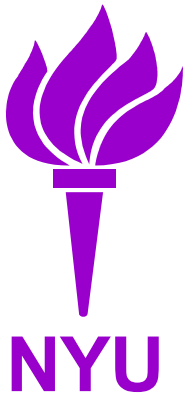


Challenges for Information Extraction

Ralph Grishman

Madrid: November 2006

Information Extraction

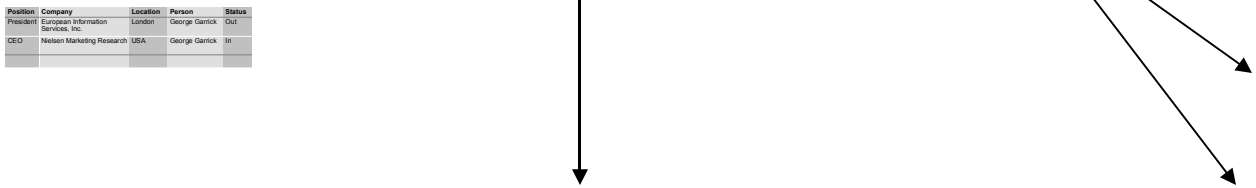
Information Extraction (IE) =

identifying the instances of the
important entities, relations, and events
for a *domain*
from *unstructured* text.

Extraction Example

Topic: executive succession

– George Garrick, 40 years old, **president of the London-based European Information Services Inc., was appointed chief executive officer of** Nielsen Marketing Research, USA.



Position	Company	Location	Person	Status
President	European Information Services, Inc.	London	George Garrick	Out
CEO	Nielsen Marketing Research	USA	George Garrick	In

Information Extraction vs. Information Retrieval

- Based on/returns
  - normalized values (e.g., dates)
  - entities
  - relations
- Must be adapted to specific task / domain

- Based on terms
- Returns documents or passages
- General technology

Power of Information Extraction

Queries involving relations or ranges of values are hard to answer using IR:

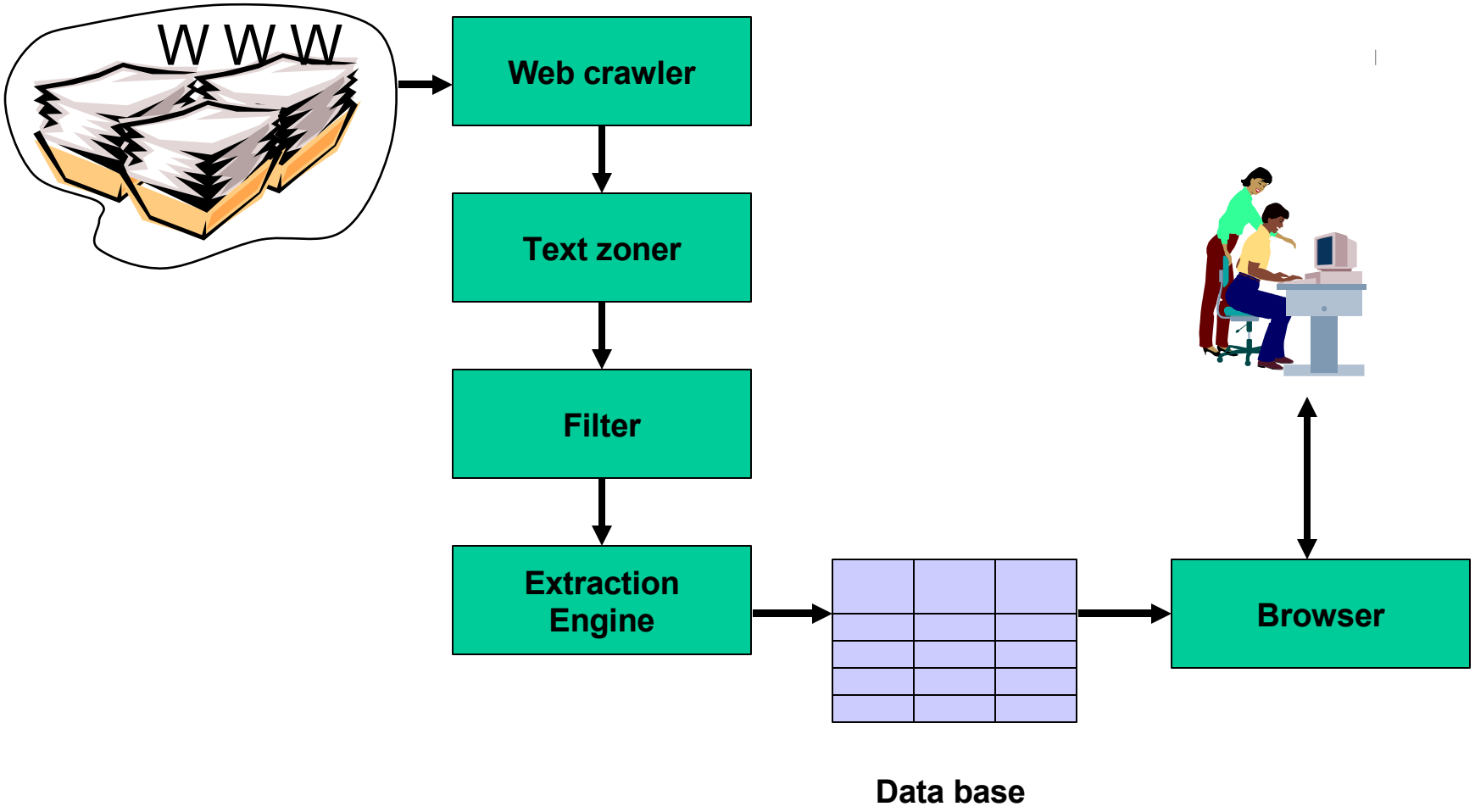
- Where has Condoleezza Rice been in the last month?
- What terrorist attacks occurred in Europe in 2004?

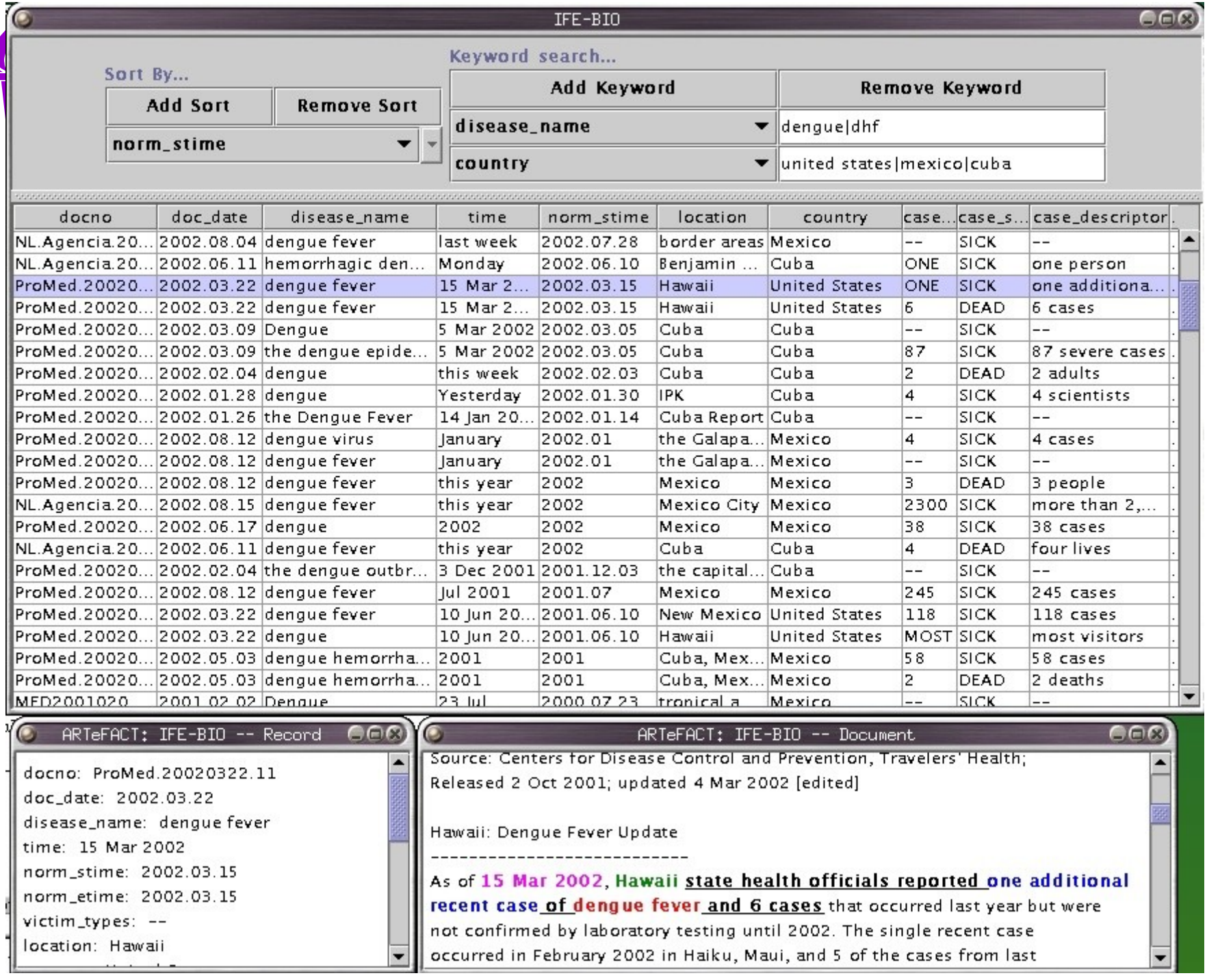
IE for Document Access

IE can provide a powerful search tool for documents in a specific domain

Approach:

- Use information extraction to create a database of events of interest
  - in our example, infectious disease outbreaks
- Update the data base on a daily basis by using a web crawler to retrieve the latest news
- Provide access to articles / web pages through the data base
  - provides targeted, spreadsheet-like search capabilities


W W W
Web crawler
Text zoner
Filter
Extraction Engine
Browser
Data base


IFE-BIO
Sort By...
Add Sort
Remove Sort
norm_stime
Keyword search...
Add Keyword
Remove Keyword
disease_name
dengue|dhf
country
united states|mexico|cuba
docno | doc_date | disease_name | time | norm_stime | location | country | case... | case_s... | case_descriptor
NL.Agencia.20... | 2002.08.04 | dengue fever | last week | 2002.07.28 | border areas | Mexico | -- | SICK | --
NL.Agencia.20... | 2002.06.11 | hemorrhagic den... | Monday | 2002.06.10 | Benjamin ... | Cuba | ONE | SICK | one person
ProMed.20020... | 2002.03.22 | dengue fever | 15 Mar 2... | 2002.03.15 | Hawaii | United States | ONE | SICK | one additiona...
ProMed.20020... | 2002.03.22 | dengue fever | 15 Mar 2... | 2002.03.15 | Hawaii | United States | 6 | DEAD | 6 cases
ProMed.20020... | 2002.03.09 | Dengue | 5 Mar 2002 | 2002.03.05 | Cuba | Cuba | -- | SICK | --
ProMed.20020... | 2002.03.09 | the dengue epide... | 5 Mar 2002 | 2002.03.05 | Cuba | Cuba | 87 | SICK | 87 severe cases
ProMed.20020... | 2002.02.04 | dengue | this week | 2002.02.03 | Cuba | Cuba | 2 | DEAD | 2 adults
ProMed.20020... | 2002.01.28 | dengue | Yesterday | 2002.01.30 | IPK | Cuba | 4 | SICK | 4 scientists
ProMed.20020... | 2002.01.26 | the Dengue Fever | 14 Jan 20... | 2002.01.14 | Cuba Report | Cuba | -- | SICK | --
ProMed.20020... | 2002.08.12 | dengue virus | January | 2002.01 | the Galapa... | Mexico | 4 | SICK | 4 cases
ProMed.20020... | 2002.08.12 | dengue fever | January | 2002.01 | the Galapa... | Mexico | -- | SICK | --
ProMed.20020... | 2002.08.12 | dengue fever | this year | 2002 | Mexico | Mexico | 3 | DEAD | 3 people
NL.Agencia.20... | 2002.08.15 | dengue fever | this year | 2002 | Mexico City | Mexico | 2300 | SICK | more than 2,...
ProMed.20020... | 2002.06.17 | dengue | 2002 | 2002 | Mexico | Mexico | 38 | SICK | 38 cases
NL.Agencia.20... | 2002.06.11 | dengue fever | this year | 2002 | Cuba | Cuba | 4 | DEAD | four lives
ProMed.20020... | 2002.02.04 | the dengue outbr... | 3 Dec 2001 | 2001.12.03 | the capital... | Cuba | -- | SICK | --
ProMed.20020... | 2002.08.12 | dengue fever | Jul 2001 | 2001.07 | Mexico | Mexico | 245 | SICK | 245 cases
ProMed.20020... | 2002.03.22 | dengue fever | 10 Jun 20... | 2001.06.10 | New Mexico | United States | 118 | SICK | 118 cases
ProMed.20020... | 2002.03.22 | dengue | 10 Jun 20... | 2001.06.10 | Hawaii | United States | MOST | SICK | most visitors
ProMed.20020... | 2002.05.03 | dengue hemorrha... | 2001 | 2001 | Cuba, Mex... | Mexico | 58 | SICK | 58 cases
ProMed.20020... | 2002.05.03 | dengue hemorrha... | 2001 | 2001 | Cuba, Mex... | Mexico | 2 | DEAD | 2 deaths
MED2001020 | 2001.02.02 | Dengue | 23 Jul | 2000.07.23 | tropical a | Mexico | -- | SICK | --
ARTeFACT: IFE-BIO -- Record
docno: ProMed.20020322.11
doc_date: 2002.03.22
disease_name: dengue fever
time: 15 Mar 2002
norm_stime: 2002.03.15
norm_etime: 2002.03.15
victim_types: --
location: Hawaii
ARTeFACT: IFE-BIO -- Document
Source: Centers for Disease Control and Prevention, Travelers' Health;
Released 2 Oct 2001; updated 4 Mar 2002 [edited]
Hawaii: Dengue Fever Update
As of 15 Mar 2002, Hawaii state health officials reported one additional recent case of dengue fever and 6 cases that occurred last year but were not confirmed by laboratory testing until 2002. The single recent case occurred in February 2002 in Haiku, Maui, and 5 of the cases from last

Application Areas

- Genomics
  - extracting information about gene/protein interactions
- Medical
  - generating statistical summaries of medical reports
- Financial
  - extracting information from financial news and company reports

Not a New Idea

- The idea of building data bases through linguistic analysis and using these data bases for document retrieval was discussed by Zellig Harris in 1958
("Linguistic Analysis for Information Retrieval",
Int'l Conf. on Scientific Information, Washington, D.C.)

[see http://www.garfield.library.upenn.edu/papers/pittsburgh92001.pdf]

- It has taken several decades for the linguistic technology to catch up to these ideas.
  - Robust, often corpus-trained analysis methods for text analysis: name recognition, chunking, parsing, coreference, …
  - Methods for *discovering* relationships from text

Challenges of IE

To understand recent progress in IE, we need to

- appreciate the basic approach and the problems which arise because of the complexities of language
- see how these difficulties are being addressed by current research
  - *including our own research at NYU* 

Basic Approach

- Build linguistic patterns:

  person was appointed as *post* of *company*

  company named *person* to *post*

- Apply patterns to text and fill data base

Why It's Hard

- Lots of different patterns
  - different words:

    named, appointed, selected, chosen, promoted, …
  - different constructions:

    IBM named Fred president
    IBM announced the appointment of Fred as president
    Fred, who was named president by IBM
  - different names:

    George H. W. Bush, former President Bush, 41

Why It's Hard

- Ambiguity
  - Fred's appointment as professor *vs.*
    Fred's 3 PM appointment with the dean
  - outbreak of typhoid *vs.*
    outbreak of violence
- Complex structures
  - For the Federal Election Commission, Bush picked Justice Department employee and former Fulton County, Ga., Republican chairman Hans von Spakovsky for one of three openings.

Why It's Hard

- Reference
  - George Garrick has served as president of Sony USA for 13 years. *The company* announced *his* retirement effective *next May*.
  - IBM announced several new appointments yesterday. Fred Smith was named head of research.

Challenges of IE

- Collecting the patterns for a given relationship

- Identifying instances of these patterns in text

Challenges of IE

- Collecting the patterns for a given relationship

- **Identifying instances of these patterns in text**

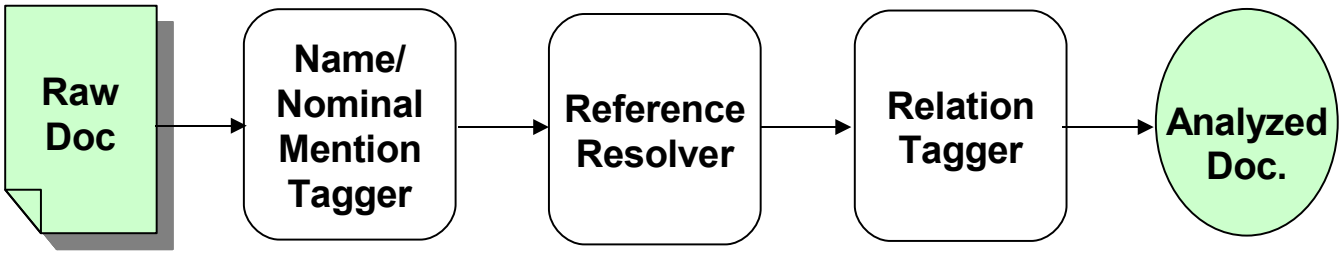
Identifying linguistic expressions

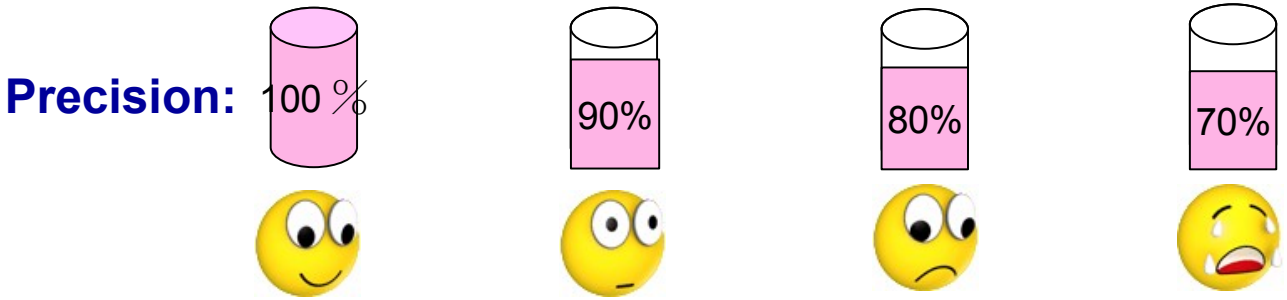
- To be at all useful, the patterns for IE must be stated structurally
  - patterns at the token level are not general enough
- So our main obstacle (as for many NLP tasks) is accurate structural analysis
  - name recognition and classification
  - syntactic structure
  - co-reference structure
  - if the analysis is wrong, the pattern won't match

Decomposing Structural Analysis

- Decomposing structural analysis into subtasks like named entities, syntactic structure, coreference has clear benefits …
  - problems can be addressed separately
  - can build separate corpus-trained models
  - can achieve fairly good levels of performance (near 90%) separately
    - well, maybe not for coreference
- But it also has problems ...

Sequential IE Framework



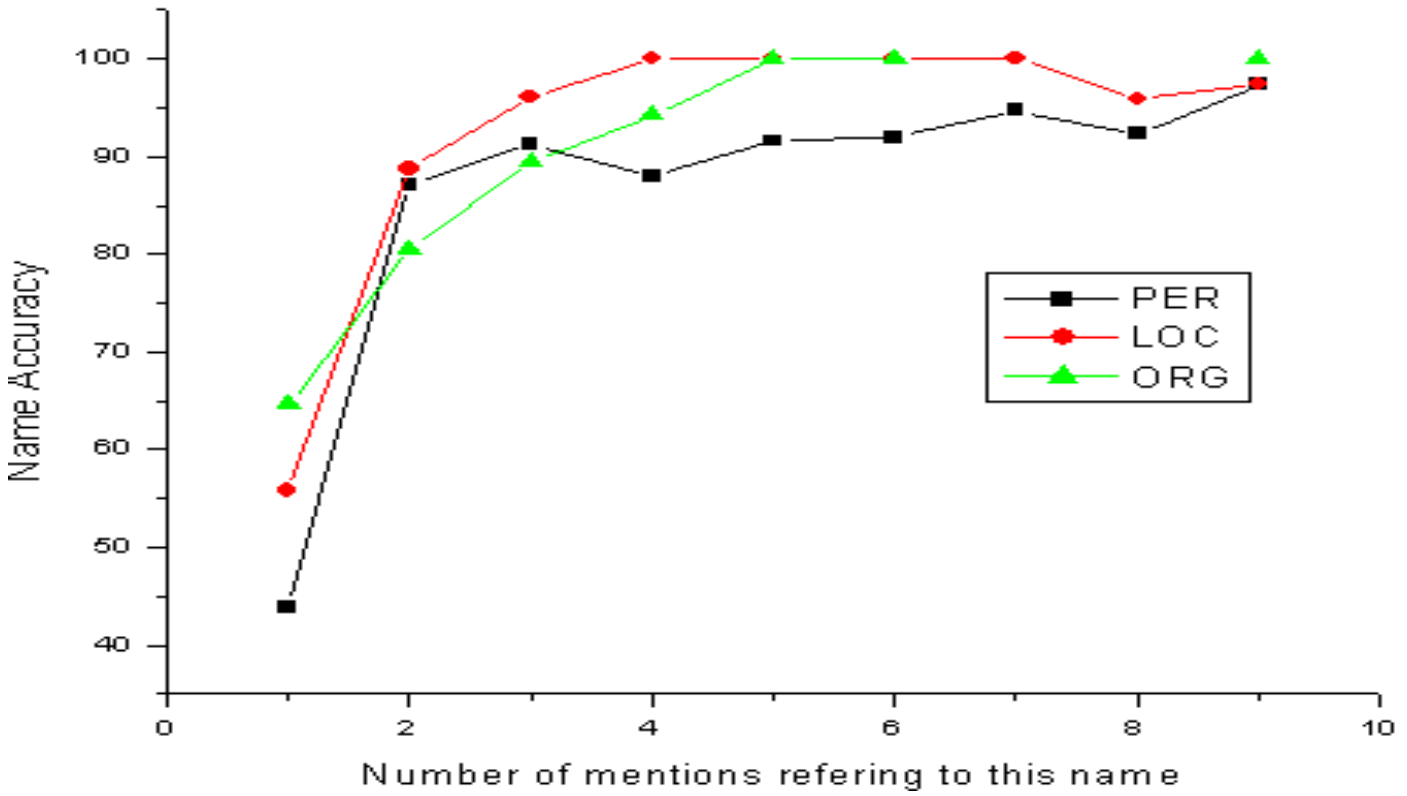


Errors are compounded from stage to stage

A More Global View

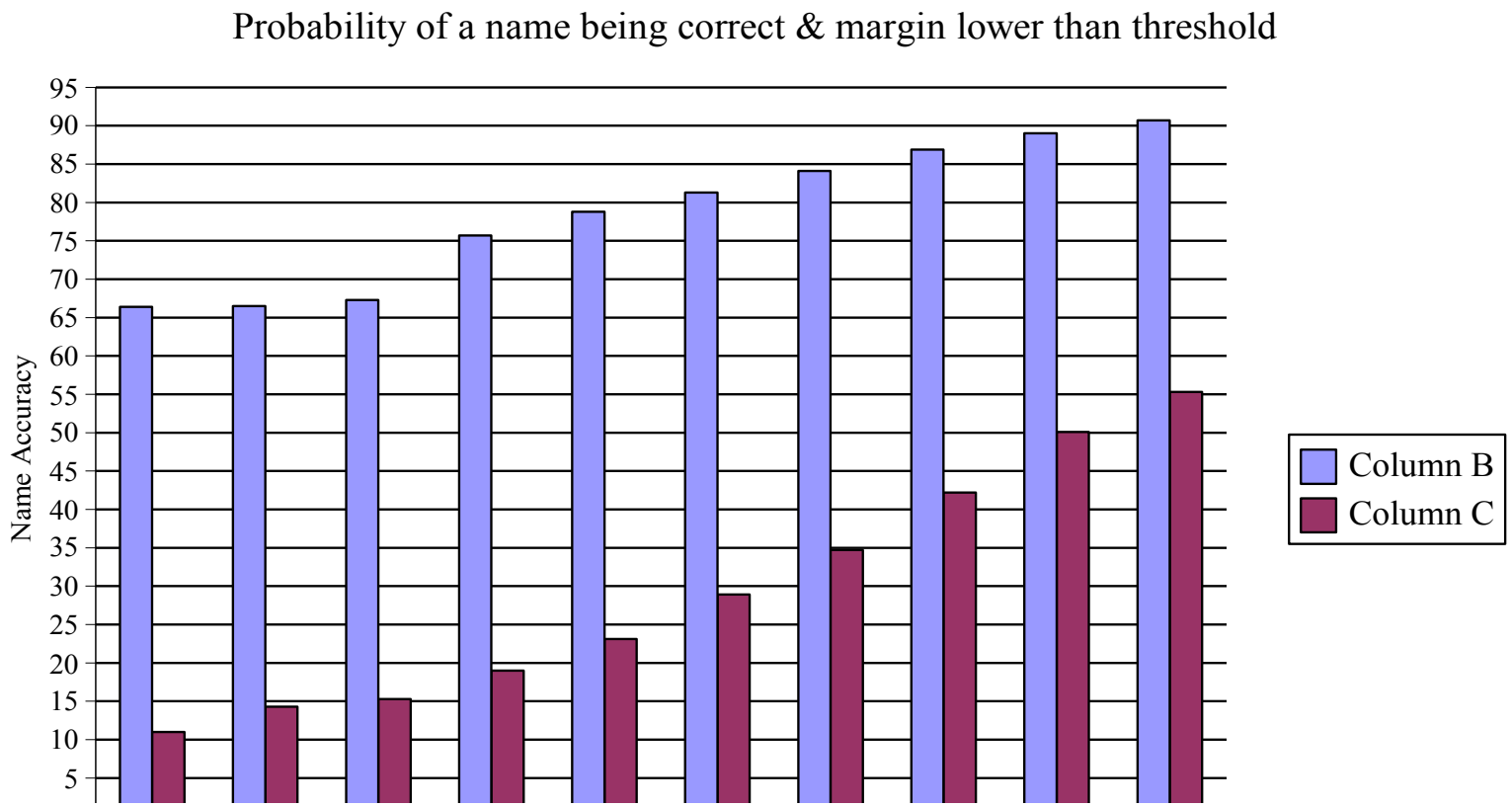
- Typical pipeline approach performs local optimization of each stage
- We can *take advantage* of interactions between stages by taking a more global view of 'best analysis'
- For example, prefer named entity analyses which allow for more coreference or more semantic relations

Names which can be coreferenced are much more likely to be correct



Counting only 'difficult' names for name tagger … small margin over 2nd hypothesis, not on list of common names

Names which can participate in semantic relations are much more likely to be correct



Sources of interaction

- Coreference and semantic relations impose type constraints (or preferences) on their arguments

- A natural discourse is more likely to be cohesive … to have ‘mentions’ (noun phrases) which are linked by coreference and semantic relations

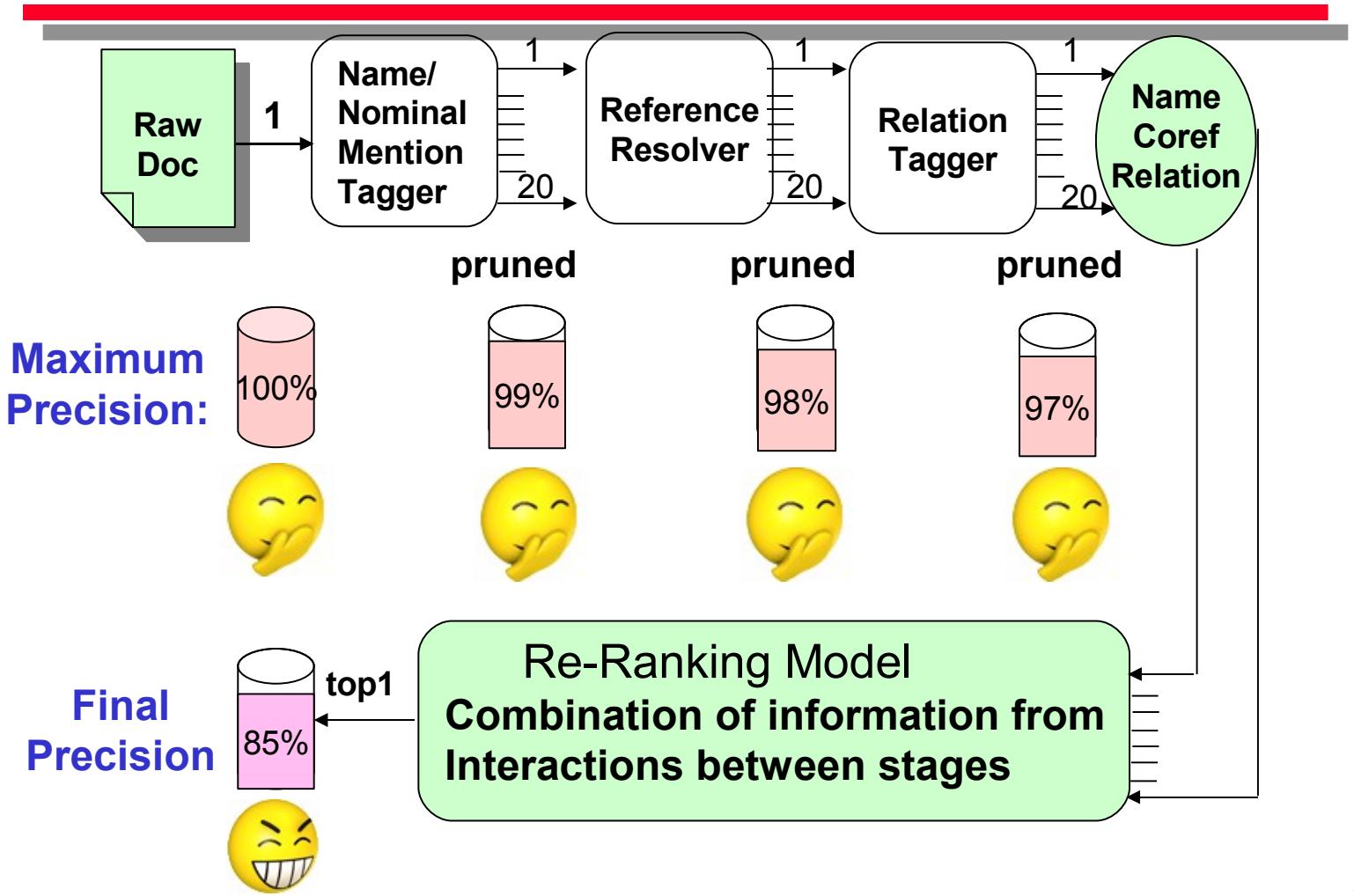
N-best

- One way to capture such global information is to use an N-best pipeline and rerank after each stage, using the additional information provided by that stage

(Ji and Grishman ACL 2005)

Reduced name tagging errors for Chinese by 20%
(F measure: 87.5 --> 89.9)

Multiple Hypotheses + Re-Ranking



Computing Global Probabilities

- Roth and Yih (CoNLL 2004) optimized a combined probability over two analysis stages
  - limited interaction to name classification and semantic relation identification
  - optimized product of name and relation probabilities, subject to constraint on types of name arguments
  - used linear programming methods
  - obtained 1%+ improvement in name tagging, and 2-4% in relation tagging, over conventional pipeline

Challenges of IE

- Collecting the patterns for a given relationship

- Identifying instances of these patterns in text

Lots of Ways of Expressing an Event

- Booth assassinated Lincoln
- Lincoln was assassinated by Booth
- The assassination of Lincoln by Booth
- Booth went through with the assassination of Lincoln
- Booth murdered Lincoln
- Booth fatally shot Lincoln

Patterns are Expensive

- It's hard to think of many patterns
- These patterns can be learned individually by annotating a large corpus
  - but consistent annotation is difficult and expensive
- Can we reduce the need for large-scale annotation?

Syntactic Paraphrases

- Some paraphrase relations involve the same words (or morphologically related words) and are broadly applicable
  - Booth assassinated Lincoln
  - Lincoln was assassinated by Booth
  - The assassination of Lincoln by Booth
  - Booth went through with the assassination of Lincoln
- These are *syntactic* paraphrases

Semantic Paraphrases

- Others paraphrase relations involve different word choices:
  - Booth assassinated Lincoln
  - Booth murdered Lincoln
  - Booth fatally shot Lincoln
- These are *semantic* paraphrases

Attacking Syntactic Paraphrases

- Syntactic paraphrases can be addressed through ‘deeper’ syntactic representations which reduce paraphrases to a common relationship:
  - chunks
  - surface syntax
  - deep structure (logical subject/object)
  - predicate-argument structure (‘semantic roles’)

Tree Banks

- Syntactic analyzers have been effectively created through training from tree banks
  - good coverage possible with a limited corpus

Predicate Argument Banks

- The next stage of syntactic analysis is being enabled through the creation of predicate-argument banks
  - PropBank (for verb arguments)
    - (Kingsbury and Palmer [Univ. of Penn.])
  - NomBank (for noun arguments)
    - (Meyers et al.)

PA Banks, cont'd

- Together these predicate-argument banks assign common argument labels to a wide range of constructs

  - The Bulgarians attacked the Turks
  - The Bulgarians' attack on the Turks
  - The Bulgarians launched an attack on the Turks

Depth vs. Accuracy

- Patterns based on deeper representations cover more examples

but

- Deeper representations are generally less accurate
- Leaves us with a dilemma …
  to use shallow (chunk) or deep (PA) patterns

Resolving the Dilemma

- The solution:
  - allow patterns at multiple levels
  - combine evidence from the different levels
  - use machine learning methods to assign appropriate weights to each level

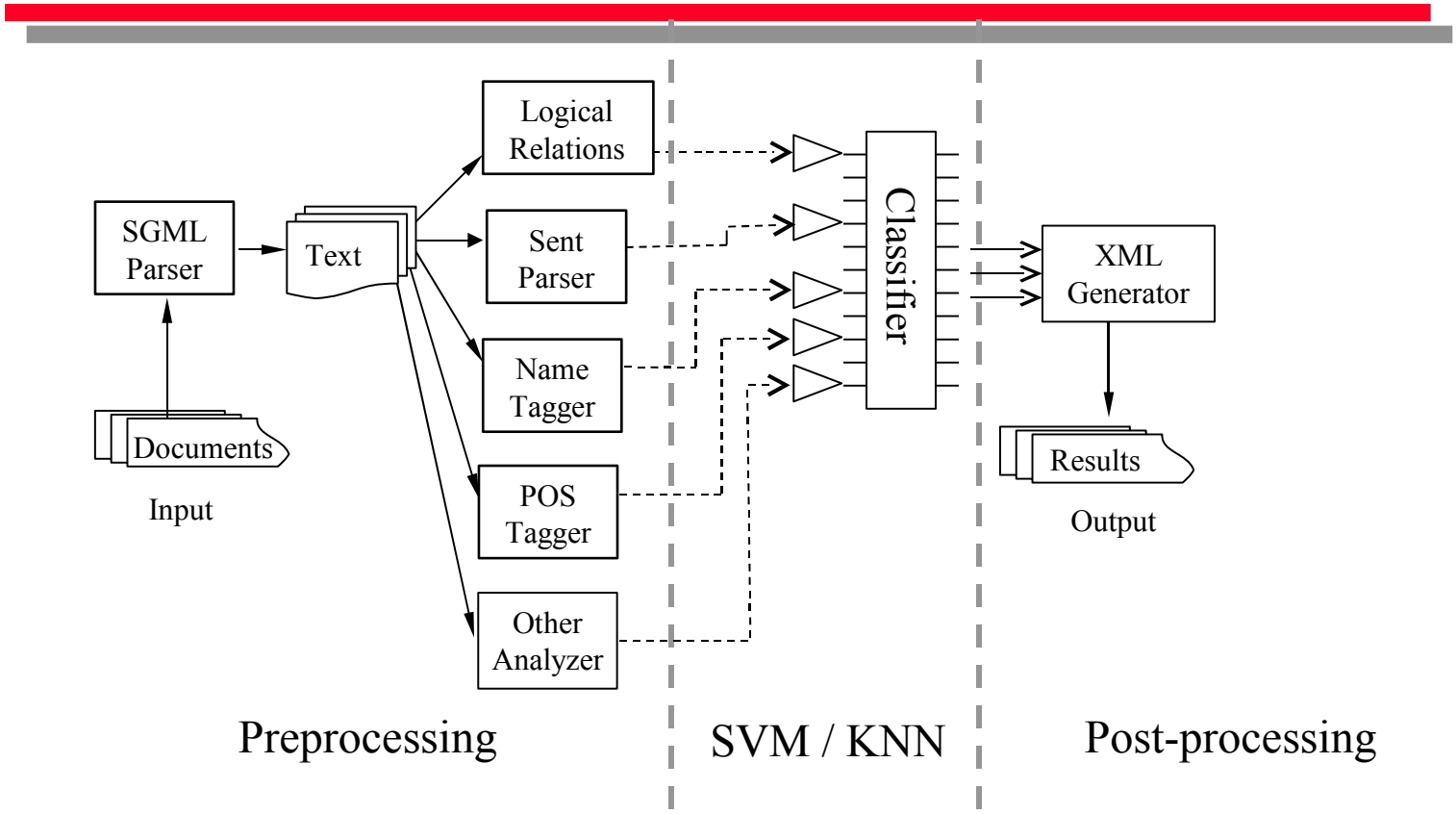
In cases where deep analysis fails, correct decision can often be made from shallow analysis

Integrating Multiple Levels

- Zhao applied this approach to relation and event detection
  - corpus-trained method
  - a ‘kernel’ measures similarity of an example in the training corpus with a test input

    separate kernels at
    - word level
    - chunk level
    - logical syntactic structure level
  - a composite kernel combines information at different levels

Kernel-based Integration



Benefits of Level Integration

- Zhao demonstrated significant performance improvements for semantic relation detection by combining
  - word,
  - chunk
  - logical syntactic relations

  over performance of individual levels

  (Zhao and Grishman ACL 2005)

Attacking Semantic Paraphrase

- Some semantic paraphrase can be addressed through manually prepared synonym sets, such as are available in WordNet
- Stevenson and Greenwood [Sheffield] (ACL 2005) measured the degree to which IE patterns could be successfully generalized using WordNet
  - measured on ‘executive succession’ task
  - started with a small ‘seed’ set of patterns

Seed Pattern Set for Executive Succession

- v-appoint = { appoint, elect, promote, name }
- v-resign = { resign, depart, quit}

Evaluating IE Patterns

- *Text filtering metric*: if we select documents / sentences containing a pattern, how many of the relevant documents / sentences do we get?

- Wordnet worked quite well for the executive succession task …

	seed		expanded	
	P	R	P	R
document filtering	100%	26%	68%	96%
sentence filtering	81%	10%	47%	64%

Challenge of Semantic Paraphrase

- But semantic paraphrase, by its nature, is more open ended and more domain-specific than syntactic paraphrase, so it is hard to prepare any comprehensive resource by hand
- Corpus-based discovery methods will be essential to improve our coverage

Two Approaches

- Predicates occurring with the same arguments

- Predicates frequently co-occurring in the same documents

Two Approaches

- Predicates occurring with the same arguments

- Predicates frequently co-occurring in the same documents

Paraphrase discovery

- Basic Intuition:
  - find pairs of passages which *probably* convey the same information
  - align structures at points of known correspondence (e.g., names which appear in both passages)

Fred *xxxxx* Harriet

Fred *yyyyy* Harriet

paraphrases

similar to MT training from bitexts

Evidence of paraphrase

- From almost parallel text:
  strong external evidence of paraphrase +
  a single aligned example
- From comparable text:
  weak external evidence of paraphrase +
  a few aligned examples
- From general text:
  using lots of aligned examples

Paraphrase from Translations

(Barzilay and McKeown ACL 01 [Columbia])

- Take multiple translations of same novel.
  - *High likelihood of passage paraphrase*
- Align sentences.
- Chunk and align sentence constituents

- Found lots of lexical paraphrases (words & phrases); a few larger (syntactic) paraphrases
- Data availability limited

Paraphrase from news sources

(Shinyama, Sekine, et al. IWP 03)

- Take news stories from multiple sources from same day
- Use word-based metric to identify stories about same topic
- Tag sentences for names; look for sentences in the two stories with several names in common
    - *moderate likelihood of sentence paraphrase*
- Look for syntactic structures in these sentences which share names
    - sharing 2 names, paraphrase precision 62% (articles about murder in Japanese)
    - sharing one name, at least four examples of a given paraphrase relation, precision 58% (2005 results, English, no topic constraint)

Relation paraphrase from multiple examples

Basic idea:

- If
  - expression R appears with several pairs of names
    - a R b, c R d, e R f, …
  - expression S appears with several of the same pairs
    - a S b, e S f, …
- Then there is a good chance that R and S are paraphrases

Relation paraphrase -- example

• Eastern Group	's agreement to buy	Hanson
• Eastern Group	to acquire	Hanson
• CBS	will acquire	Westinghouse
• CBS	's purchase of	Westinghouse
• CBS	agreed to buy	Westinghouse

(example based on Sekine 2005)

Relation paraphrase -- example

- Eastern Group 's agreement to **buy** Hanson
- Eastern Group to **acquire** Hanson

- CBS will **acquire** Westinghouse
- CBS 's **purchase** of Westinghouse
- CBS agreed to **buy** Westinghouse

select main linking predicate

Relation paraphrase -- example

- Eastern Group 's agreement to **buy** Hanson
- Eastern Group to **acquire** Hanson

- CBS will **acquire** Westinghouse
- CBS 's **purchase** of Westinghouse
- CBS agreed to **buy** Westinghouse

2 shared pairs ⇨ paraphrase link (buy ⇔ acquire)

Relation paraphrase, cont'd

- Brin (1998); Agichtein and Gravano (2000):
  - acquired individual relations (authorship, location)
- Lin and Pantel (2001)
  - patterns for use in QA
- Sekine (IWP 2005)
  - acquire all relations between two types of names
  - paraphrase precision 86% for person-company pairs, 73% for company-company pairs

Two Approaches

- Predicates occurring with the same arguments

- Predicates frequently co-occurring in the same documents

- Topic

- Set of documents on topic

- Set of patterns characterizing topic

Riloff Metric

- Divide corpus into relevant (on-topic) and irrelevant (off-topic) documents
- Classify (some) words into major semantic categories (people, organizations, …)
- Identify predication structures in document (such as verb-object pairs)
- Count frequency of each structure in relevant (R) and irrelevant (I) documents
- Score structures by (R/I) log R
- Select top-ranked patterns

Bootstrapping

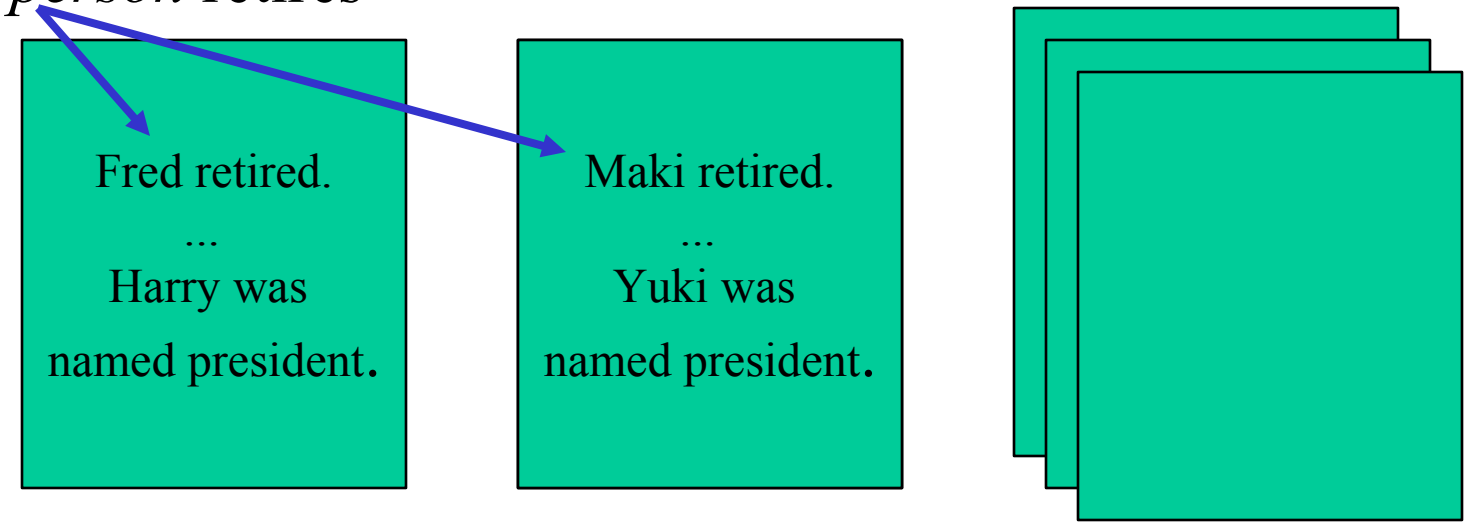
- Goal:
  find examples / patterns relevant to a given topic
  without any corpus tagging (Yangarber '00)
- Method:
  - identify a few seed patterns for topic
  - retrieve documents containing patterns
  - find additional structures with high Riloff metric
  - add to seed and repeat

#1: pick seed pattern

Seed: < *person* retires >

#2: retrieve relevant documents

Seed: < *person* retires >

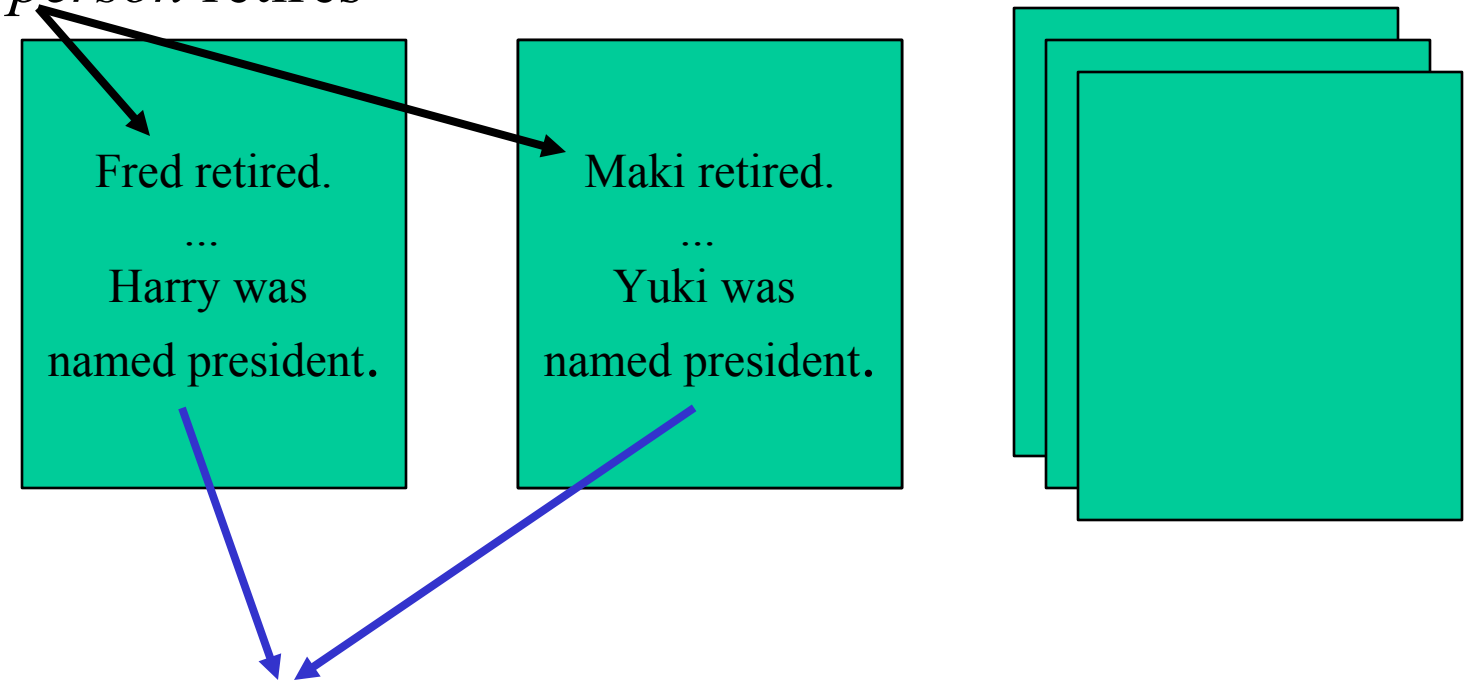


Relevant documents

Other documents

#3: pick new pattern

Seed: < *person* retires >



< *person* was named president >

appears in several relevant documents (top-ranked by Riloff metric)

#4: add new pattern to pattern set

Pattern set: < *person* retires >

< *person* was named president >

Applied to Executive Succession task

seed

- v-appoint = { appoint, elect, promote, name }
- v-resign = { resign, depart, quit, step-down }

- Run discovery procedure for 80 iterations

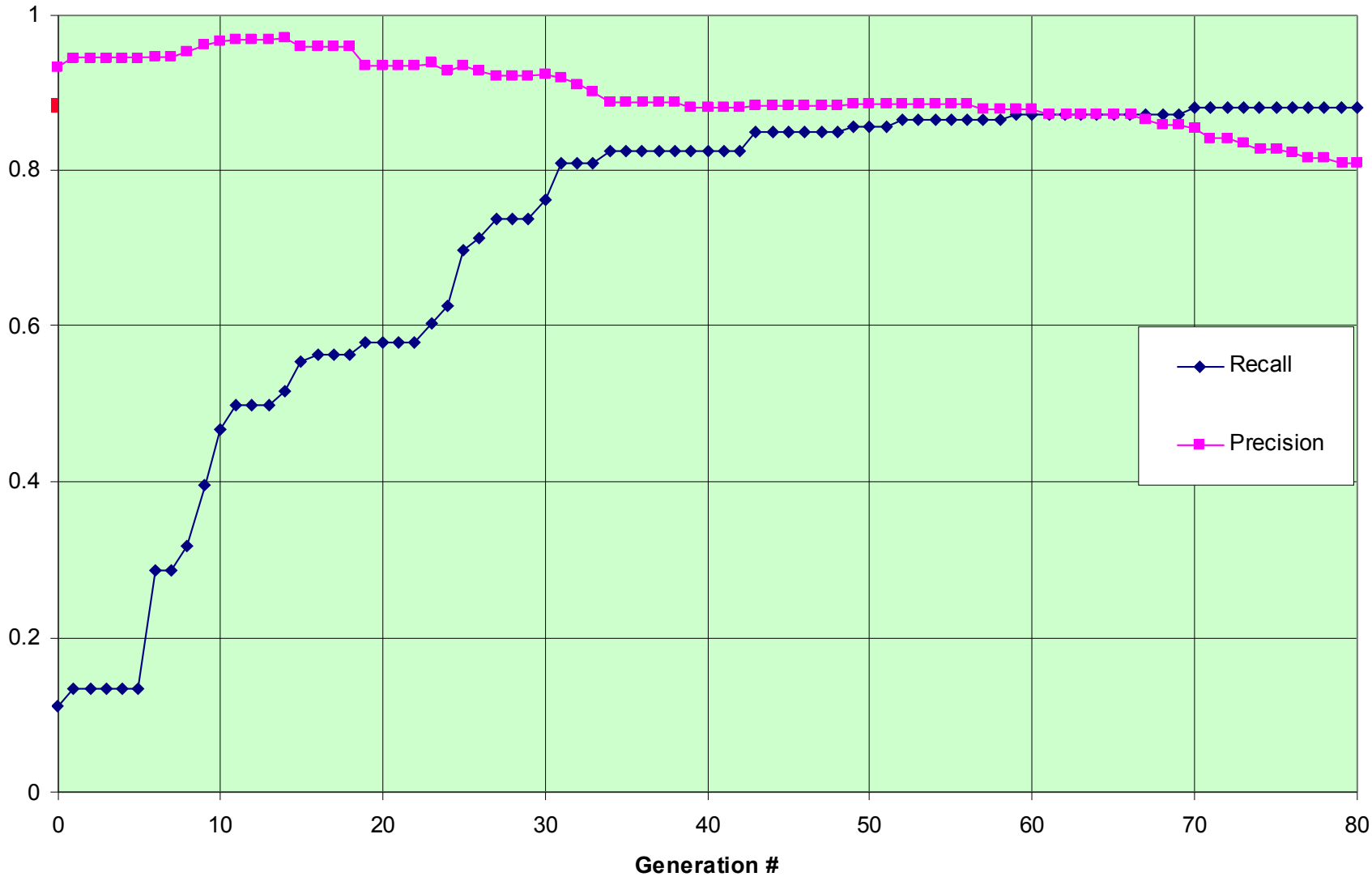
Discovered patterns

Evaluation: Text Filtering

- Evaluated using document-level text filtering
–

- Comparable to WordNet-based expansion
- Successful for a variety of extraction tasks

Document Recall / Precision



Evaluation: Slot filling

- How effective are patterns within a complete IE system?
- MUC-style IE on MUC-6 corpora

	training			test		
pattern set	*recall*	*precision*	*F*	*recall*	*precision*	*F*
seed	28	78	41	27	74	40
+ discovered	51	76	61	52	72	60
manual–MUC	54	71	62	47	70	56
manual–now	69	79	74	56	75	64

- Caveat: filtered / aligned by hand

Topical Patterns vs. Paraphrases

- These methods gather the main expressions about a particular topic
- These include sets of paraphrases
  - name, appoint, select
- But also include topically related phrases which are not paraphrases
  - appoint & resign
  - shoot & die

Pattern Discovery + Paraphrase Discovery

- We can couple topical pattern discovery and paraphrase discovery
  - first discover patterns from topic description (Sudo)
  - then group them into paraphrase sets (Shinyama)
- Result are semantically coherent extraction pattern groups (Shinyama 2002)
  - although not all patterns are grouped
  - paraphrase detection works better because patterns are already semantically related

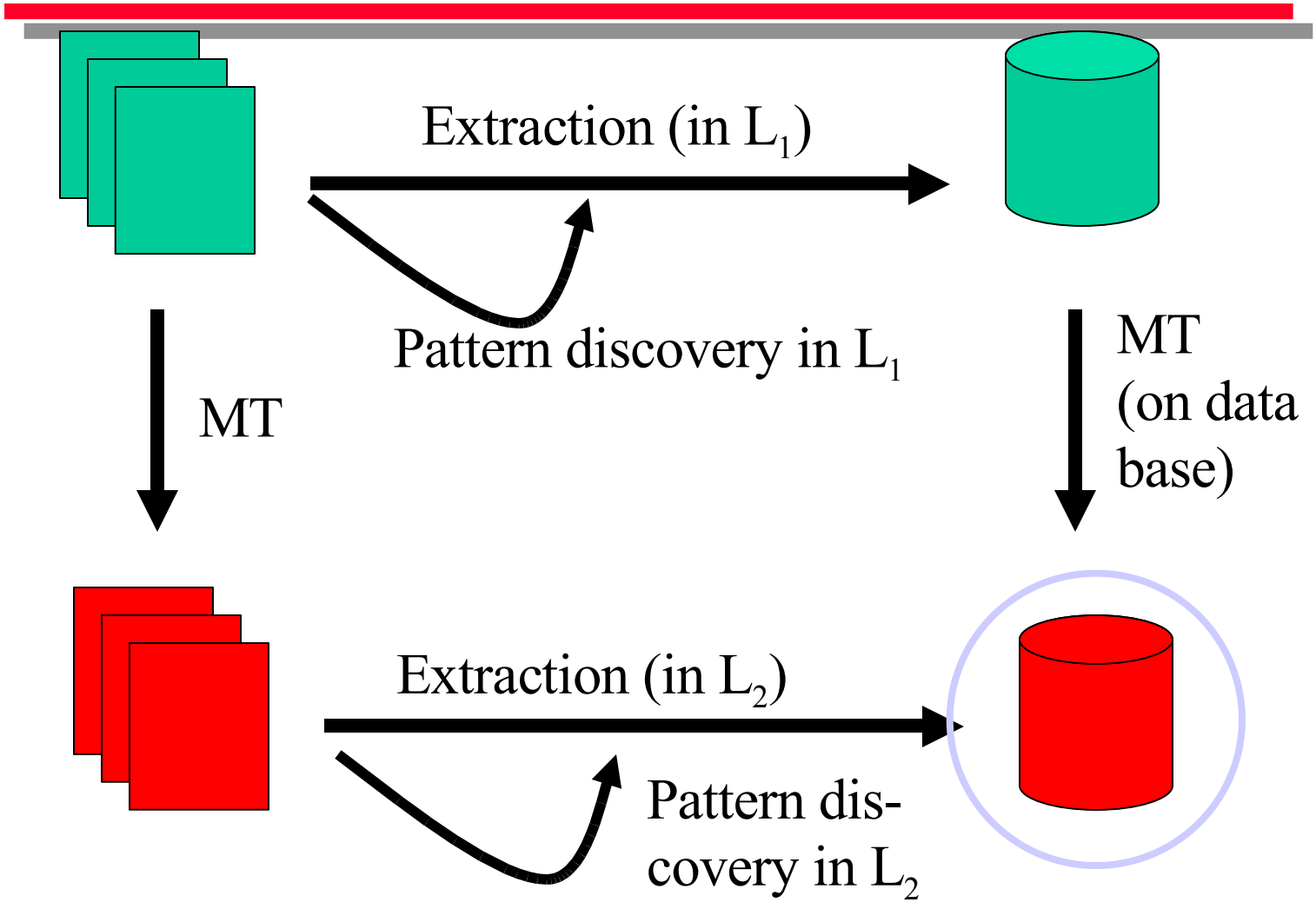
Paraphrase identification for discovered patterns (Shinyama et al 2002)

- worked well for executive succession task (in Japanese): precision 94%, coverage 47%
  - coverage = number of paraphrase pairs discovered / number of pairs required to link all paraphrases
- didn't work as well for arrest task … fewer names, multiple sentences with same name led to alignment errors

Cross-Language IE

- Suppose we want to build a data base in language L_1 from texts in L_2

Two paths to the goal



Experimental results

- Japanese --> English (Sudo et al. 2004)
  - Management succession task
  - Compared on slot recall
- Extraction on MT output: 41%
  - improved to 52% by identifying names in L_1 and then projecting them to L_2
  - argument structure sometimes lost by MT
- L1 extraction followed by MT: 60%

Multilingual IE Programs in U.S.

- ACE evaluation (open)
  - previously trilingual: Arabic, Chinese, English
  - Spanish added for 2007
  - cross-language entity translation task added for 2007
  - http://www.nist.gov/speech/tests/ace/ace07/index.htm
- tracks
  - entities
  - relations between entities
  - events

Multilingual IE Programs in U.S.

GALE Program (begun fall 2005)

- trilingual: Arabic, Chinese, English
- involves ASR, MT, IR, and IE
  - IR / IE component involves answering questions about multi-lingual corpus
    - "describe reaction of <country> to <event>"
    - "list attacks in <location> during <time> and related deaths"
  - inherently cross-lingual … responses only in English
    - looking to combine cross-lingual IE paths for maximum performance

Conclusion

- Current IE technology provides a powerful tool for *targeted* searches in large text collections
- Recent basic research on NLP methods offers significant opportunities for improved IE performance and portability
  - global optimization to improve analysis performance
  - richer treebanks to support greater coverage of syntactic paraphrase
  - corpus-based discovery methods to support greater coverage of semantic paraphrase