specific word
  - noun (nn), proper noun (nnp), adjectve (jj), adverb (rb), determiner (dt), verb (vb), "." ("."), …
  - NLP: Well-known algorithms for automatically assigning POS tags to English, French, Japanese, … (>95% accuracy)
- Semantic Classes: Synonyms or other related words
  - "Price" class: price, cost, amount, …
  - "Month" class: January, February, March, …, December
  - "US State" class: Alaska, Alabama, …, Washington, Wyoming
  - WordNet: large on-line thesaurus containing (among other things) semantic classes

Rapier rule matching example

"…sold to the bank for an undisclosed amount…"
POS: vb pr det nn pr det jj nn
SClass: price

Pre-filler:	Filler:	Post-filler:
1) tag: {nn,nnp}	1) word: undisclosed	1) sem: price
2) list: length 2	tag: jj	

"…paid Honeywell an undisclosed price…"
POS: vb nnp det jj nn
SClass: price

Rapier Rules: Details

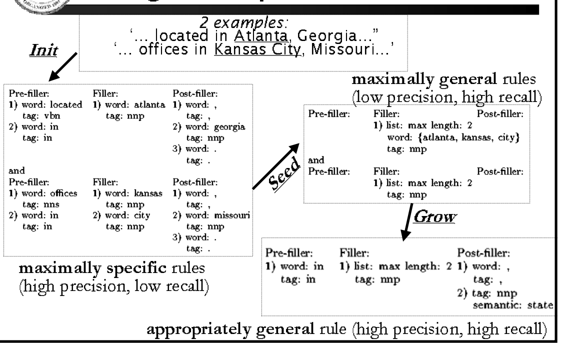
- Rapier rule :=
  - pre-filler pattern
  - filler pattern
  - post-filler pattern
- pattern := subpattern +
- subpattern := constraint +
- constraint :=
  - *Word* - exact word that must be present
  - *Tag* - matched word must have given POS tag
  - *Class* - semantic class of matched word
  - Can specify disjunction with "{...}"
  - *List length N* - between 0 and N words satisfying other constraints

Pre-filler:	Filler:	Post-filler:
1) tag: {nn,nnp}	1) word: undisclosed	1) sem: price
2) list: length 2	tag: jj	

Rapier's Learning Algorithm

- **Input**: set of training examples (list of documents annotated with "extract this substring")
- **Output**: set of rules
- ***Init***: Rules = a rule that exactly matches each training example
- Repeat several times:
  - ***Seed***: Select M examples randomly and generate the K most-accurate maximally-general filler-only rules (prefiller = postfiller = "true").
  - ***Grow***:
    Repeat For N = 1, 2, 3, ...
    Try to improve K best rules by adding N context words of prefiller or postfiller context
  - ***Keep***:
    Rules = Rules ∪ the best of the K rules – subsumed rules

Learning example (one iteration)



Information Extraction: Learning Lexico-Syntactic Patterns (Autoslog-TS, Riloff)

- Reduces greatly the needed resources: just classifications
- Start with known relevant articles; Extract noun phrases
- Create huge set of patterns of form:
  - **<subj> passive-verb** <victim> was murdered
  - **active-verb <dobj>** bombed <target>
- Use: key lexical items, syntactic frames, semantic sorts
- Compute relevance rate:
  - Pr(relevant|text has pattern$_i$) = rel-freq$_i$ / total-freq$_i$
- Rank patterns in order of importance
  - relevance rate * log2(freq$_i$), rr > 0.5
- Human judge reviewed top patterns
- Apply all patterns to each text. Worked almost as well

Statistical generative models

- Pros:
  - Well-understood underlying statistical model makes it easy to used wide range of tools from statistical decision theory
  - Portable across domains
  - Tend to have broad coverage, robustness, good recall
  - Data-driven
- Cons:
  - Range of features and patterns available may be limited
  - Not necessarily as good for complex multi-slot patterns
- Good current e.g.: Freitag & McCallum (CMU, JustSystems, Whizbang! Labs) statistically-learned models using Bayes classifiers, HMMs, etc.

Freitag and McCallum details

- Partly fixed structure, partly hidden (constrained EM using remote supervision)
- Parameter tying and shrinkage smoothing techniques
  - Better just to use a good unknown model?
- Structure learning of transition structure
  - Why not just plain EM?
- Results good on semi-structured data
  - Still rather modest on free form text
    - Need richer model class?