Relation Extraction and Machine Learning for IE

Feiyu Xu

feiyu@dfki.de

Language Technology-Lab
DFKI, Saarbrücken

Relation in IE

Information Extraction: A Pragmatic Approach

- Identify the types of entities that are relevant to a particular task
- Identify the range of facts that one is interested in for those entities
- Ignore everything else

[Appelt, 2003]

Message Understanding Conferences [MUC-7 98]

- U.S. Government sponsored conferences with the intention to coordinate multiple research groups seeking to improve IE and IR technologies (since 1987)
- defined several generic types of information extraction tasks (MUC Competition)
- MUC 1-2 focused on automated analysis of military messages containing textual information
- MUC 3-7 focused on information extraction from newswire articles
  - terrorist events
  - international joint-ventures
  - management succession event

Evaluation of IE systems in MUC

- Participants receive description of the scenario along with the annotated *training corpus* in order to adapt their systems to the new scenario (1 to 6 months)
- Participants receive new set of documents (*test corpus*) and use their systems to extract information from these documents and return the results to the conference organizer
- The results are compared to the manually filled set of templates
  (*answer key*)

Evaluation of IE systems in MUC

- precision and recall measures were adopted from the information retrieval research community

$$recall = \frac{N_{correct}}{N_{key}} \qquad precision = \frac{N_{correct}}{N_{correct} + N_{incorrect}}$$

$$F = \frac{(\beta^2 + 1) \times precision \times recall}{\beta^2 \times precision + recall}$$

- Sometimes an *F*-meassure is used as a combined recall-precision score

Generic IE tasks for MUC-7

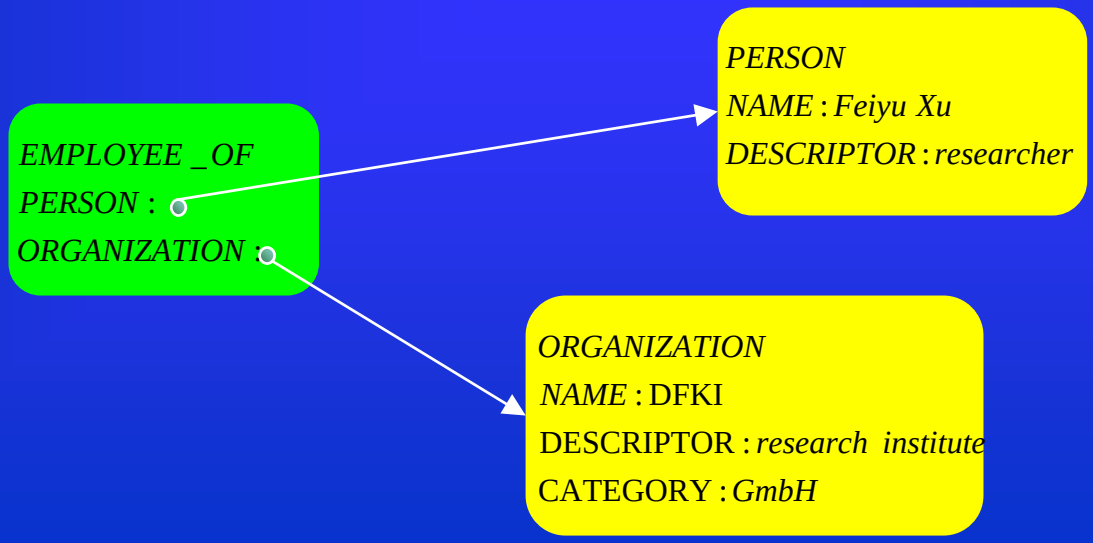
- (NE) Named Entity Recognition Task requires the identification an classification of named entities
  - organizations
  - locations
  - persons
  - dates, times, percentages and monetary expressions

- (TE) Template Element Task requires the filling of small scale templates for specified classes of entities in the texts
  - Attributes of entities are slot fills (identifying the entities beyond the name level)
  - Example: Persons with slots such as name (plus name variants), title, nationality, description as supplied in the text, and subtype.

    “Capitan Denis Gillespie, the comander of Carrier Air Wing 11”

Generic IE tasks for MUC-7

- (TR) Template Relation Task requires filling a two slot template representing a binary relation with pointers to template elements standing in the relation, which were previously identified in the TE task
  - subsidiary relationship between two companies (employee_of, product_of, location_of)



Generic IE tasks for MUC-7

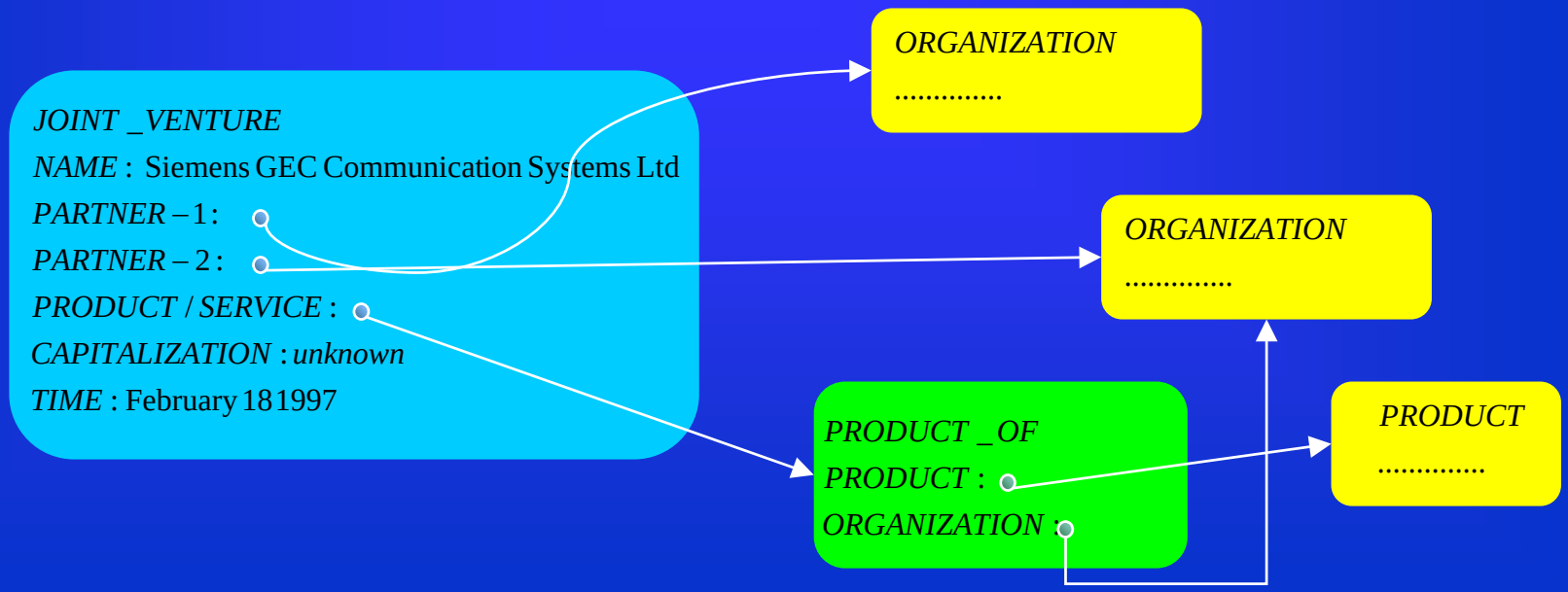
- (CO) Coreference Resolution requires the identification of expressions in the text that refer to the same object, set or activity
  - variant forms of name expressions
  - definite noun phrases and their antecedents
  - pronouns and their antecedents

"The U.K. satellite television broadcaster said its subscriber base grew 17.5 percent during the past year to 5.35 million"

- bridge between NE task and TE task

Generic IE tasks for MUC-7

- (ST) Scenario Template requires filling a template structure with extracted information involving several relations or events of interest
  - intended to be the MUC approximation to a real-world information extraction problem
  - identification of partners, products, profits and capitalization of joint ventures



Tasks evaluated in MUC 3-7 [Chinchor, 98]

EVAL\TASK	NE	CO	RE	TR	ST
MUC-3					YES
MUC-4					YES
MUC-5					YES
MUC-6	YES	YES	YES		YES
MUC-7	YES	YES	YES	YES	YES

Maximum Results Reported in MUC-7

MEASSURE\TASK	NE	CO	TE	TR	ST
RECALL	92	56	86	67	42
PRECISION	95	69	87	86	65

MUC and Scenario Templates

- Define a set of “interesting entities”
  - Persons, organizations, locations…
- Define a complex scenario involving interesting events and relations over entities
  - Example:
    - *management succession*:
      - persons, companies, positions, reasons for succession
- This collection of entities and relations is called a “scenario template.”

[Appelt, 2003]

Problems with Scenario Template

- Encouraged development of highly domain specific ontologies, rule systems, heuristics, etc.
- Most of the effort expended on building a scenario template system was not directly applicable to a different scenario template.

[Appelt, 2003]

Addressing the Problem

- Address a large number of smaller, more focused scenario templates (Event-99)
- Develop a more systematic ground-up approach to semantics by focusing on elementary entities, relations, and events (ACE)

[Appelt, 2003]

The ACE Program

- “Automated Content Extraction”
- Develop core information extraction technology by focusing on extracting specific semantic entities and relations over a very wide range of texts.
- Corpora: Newswire and broadcast transcripts, but broad range of topics and genres.
  - Third person reports
  - Interviews
  - Editorials
  - Topics: foreign relations, significant events, human interest, sports, weather
- Discourage highly domain- and genre-dependent solutions

[Appelt, 2003]

Components of a Semantic Model

- Entities - Individuals in the world *that are mentioned in a text*
  - Simple entities: singular objects
  - Collective entities: sets of objects of the same type *where the set is explicitly mentioned in the text*
- Relations – Properties that hold of tuples of entities.
- Complex Relations – Relations that hold among entities and relations
- Attributes – one place relations are attributes or individual properties

Components of a Semantic Model

- Temporal points and intervals
- Relations may be timeless or bound to time intervals
- Events – A particular kind of simple or complex relation among entities involving a change in at least one relation

Relations in Time

- timeless attribute: gender(x)
- time-dependent attribute: age(x)
- timeless two-place relation: father(x, y)
- time-dependent two-place relation: boss(x, y)

Relations vs. Features or Roles in AVMs

- Several two place relations between an entity *x* and other entities y_i can be bundled as properties of x. In this case, the relations are called roles (or attributes) and any pair <relation : y_i> is called a role assignment (or a feature).
- name <x, CR>

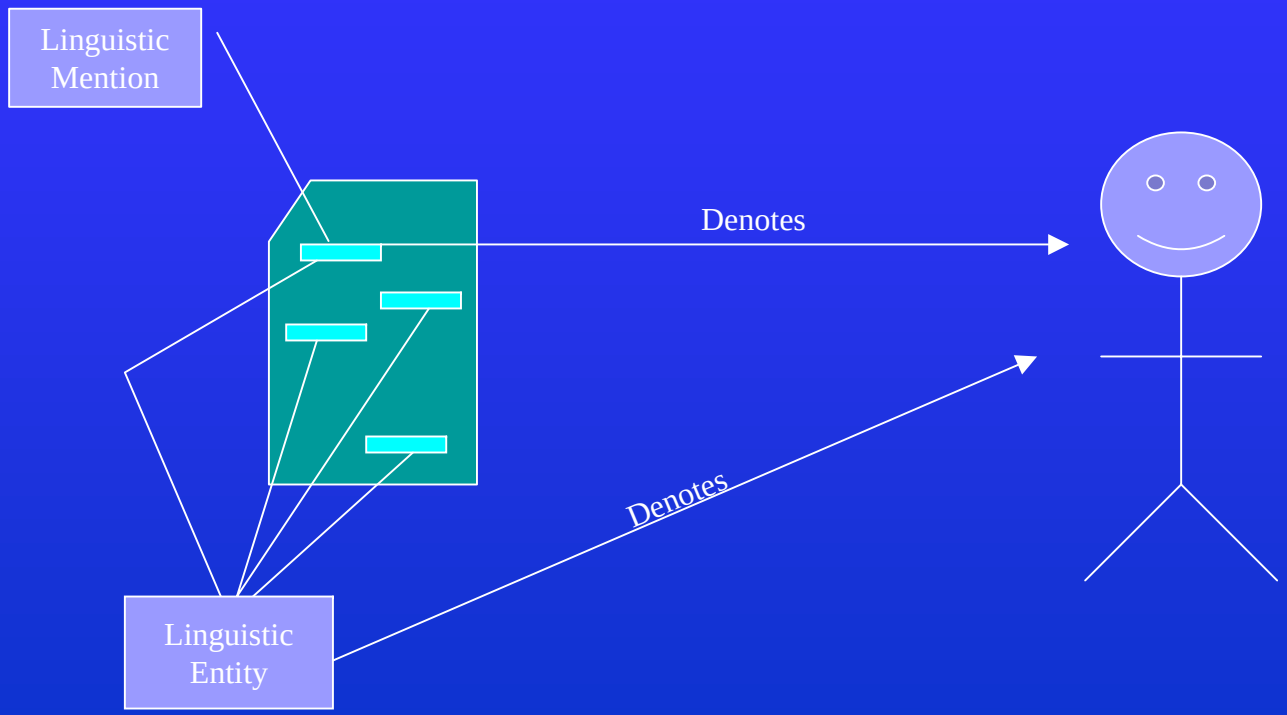
name: Condoleezza Rice
office: National Security Advisor
age: 49
gender: female

Semantic Analysis: Relating Language to the Model

- Linguistic Mention
  - A particular linguistic phrase
  - Denotes a particular entity, relation, or event
    - A noun phrase, name, or possessive pronoun
    - A verb, nominalization, compound nominal, or other linguistic construct relating other linguistic mentions
- Linguistic Entity
  - Equivalence class of mentions with same meaning
    - Coreferring noun phrases
    - Relations and events derived from different mentions, but conveying the same meaning

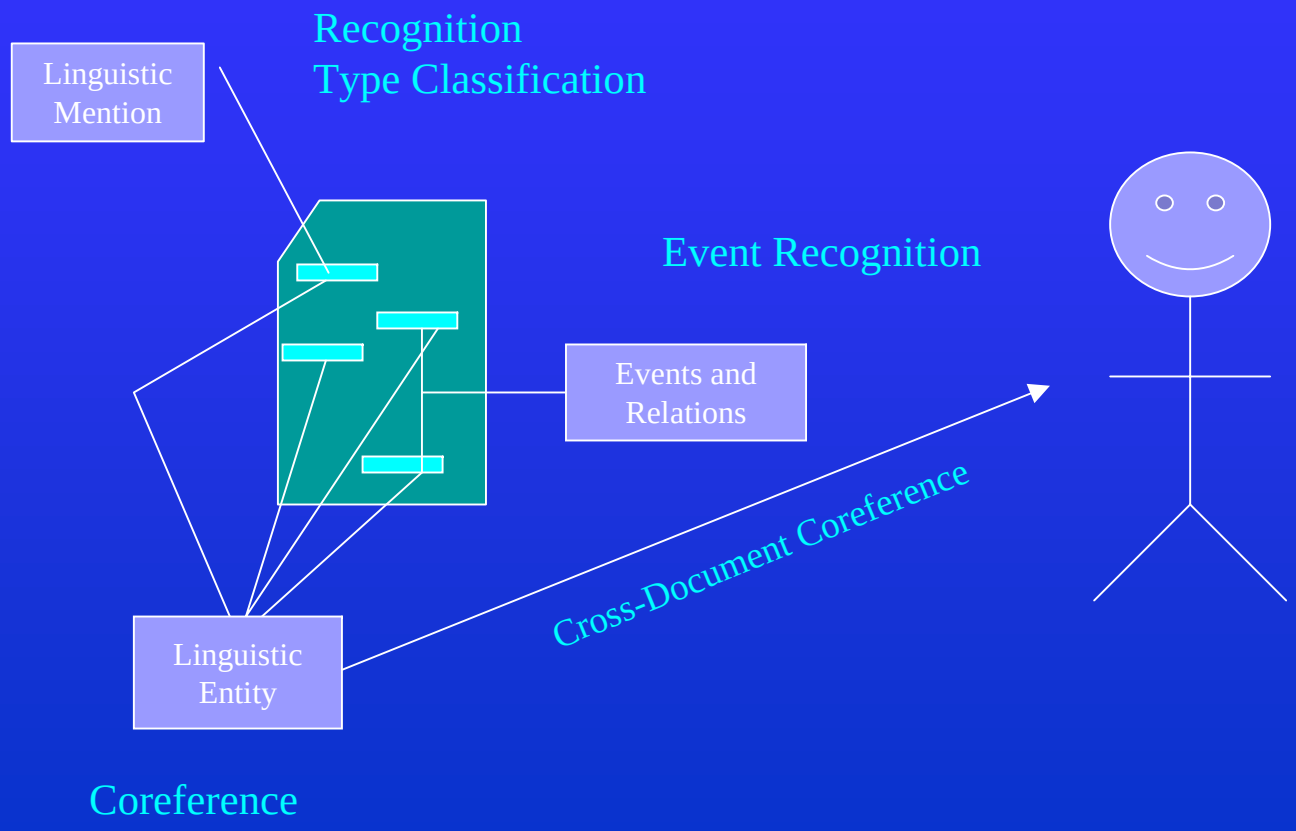
[Appelt, 2003]

Language and World Model



[Appelt, 2003]

NLP Tasks in an Extraction System



[Appelt, 2003]

The Basic Semantic Tasks of an IE System

- Recognition of linguistic entities
- Classification of linguistic entities into semantic types
- Identification of coreference equivalence classes of linguistic entities
- Identifying the actual individuals that are mentioned in an article
  - Associating linguistic entities with predefined individuals (e.g. a database, or knowledge base)
  - Forming equivalence classes of linguistic entities from different documents.

[Appelt, 2003]

The ACE Ontology

- Persons
  - A natural kind, and hence self-evident
- Organizations
  - Should have some persistent existence that transcends a mere set of individuals
- Locations
  - Geographic places with no associated governments
- Facilities
  - Objects from the domain of civil engineering
- Geopolitical Entities
  - Geographic places with associated governments

[Appelt, 2003]

Why GPEs

- An ontological problem: certain entities have attributes of physical objects in some contexts, organizations in some contexts, and collections of people in others
- Sometimes it is difficult to impossible to determine which aspect is intended
- It appears that in some contexts, the same phrase plays different roles in different clauses

Aspects of GPEs

- Physical
  - San Francisco has a mild climate
- Organization
  - The United States is seeking a solution to the North Korean problem.
- Population
  - France makes a lot of good wine.

Types of Linguistic Mentions

- Name mentions
  - The mention uses a proper name to refer to the entity
- Nominal mentions
  - The mention is a noun phrase whose head is a common noun
- Pronominal mentions
  - The mention is a headless noun phrase, or a noun phrase whose head is a pronoun, or a possessive pronoun

Entity and Mention Example

[COLOGNE, [Germany]] (AP) _ [A [Chilean] exile] has filed a complaint against [former [Chilean] dictator Gen. Augusto Pinochet] accusing [him] of responsibility for [her] arrest and torture in [Chile] in 1973, [prosecutors] said Tuesday.

[[The woman, [a Chilean] who has since gained [German] citizenship],] accused [Pinochet] of depriving [her] of personal liberty and causing bodily harm during [her] arrest and torture.

Person
Organization
Geopolitical Entity

Explicit and Implicit Relations

- Many relations are true in the world. Reasonable knoweldge bases used by extraction systems will include many of these relations. Semantic analysis requires focusing on certain ones that are directly motivated by the text.

- Example:
  - Baltimore is in Maryland is in United States.
  - “Baltimore, MD”
  - Text mentions Baltimore and United States. Is there a relation between Baltimore and United States?

Another Example

- *Prime Minister Tony Blair attempted to convince the British Parliament of the necessity of intervening in Iraq.*

- Is there a role relation specifying Tony Blair as prime minister of Britain?

- A test: a relation is implicit in the text if the text provides convincing evidence that the relation actually holds.

Explicit Relations

- Explicit relations are expressed by certain surface linguistic forms
  - Copular predication - Clinton was the president.
  - Prepositional Phrase - The CEO of Microsoft…
  - Prenominal modification - The American envoy…
  - Possessive - Microsoft's chief scientist…
  - SVO relations - Clinton arrived in Tel Aviv…
  - Nominalizations - Anan's visit to Baghdad…
  - Apposition - Tony Blair, Britain's prime minister…

Types of ACE Relations

- ROLE - relates a person to an organization or a geopolitical entity
  - Subtypes: member, owner, affiliate, client, citizen
- PART - generalized containment
  - Subtypes: subsidiary, physical part-of, set membership
- AT - permanent and transient locations
  - Subtypes: located, based-in, residence
- SOC - social relations among persons
  - Subtypes: parent, sibling, spouse, grandparent, associate

Event Types (preliminary)

- Movement
  - Travel, visit, move, arrive, depart …
- Transfer
  - Give, take, steal, buy, sell…
- Creation/Discovery
  - Birth, make, discover, learn, invent…
- Destruction
  - die, destroy, wound, kill, damage…

Machine Learning for Relation Extraction

Motivations of ML

- Porting to new domains or applications is expensive
- Current technology requires IE experts
  - Expertise difficult to find on the market
  - SME cannot afford IE experts
- Machine learning approaches
  - Domain portability is relatively straightforward
  - System expertise is not required for customization
  - “Data driven” rule acquisition ensures full coverage of examples

Problems

- Training data may not exist, and may be very expensive to acquire
- Large volume of training data may be required
- Changes to specifications may require reannotation of large quantities of training data
- Understanding and control of a domain adaptive system is not always easy for non-experts

Parameters

- Document structure
  - Free text
  - Semi-structured
  - Structured
- Richness of the annotation
  - Shallow NLP
  - Deep NLP
- Complexity of the template filling rules
  - Single slot
  - Multi slot
- Amount of data
- Degree of automation
  - Semi-automatic
  - Supervised
  - Semi-Supervised
  - Unsupervised
- Human interaction/contribution
- Evaluation/validation
  - during learning loop
  - Performance: recall and precision

Learning Methods for Template Filling Rules

- Inductive learning
- Statistical methods
- Bootstrapping techniques
- Active learning

Documents

- Unstructured (Free) Text
  - Regular sentences and paragraphs
  - Linguistic techniques, e.g., NLP
- Structured Text
  - Itemized information
  - Uniform syntactic clues, e.g., table understanding
- Semi-structured Text
  - Ungrammatical, telegraphic (e.g., missing attributes, multi-value attributes, …)
  - Specialized programs, e.g., wrappers

“Information Extraction” From Free Text

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

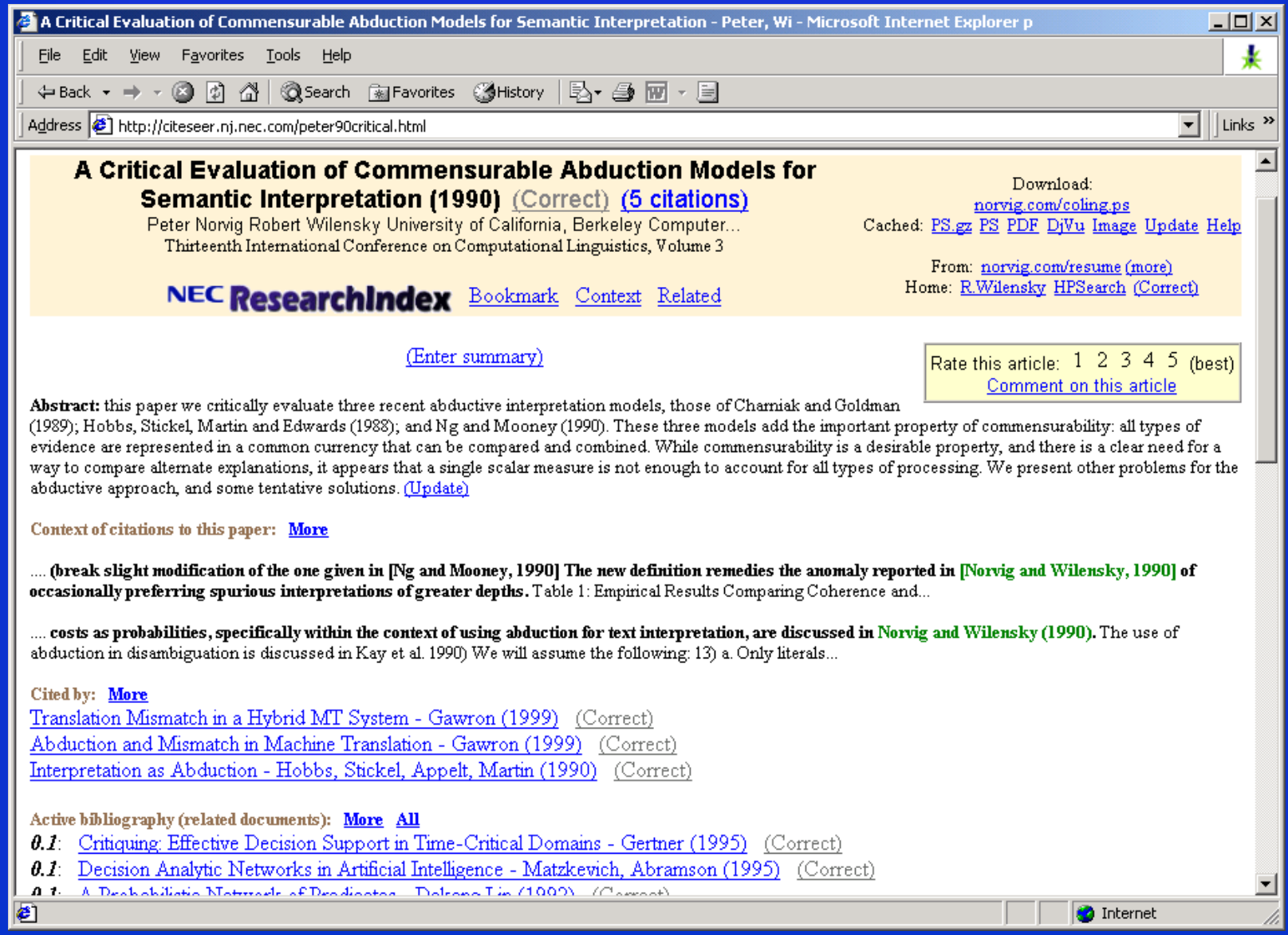
* Microsoft
Gates
Microsoft

* Bill Veghte
Microsoft
VP

* Richard Stallman
founder
Free Software Foundation

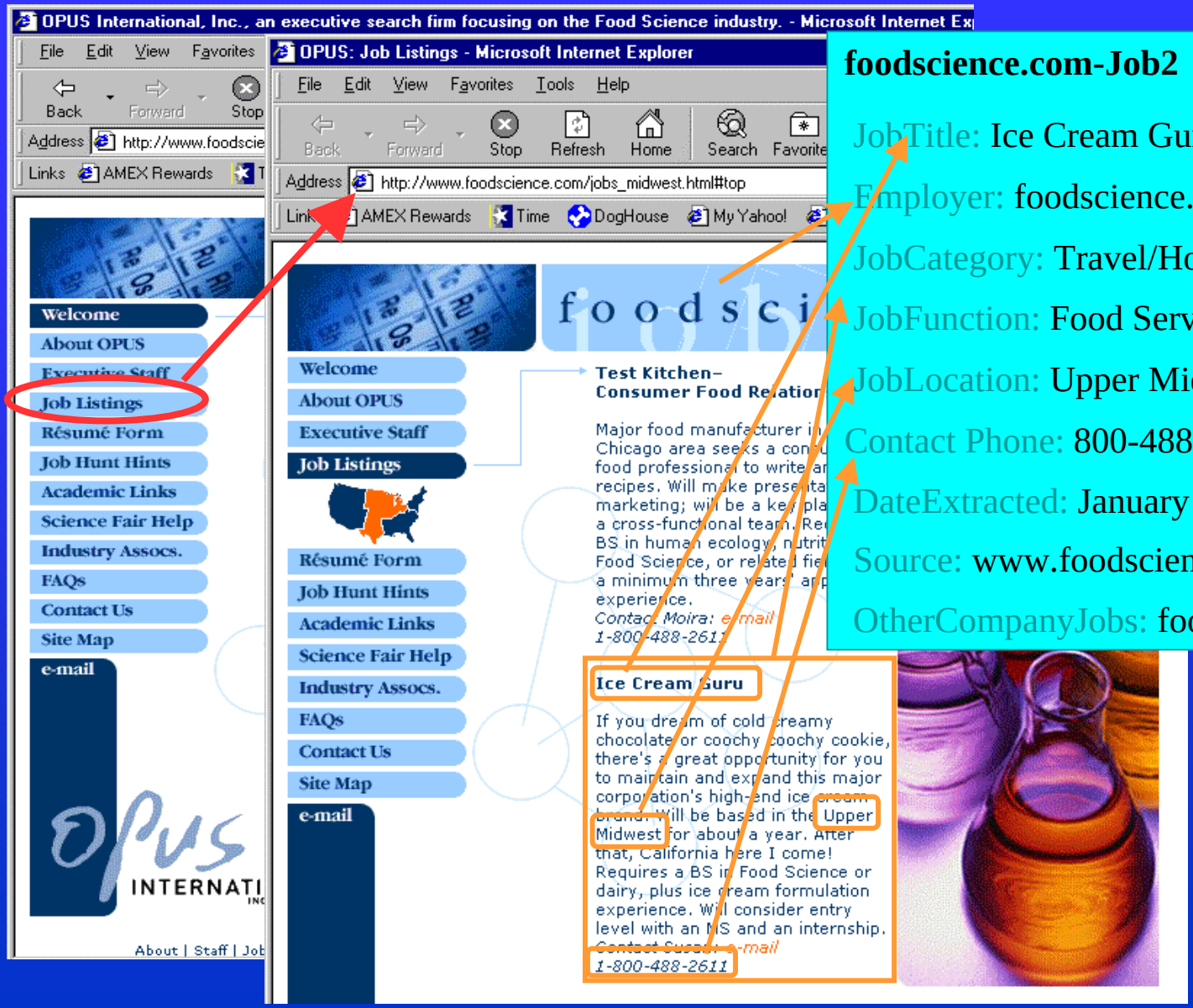
NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
BillVeghte	VP	Microsoft
RichardStallman	founder	Free Soft..

IE from Research Papers



Extracting Job Openings from the Web:

Semi-Structured Data



foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

Contact Phone: 800-488-2611

DateExtracted: January 8, 2001

Source: www.foodscience.com/jobs_midwest.html

OtherCompanyJobs: foodscience.com-Job1

Outline

- Free text
  - Supervised and semi-automatic
    - AutoSlog
  - Semi-Supervised
    - AutoSlog-TS
  - Unsupervised
    - ExDisco
- Semi-structured and unstructured text
  - NLP-based wrapping techniques
    - RAPIER

Free Text

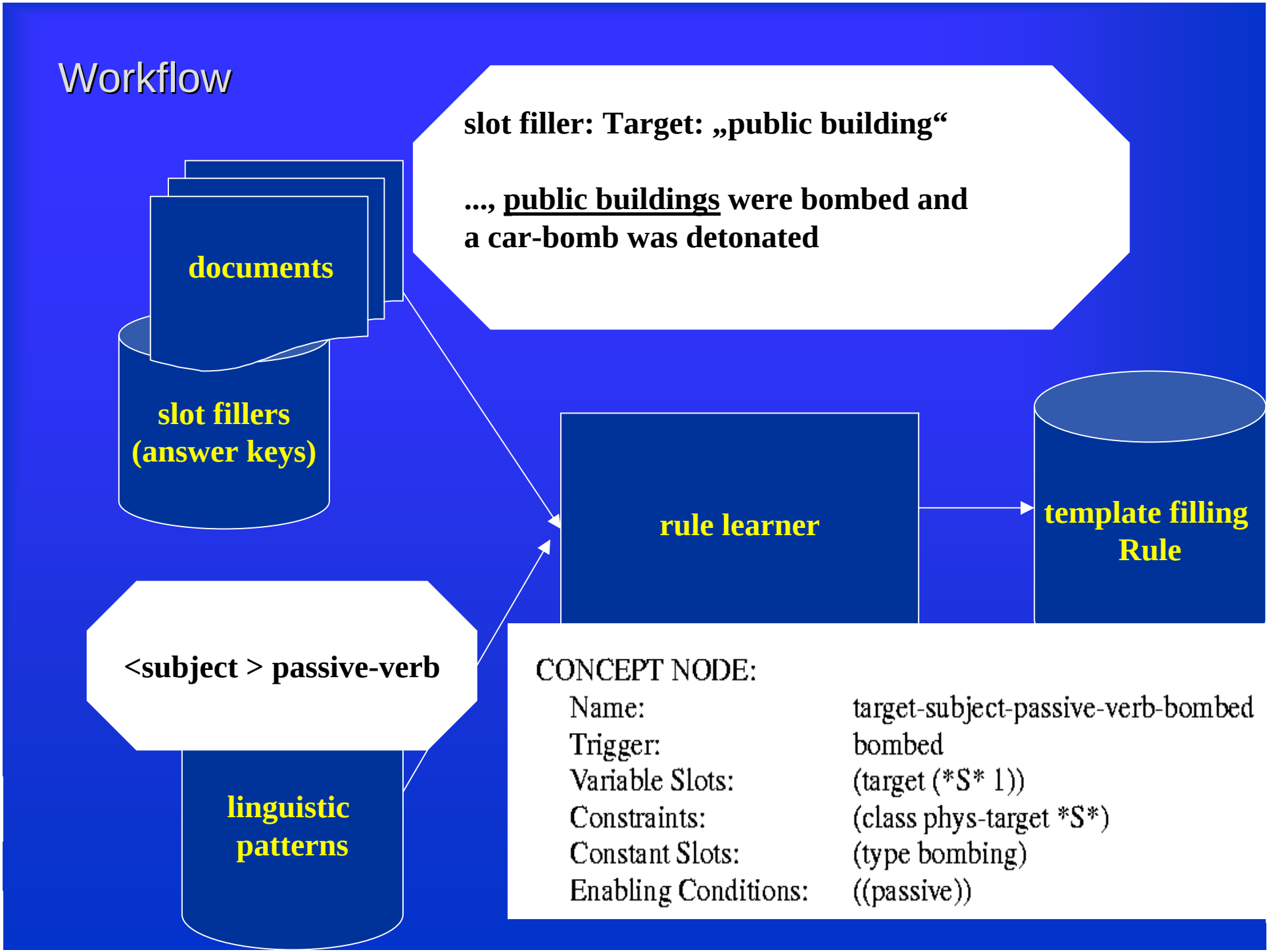
NLP-based Supervised Approaches

- Input is an annotated corpus
  - Documents with associated templates
- A parser
  - Chunk parser
  - Full sentence parser
- Learning the mapping rules
  - From linguistic constructions to template fillers

AutoSlog (1993)

- Extracting a concept dictionary for template filling
- Full sentence parser
- One slot filler rules
- Domain adaptation performance
  - Before AutoSlog: hand-crafted dictionary
    - two highly skilled graduate students
    - 1500 person-hours
  - AutoSlog:
    - A dictionary for the terrorist domain: 5 person hours
    - 98% performance achievement of the hand-crafted dictionary

Workflow


slot filler: Target: „public building“
..., public buildings were bombed and
a car-bomb was detonated
documents
slot fillers
(answer keys)
rule learner
template filling
Rule
<subject > passive-verb
linguistic
patterns
CONCEPT NODE:
Name: target-subject-passive-verb-bombed
Trigger: bombed
Variable Slots: (target (*S* 1))
Constraints: (class phys-target *S*)
Constant Slots: (type bombing)
Enabling Conditions: ((passive))

Linguistic Patterns

Linguistic Pattern	Example
<subject> passive-verb	<victim> was murdered
<subject> active-verb	<perpetrator> bombed
<subject> verb infinitive	<perpetrator> attempted to kill
<subject> auxiliary noun	<victim> was victim
passive-verb <dobj>[1]	killed <victim>
active-verb <dobj>	bombed <target>
infinitive <dobj>	to kill <victim>
verb infinitive <dobj>	threatened to attack <target>
gerund <dobj>	killing <victim>
noun auxiliary <dobj>	fatality was <victim>
noun prep <np>	bomb against <target>
active-verb prep <np>	killed with <instrument>

Id: DEV-MUC4-1192 **Slot filler:** "gilberto molasco"
Sentence: (they took 2-year-old gilberto molasco, son of patricio rodriguez, and 17-year-old andres argueta, son of emimesto argueta.)

CONCEPT NODE

Name:	victim-active-verb-dobj-took
Trigger:	took
Variable Slots:	(victim (*DOBJ* 1))
Constraints:	(class victim *DOBJ*)
Constant Slots:	(type kidnapping)
Enabling Conditions:	((active))

A bad concept node definition

Error Sources

- A sentence contains the answer key string but does not contain the event
- The sentence parser delivers wrong results
- A heuristic proposes a wrong conceptual anchor

Training Data

- MUC-4 corpus
- 1500 texts
- 1258 answer keys
- 4780 string fillers
- 1237 concept node definition

- Human in loop for validation to filter out bad and wrong definitions: 5 hours

- 450 concept nodes left after human review

System/Test Set	Recall	Precision	F-measure
MUC-4/TST3	46	56	50.51
AutoSlog/TST3	43	56	48.65
MUC-4/TST4	44	40	41.90
AutoSlog/TST4	39	45	41.79

Comparative Results

Summary

- Advantages
  - Semi-automatic
  - Less human effort
- Disadvantages
  - Human interaction
  - Still very naive approach
  - Need a big amount of annotation
  - Domain adaptation bottelneck is shifted to human annotation
  - No generation of rules
  - One slot filling rule
  - No mechanism for filtering out bad rules

NLP-based ML Approaches

- LIEP (Huffman, 1995)
- PALKA (Kim & Moldovan, 1995)
- HASTEN (Krupka, 1995)
- CRYSTAL (Soderland et al., 1995)

LIEP [1995]

The Parliament building was bombed by ***Carlos***.

TARGET-was-bombed-by-PERPETRATOR:
  noun-group(TRGT, head(isa(physical-target))),
  noun-group(PERP, head(isa(perpetrator)))
  verb-group(VG, type(passive), head(bombed))
  preposition(PREP, head(by))

  subject(TRGT, VG),
  post-verbal-prep(VG, PREP),
  prep-object(PREP, PERP)
  $\Longrightarrow$ bombing-event(BE, target(TRGT), agent(PERP))

PALKA [1995]

The Parliament building was bombed by ***Carlos***.

FP-structure = MeaningFrame + PhrasalPattern

Meaning Frame: (BOMBING agent: ANIMATE
target: PHYS-OBJ
instrument: PHYS-OBJ
effect: STATE)

Phrasal Pattern: ((PHYS-OBJ) was bombed by (PERP))

FP-structure:
(BOMBING target: PHYS-OBJ
agent: PERP
pattern: ((target) was bombed by (agent))

HASTEN [1995]

The Parliament building was bombed by ***Carlos***.

BOMBING:	
TARGET:	NP "semantic = physical-object"
ANCHOR:	VG "root = bomb"
PERPETRATOR:	NP "semantic = terrorist-group"

- Egraphs
- (*SemanticLabel, StructuralElement*)

CRYSTAL [1995]

The Parliament building was bombed by ***Carlos***.

Concept type: BUILDING BOMBING

SUBJECT:	Classes include:	<PhysicalTarget>
	Terms include:	BUILDING
	Extract:	*target*
VERB:	Root:	BOMB
	Mode:	passive
PREPOS-PHRASE:	Preposition:	BY
	Classes include:	<PersonName>
	Extract:	*perpetrator name*

A Few Remarks

- Single slot vs. multi.-solt rules
- Semantic constraints
- Exact phrase match

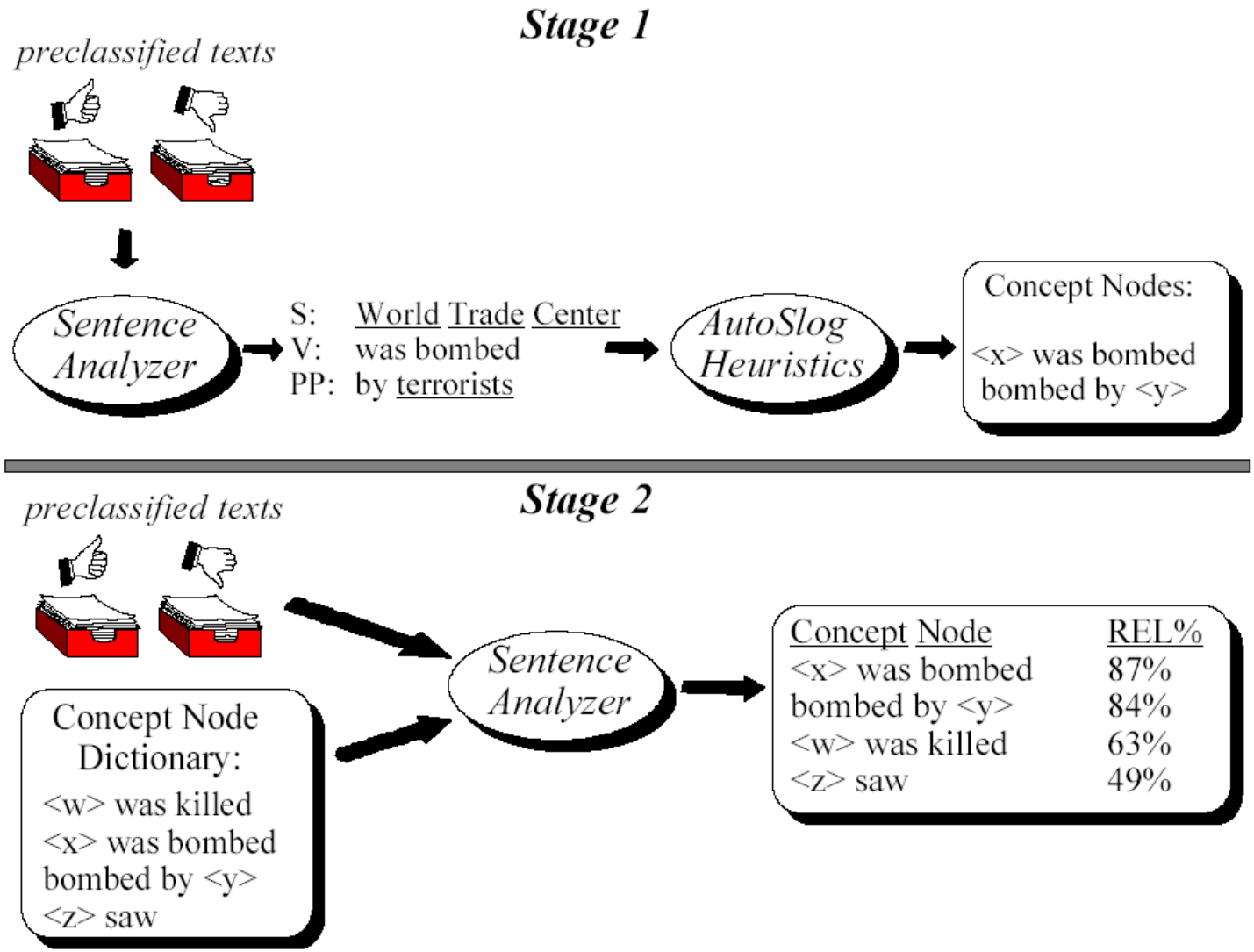
Semi-Supervised Approaches

AutoSlog TS [Riloff, 1996]

- Input: pre-classified documents (relevant vs. irrelevant)
- NLP as preprocessing: full parser for detecting subject-v-object relationships
- Principle
  - Relevant patterns are patterns occuring more often in the relevant documents
- Output: ranked patterns, but not classified, namely, only the left hand side of a template filling rule
- The dictionary construction process consists of two stages:
  - pattern generation and
  - statistical filtering
- Manual review of the results

Linguistic Patterns

Linguistic Pattern	Example
<subject> passive-verb	<victim> was murdered
<subject> active-verb	<perpetrator> bombed
<subject> verb infinitive	<perpetrator> attempted to kill
<subject> auxiliary noun	<victim> was victim
passive-verb <dobj>[1]	killed <victim>
active-verb <dobj>	bombed <target>
infinitive <dobj>	to kill <victim>
verb infinitive <dobj>	threatened to attack <target>
gerund <dobj>	killing <victim>
noun auxiliary <dobj>	fatality was <victim>
noun prep <np>	bomb against <target>
active-verb prep <np>	killed with <instrument>



AutoSlog-TS flowchart

Pattern Extraction

The sentence analyzer produces a syntactic analysis for each sentence and identified noun phrases. For each noun phrase, the heuristic rules generate a pattern to extract noun phrase.

<subject> bombed

Relevance Filtering

- the whole text corpus will be processed a second time using the extracted patterns obtained by stage 1.
- Then each pattern will be assigned with a relevance rate based on its occurring frequency in the relevant documents relatively to its occurrence in the total corpus.
- A preferred pattern is the one which occurs more often in the relevant documents.

Statistical Filtering

Relevance Rate:

$$Pr(\text{relevant text} \setminus \text{text contains case frame}_i) = \frac{\text{rel-freq}_i}{\text{total-freq}_i}$$

rel-freq_i : number of instances of case-frame_i *in the relevant documents*

total-freq_i : total number of instances of case-frame_i

Ranking Function:

$$\text{score}_i = \text{relevance rate}_i * \log_2(\text{frequency}_i)$$

Pr < 0,5 negatively correlated with the domain

„Top“

1. <subj> exploded	14. <subj> occurred
2. murder of <np>	15. <subj> was located
3. assassination of <np>	16. took_place on <np>
4. <subj> was killed	17. responsibility for <np>
5. <subj> was kidnapped	18. occurred on <np>
6. attack on <np>	19. was wounded in <np>
7. <subj> was injured	20. destroyed <dobj>
8. exploded in <np>	21. <subj> was murdered
9. death of <np>	22. one of <np>
10. <subj> took_place	23. <subj> kidnapped
11. caused <dobj>	24. exploded on <np>
12. claimed <dobj>	25. <subj> died
13. <subj> was wounded	

The Top 25 Extraction Patterns

Empirical Results

- 1500 MUC-4 texts
  - 50% are relevant.
- In stage 1, 32,345 unique extraction patterns.
- A user reviewed the top 1970 patterns in about 85 minutes and kept the best 210 patterns.
- Evaluation
  - AutoSlog and AutoSlog-TS systems return comparable performance.

Conclusion

- Advantages
  - Pioneer approach to automatic learning of extraction patterns
  - Reduce the manual annotation
- Disadvantages
  - Ranking function is too dependent on the occurrence of a pattern, relevant patterns with low frequency can not float to the top
  - Only patterns, not classification

Unsupervised

ExDisco (Yangarber 2001)

- Seed
- Bootstrapping
- Duality/Density Principle for validation of each iteration

Input

- a corpus of unclassified and unannotated documents
- a seed of patterns, e.g.,

subject(company)-verb(appoint)-object(person)

NLP as Preprocessing

- full parser for detecting subject-v-object relationships
  - NE recognition
  - Functional Dependency Grammar (FDG) formalism (Tapannaien & Järvinen, 1997)

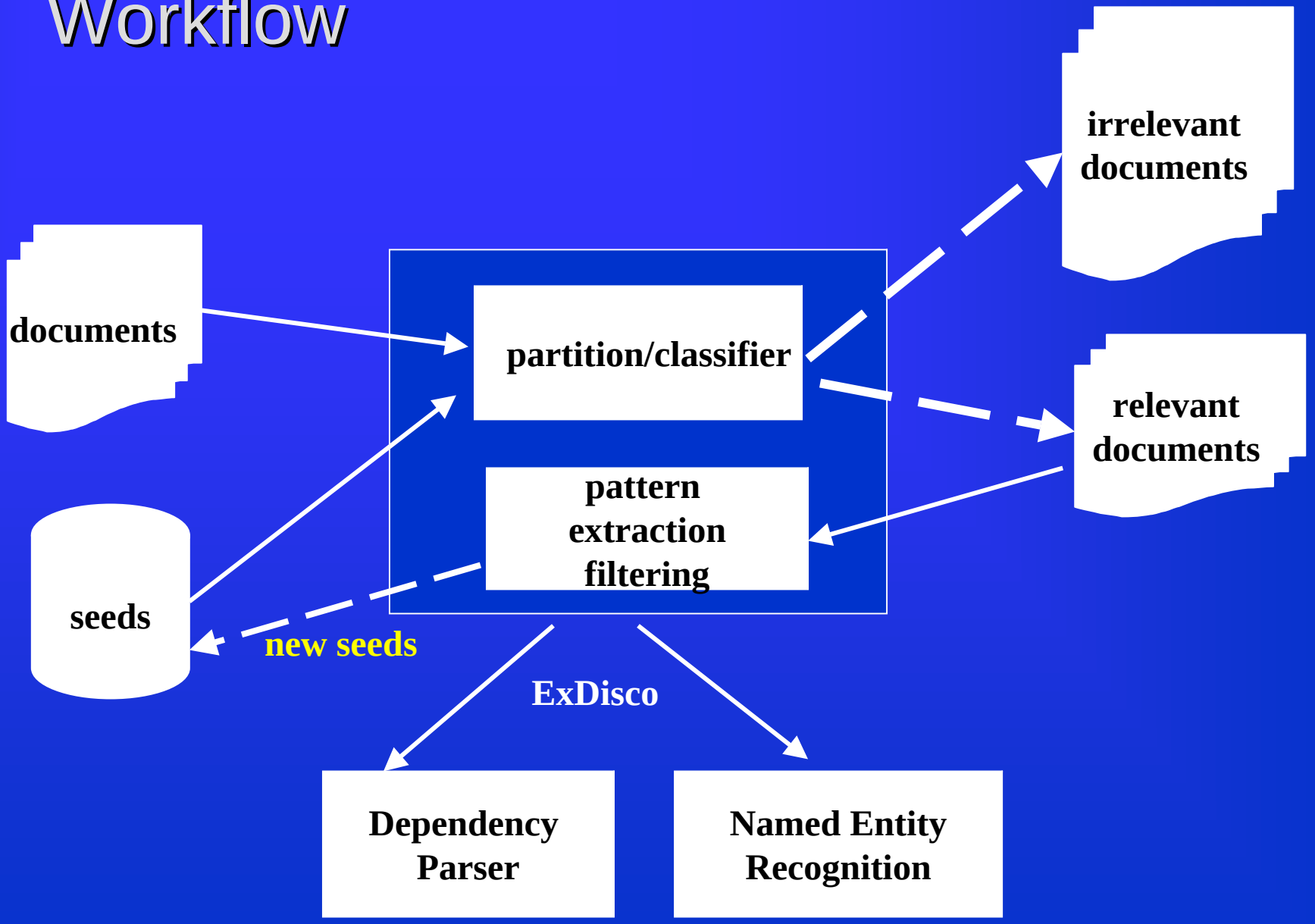
Duality/Density Principle (boostrapping)

- Density:
  - Relevant documents contain more relevant patterns
- Duality:
  - documents that are relevant to the scenario are strong indicators of good patterns
  - good patterns are indicators of relevant documents

Algorithm

- Given:
  - a large corpus of un-annotated and un-classified documents
  - a trusted set of scenario patterns, initially chosen ad hoc by the user, the seed. Normally is the seed relatively small, two or three
  - (possibly empty) set of concept classes
- Partition
  - applying seed to the documents and divide them into relevant and irrelevant documents
- Search for new candidate patterns:
  - automatic convert each sentence into a set of candidate patterns.
  - choose those patterns which are strongly distributed in the relevant documents
  - Find new concepts
- User feedback
- Repeat

Workflow


irrelevant documents
documents
partition/classifier
relevant documents
pattern extraction filtering
seeds
new seeds
ExDisco
Dependency Parser
Named Entity Recognition

Pattern Ranking

$$Score(P)=\frac{|H\cap R|}{|H|}\cdot \mathbf{LOG}\,(|H\cap R|)$$

Evaluation of Event Extraction

Pattern Base	*Recall*	*Precision*	*F*
Seed	27	74	39.58
ExDisco	52	72	60.16
Union	57	73	63.56
Manual-MUC	47	70	56.40
Manual-NOW	56	75	64.04

ExDisco

- Advantages
  - Unsupervised
  - Multi-slot template filler rules
- Disadvantages
  - Only subject-verb-object patterns, local patterns are ignored
  - No generalization of pattern rules (see inductive learning)
  - Collocations are not taken into account, e.g., *PN take responsibility of Company*
- Evaluation methods
  - Event extraction: integration of patterns into IE system and test recall and precision
  - Qualitative observation: manual evaluation
  - Document filtering: using ExDisco as document classifier and document retrieval system

Relational learning and Inductive Logic Programming (ILP)

- Allow induction over structured examples that can include first-order logical representations and unbounded data structures

Semi-Structured and Un-Structured Documents

RAPIER [Califf, 1998]

- Inductive Logic Programming
- Extraction Rules
  - Syntactic information
  - Semantic information
- Advantage
  - Efficient learning (bottom-up)
- Drawback
  - Single-slot extraction

RAPIER [Califf, 1998]

- Uses relational learning to construct unbounded pattern-match rules, given a database of texts and filled templates
- Primarily consists of a bottom-up search
- Employs limited syntactic and semantic information
- Learn rules for the complete IE task

Filled template of RAPIER

Posting from Newsgroup

```
Telecommunications. SOLARIS Systems
Administrator. 38-44K. Immediate need

Leading telecommunications firm in need
of an energetic individual to fill the
following position in the Atlanta
office:

   SOLARIS SYSTEMS ADMINISTRATOR
   Salary: 38-44K with full benefits
   Location: Atlanta Georgia, no
         relocation assistance provided
```

Filled Template

```
computer_science_job
title: SOLARIS Systems Administrator
salary: 38-44K
state: Georgia
city: Atlanta
platform: SOLARIS
area: telecommunications
```

Figure 1: Sample Message and Filled Template

RAPIER's rule representation

- Indexed by template name and slot name
- Consists of three parts:
  1. A pre-filler pattern
  2. Filler pattern (matches the actual slot)
  3. Post-filler

Pattern

- Pattern item: matches exactly one word
- Pattern list: has a maximum length N and matches 0..N words.
- Must satisfy a set of constraints
  1. Specific word, POS, Semantic class
  2. Disjunctive lists

RAPIER Rule

ORIGINAL DOCUMENT:

Al. C Programmer. 38-44K.
Leading AI firm in need of
an energetic individual to
fill the following position:

EXTRACTED DATA:

computer-science-job
title: C Programmer
salary: 38-44K
area: AI

AREA extraction pattern:

Pre-filler pattern: word: leading
Filler pattern: list: len: 2
tags: [nn, nns]
Post-filler pattern: word: [firm, company]

RAPIER'S Learning Algorithm

- Begins with a most specific definition and compresses it by replacing with more general ones
- Attempts to compress the rules for each slot
- Preferring more specific rules

Implementation

- Least general generalization (LGG)
- Starts with rules containing only generalizations of the filler patterns
- Employs top-down beam search for pre and post fillers
- Rules are ordered using an information gain metric and weighted by the size of the rule (preferring smaller rules)

Example

Located in Atlanta, Georgia.
Offices in Kansas City, Missouri

Pre-filler:	Filler:	Post-filler:
1) word: located tag: vbn 2) word: in tag: in	1) word: atlanta tag: nnp	1) word: , tag: , 2) word: georgia tag: nnp 3) word: . tag: .

and

Pre-filler:	Filler:	Post-filler:
1) word: offices tag: nns 2) word: in tag: in	1) word: kansas tag: nnp 2) word: city tag: nnp	1) word: , tag: , 2) word: missouri tag: nnp 3) word: . tag: .

Example (cont)

```
Pre-filler:     Filler:                          Post-filler:
                1) list: max length: 2
                    word: {atlanta, kansas, city}
                    tag: nnp
and
Pre-filler:     Filler:                          Post-filler:
                1) list: max length: 2
                    tag: nnp
```

```
Pre-filler:     Filler:                          Post-filler:
1) word: in     1) list: max length: 2           1) word: ,
    tag: in         word: {atlanta,                  tag: ,
                    kansas, city}
                    tag: nnp
and
Pre-filler:     Filler:                          Post-filler:
1) word: in     1) list: max length: 2           1) word: ,
    tag: in         tag: nnp                         tag: ,
```

Example (cont)

Final best rule:

Pre-filler:	Filler:	Post-filler:
1) word: in	1) list: max length: 2	1) word: ,
tag: in	tag: nnp	tag: ,
		2) tag: nnp
		semantic: state

Experimental Evaluation

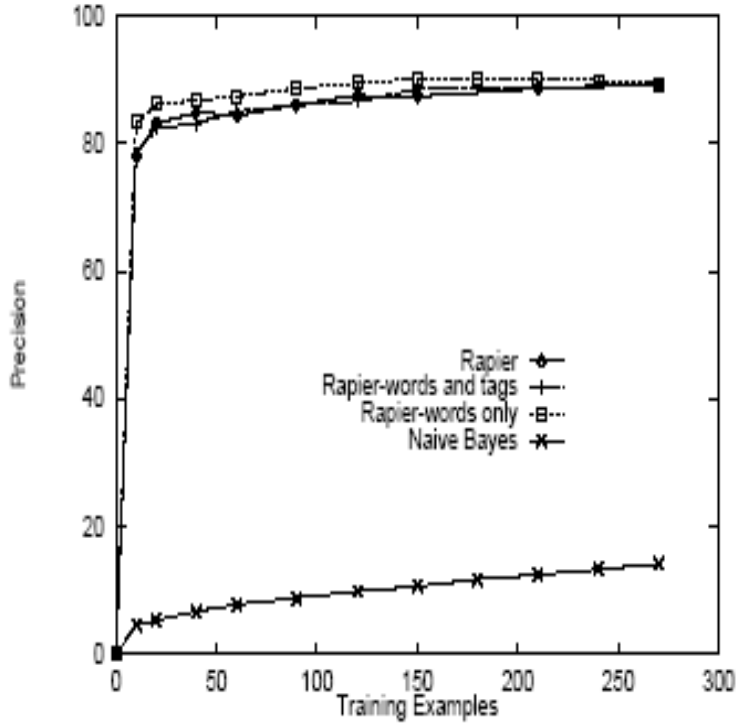
- A set of 300 computer-related job posting from austin.jobs
- A set of 485 seminar announcements from CMU.
- Three different versions of RAPIER were tested

  1.words, POS tags, semantic classes

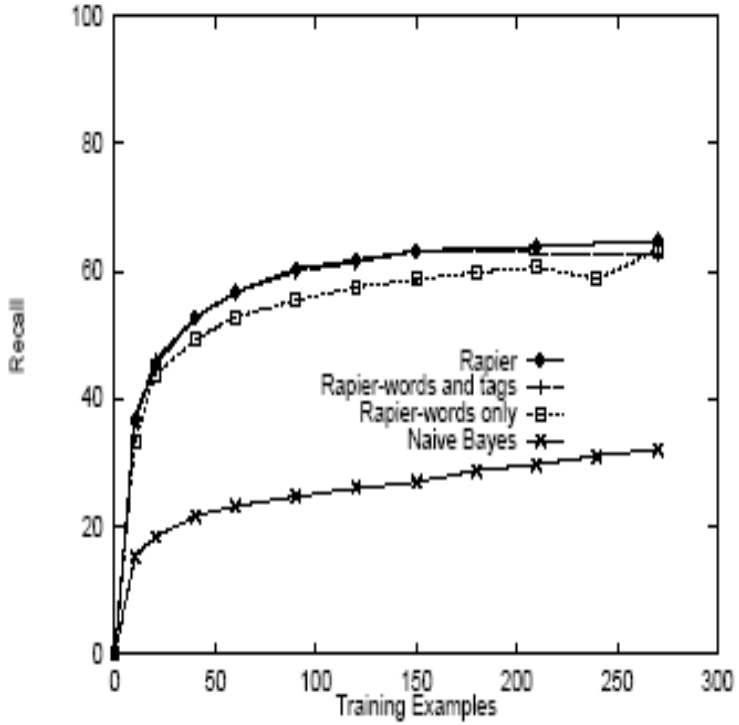
  2. words, POS tags

  3. words

Performance on job postings



Precision



Recall

Results for seminar announcement task

System	stime		etime		loc		speaker	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
RAPIER	93.9	92.9	95.8	94.6	91.0	60.5	80.9	39.4
RAP-WT	96.5	95.3	94.9	94.4	91.0	61.5	79.0	40.0
RAP-W	96.5	95.9	96.8	96.6	90.0	54.8	76.9	29.1
NAIBAY	98.2	98.2	49.5	95.7	57.3	58.8	34.5	25.6
SRV	98.6	98.4	67.3	92.6	74.5	70.1	54.4	58.4
WHISK	86.2	100.0	85.0	87.2	83.6	55.4	52.6	11.1
WH-PR	96.2	100.0	89.5	87.2	93.8	36.1	0.0	0.0

Conclusion

- Pros
  - Have the potential to help automate the development process of IE systems.
  - Work well in locating specific data in newsgroup messages
  - Identify potential slot fillers and their surrounding context with limited syntactic and semantic information
  - Learn rules from relatively small sets of examples in some specific domain
- Cons
  - single slot
  - regular expression
  - Unknown performances for more complicated situations

References

1. N. Kushmerick. Wrapper induction: Efficiency and Expressiveness, Artificial Intelligence, 2000.
2. I. Muslea. Extraction Patterns for Information Extraction. AAAI-99 Workshop on Machine Learning for Information Extraction.
3. Riloff, E. and R. Jones. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99) , 1999, pp. 474-479.
4. R. Yangarber, R. Grishman, P. Tapanainen and S. Huttunen. Automatic Acquisition of Domain Knowledge for Information Extraction. In Proceedings of the 18th International Conference on Computational Linguistics: COLING-2000, Saarbrücken.

http://www.dfki.de/~neumann/ie-esslli04.html