

Information Extraction: Building Graphs from Natural Language Text

Ralph Grishman

April 2007



What is Information Extraction?

building
structured information
from
(unstructured) text



What is Information Extraction?

- Information extraction is the identification, in text, of specified classes of
 - names / entities
 - relations
 - events
- For relations and events, this includes finding the participants and modifiers (date, time, location, etc.).



Information Extraction

- In other words, we build a data base or network with the information on a given relation or event:
 - people's jobs
 - people's whereabouts
 - merger and acquisition activity
 - disease outbreaks



Extraction Example

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



Extraction Example

- 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



IE Evaluations

Work in IE has been prodded and shaped by a series of IE evaluations

- MUC [Message Understanding Conference] evaluations (second millenium)
- ACE [Automated Content Extraction] evaluations (third millenium)

IE has gradually expanded from a single task (event extraction) to an integrated set of tasks



ACE Task Hierarchy

- Entities
 - Participants in relations and events:
people, organizations, political entities, other locations, ...
 - Referred to by *mentions*: names and nominal phrases
- Time expressions
- Relations
 - Link pairs of entities
 - Mostly long-term relationships:
 - person-person (family, acquaintance),
 - person-org (employment, citizenship)
- Events
 - Multiple entity arguments
 - Time and location modifiers



Building an Ace Graph

George W. Bush vacationed last week with his brother Jeb, the former governor of Florida. He flew to Mexico City today to meet with Felipe Calderon, the new president of Mexico.



Entity Mentions

George W. Bush vacationed last week with his brother Jeb. He flew to Mexico City today to meet with Felipe Calderon, the new president of Mexico.

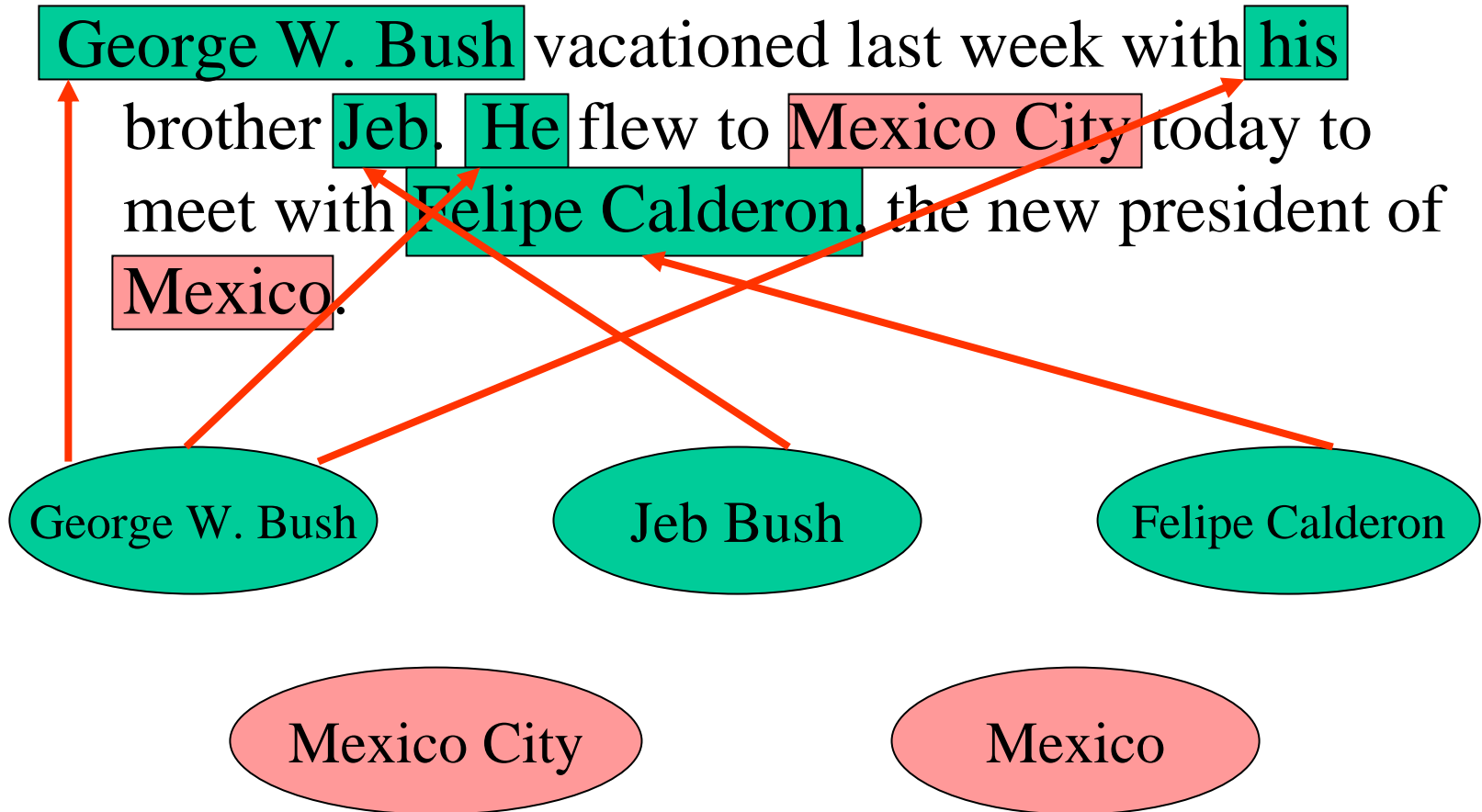
Entity mentions:

People

Political entities

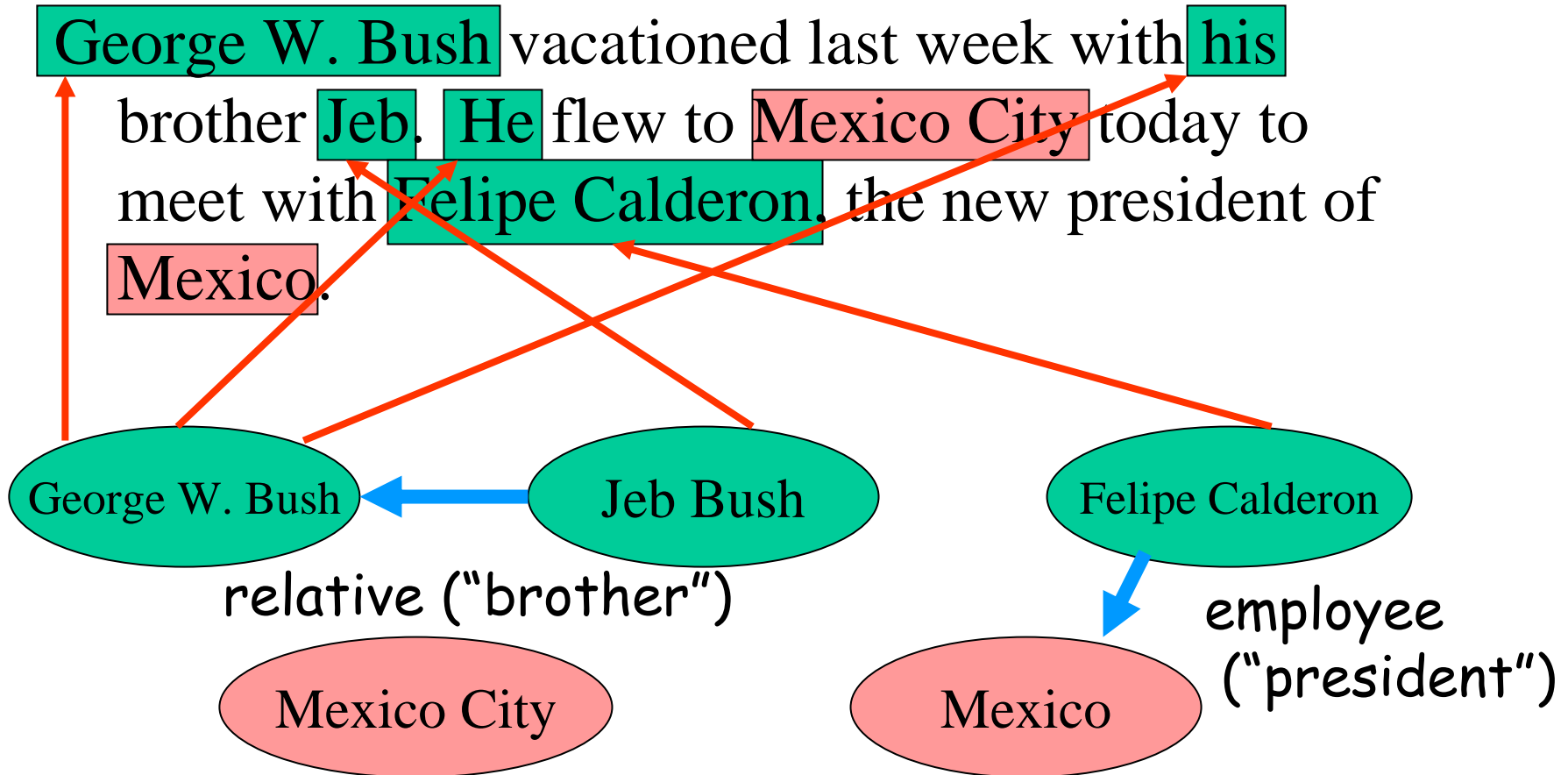


Entities





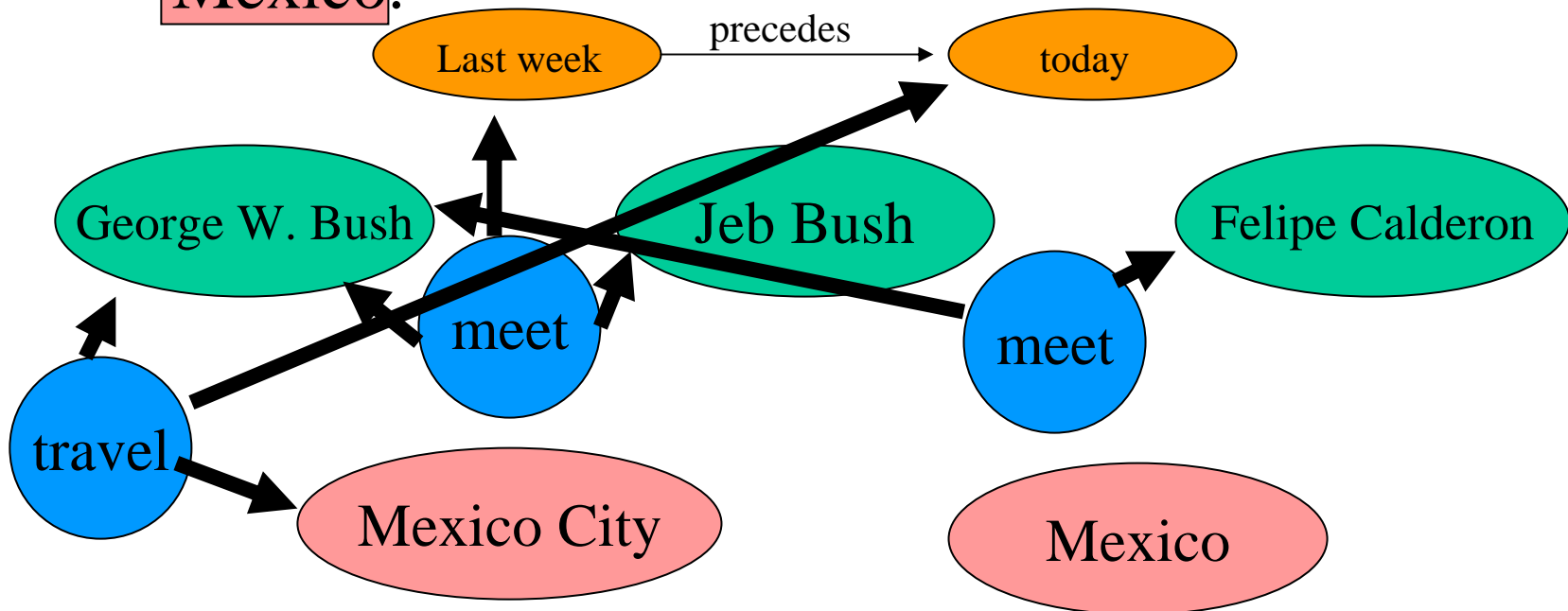
Relations





Events

George W. Bush vacationed last week with his brother Jeb. He flew to Mexico City today to meet with Felipe Calderon, the new president of Mexico.





Performance

Typical performance (F measure):

- Named entity mentions, 90%
- Entity mentions, 85%
- Entities, 75-80%
- Relation mentions, 40-60% (type & argument accuracy)
- Event mentions, 40% (type & argument accuracy)
 - Precision better than recall for relations and events
 - Ace metrics strict, event task new ...
MUC performance up to 50-60% on events, 60-70% on relations

Despite errors, relations and events can provide significant information beyond that available through IR / bag of word / name proximity methods for building networks of nodes



Building an IE System

Old-fashioned way ...

hand-coded rules for each stage

- For names
 - person = <first name> [<initial>] <capitalized token>
- For entities (coreference rules)
- For relations and events

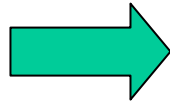


-
- Build event patterns:

person was appointed as *post* of *company*
company named *person* to *post*

- Apply patterns to text and fill data base

*person*₁ was appointed as
*post*₂ of *company*₃



PERSON	#1
POSITION	#2
COMPANY	#3
STATUS	in



IE by Supervised Learning

- Create a corpus-trainable statistical model for each stage
 - Sequential model (HMM, CRF, ...) for name recognition
 - Sequential model for NPs (nominal entities)
 - Classifier for entity types
 - Probabilistic coreference model
 - Probabilistic relation model based on heads of two entity mentions, syntactic relation, intervening terms
- Annotate a lot of data and train model
- Apply models in succession



Problems with this Approach

- Annotation quite expensive
- Zipfian distribution of patterns means that annotation of consecutive text is inefficient ... the same pattern is annotated many times
- Succession of slightly inaccurate models produces quite inaccurate results



Unsupervised learning for events?

- The intuition:

if we collect documents D_R relevant to the scenario, patterns relevant to the scenario will occur more frequently in D_R than in the language as a whole

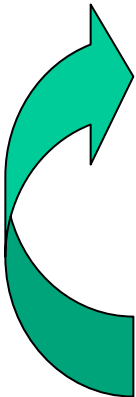
First used by Riloff (1996)

Developed at NYU by Yangarber (2000, 2003), Sudo (2001, 2003)



Automated Extraction Pattern Discovery

- Goal:
find examples / patterns relevant to a given scenario
without any corpus tagging
- Method:
 - identify a few seed patterns for scenario
 - retrieve documents containing patterns
 - find subject-verb-object pattern with
 - high frequency in retrieved documents
 - relatively high frequency in retrieved docs vs. other docs
 - add pattern to seed and repeat





#1: pick seed pattern

Seed: < *person* retires >

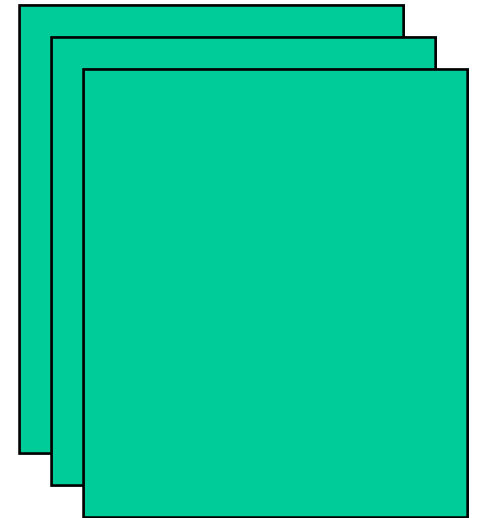


#2: retrieve relevant documents

Seed: < *person* retires >

Fred retired.
...
Harry was
named president.

Maki retired.
...
Yuki was
named president.



Relevant documents

Other
documents

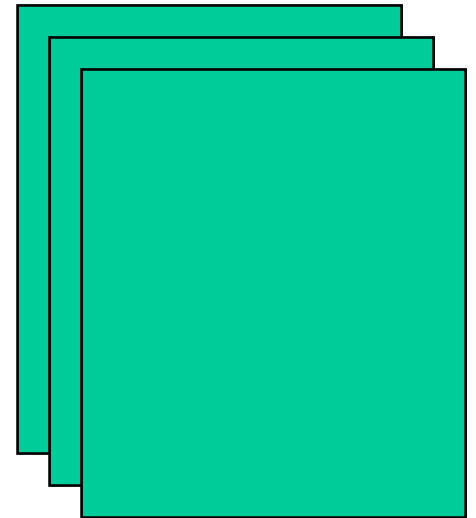


#3: pick new pattern

Seed: < *person* retires >

Fred retired.
...
Harry was
named president.

Maki retired.
...
Yuki was
named president.



< *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 >



Experiment: two seed patterns

Subject	Verb	Object
company	v-appoint	person
person	v-resign	-

- v-appoint = { appoint, elect, promote, name }
- v-resign = { resign, depart, quit, step-down }
- Run discovery procedure for 80 iterations

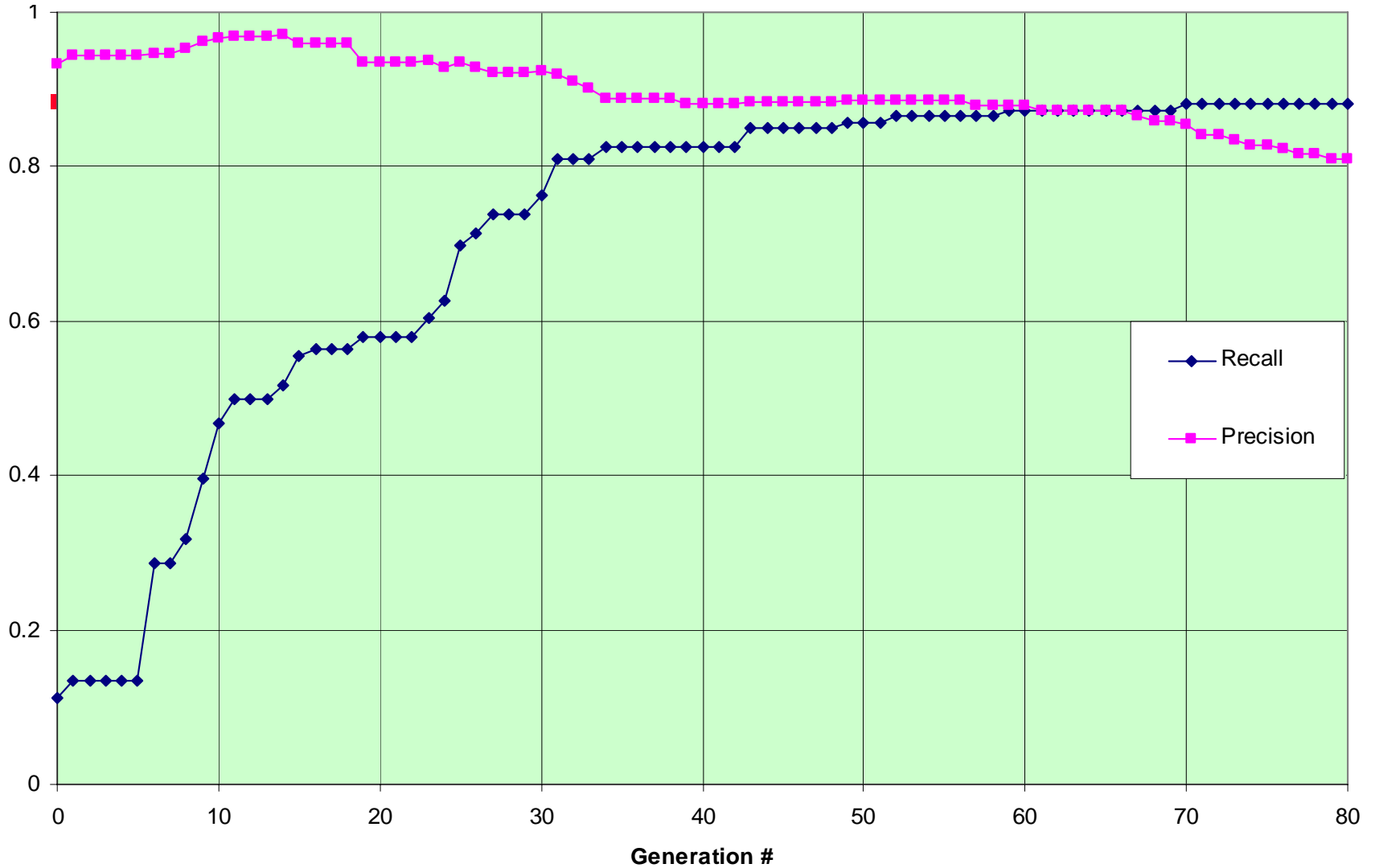


Discovered patterns

Subject	Verb	Object
company	v-appoint	person
person	v-resign	-
person	succeed	person
person	be become	president officer chairman executive
company	name	president ...
person	join run leave	company
person	serve	board company
person	leave	post



Effectiveness in Finding Relevant Documents



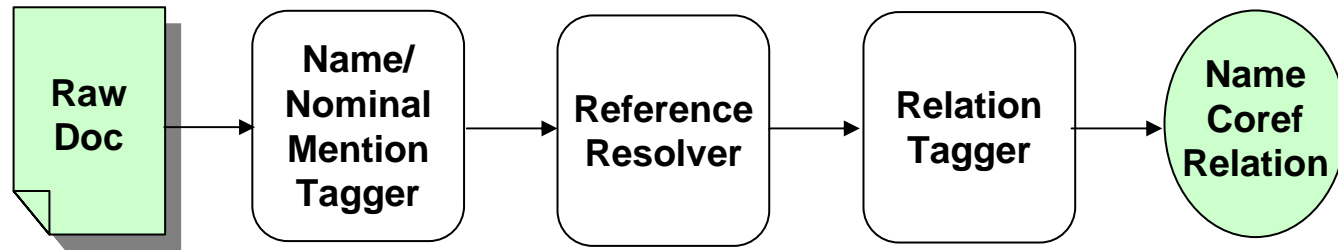


Problems with this Approach

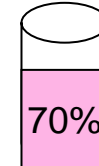
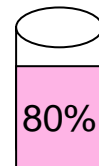
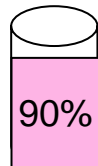
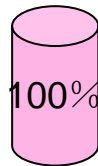
- Annotation quite expensive
- Zipfian distribution of patterns means that annotation of consecutive text is inefficient ... the same pattern is annotated many times
- Succession of slightly inaccurate models produces quite inaccurate results



Sequential IE Framework



Precision:



Errors are compounded from stage to stage



Intuition for Joint Inference

Later stages may provide clues to help decisions of earlier stages ...

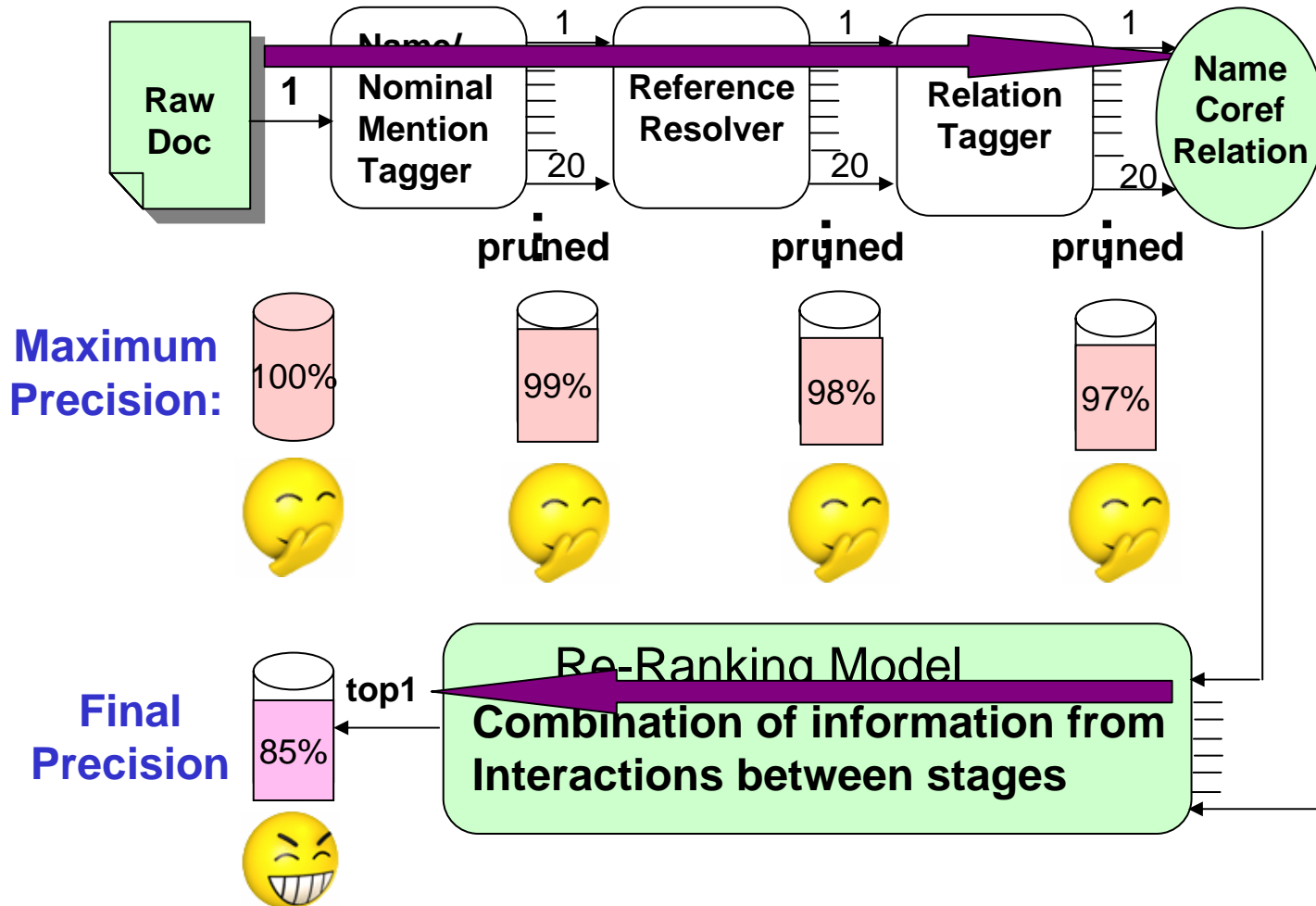
Relations and events impose type preferences which can resolve name ambiguities ...

- Mumble Crumble got married
- His brother, Mumble Crumble
- Harry flew to Mumble Crumble for vacation
- Mumble Crumble is delicious

In languages without case or token delimiters, relation and event patterns can also help resolve name boundaries



New IE Framework: Multiple Hypotheses + Re-Ranking





New IE Framework: Multiple Hypotheses + Re-Ranking

1. Each stage generates and propagates N-Best multiple hypotheses, after pruning
2. Add a re-ranking model based on stage-interaction features
3. Determine a new ranking for the n-best extraction hypotheses
4. The new top hypothesis is the final extraction result



Experiment Results (Chinese Names)

System		Performance	Precision	Recall	F-measure
(1)	Baseline		88.4	86.7	87.5
(2)	(1) + Word Clustering by Relation		89.0	87.4	88.2
(3)	(2) + Re-ranking by Coreference		90.0	88.8	89.4
(4)	(3) + Re-ranking by Relation		91.2	88.6	89.9



In Summary ...

- IE using Ace or similar hierarchy can provide a rich graph structure for text information
- Accuracy (coverage and precision) for relations and events remains limited, due to ‘combination of small errors’ effect
 - But still extracts significant value beyond that available from bag-of-words / IR methods
 - Will improve through use of
 - Unsupervised and semi-supervised learning methods
 - Joint inference to combine information across stages