COMP5046: Information Extraction

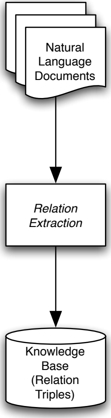
Joel Nothman

joel.nothman@sydney.edu.au

School of Information Technologies
University of Sydney

2018-05-01

Information extraction: structured data from text



Example Documents

```
...
American saxophonist David Murray recruited Amidu Berry.
...
```

```
...
Cdc3+ encodes profilin, an actin-monomer-binding protein.
...
```

Example Relation Triples

Entity1	**Entity2**	**Relation Type**
“David Murray”	“American”	CITIZEN-OR-RESIDENT
“David Murray”	“Amidu Berry”	BUSINESS

Entity1	**Entity2**	**Relation Type**
“Cdc3+”	“profilin”	ENCODE
“profilin”	“actin-monomer”	BIND

Common Sub-tasks of Information Extraction

1. named entity recognition
2. coreference resolution
3. relation extraction
4. temporal expression recognition
5. event/fact extraction
6. temporal/event interrelation

How are extracted entities related to each other?

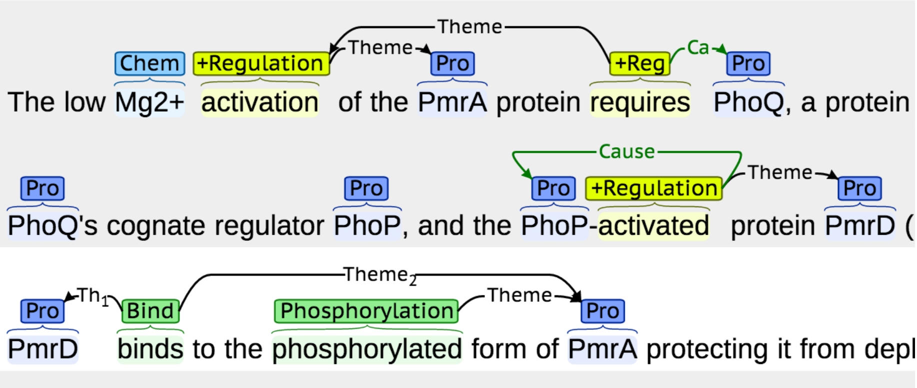
- The *identity* relation ⇒ coreference resolution
- Other relations (*ORG headquartered in LOC*, *PER member of ORG*)
- Entities may be participants or attributes of an event:
  an event's *who*, *what*, *when*, *where*
- Events may also be interrelated (*identity*, *part of*, *precedes*, *causes*)

Shared tasks

- Message Understanding Conference (MUC): 1990s, DARPA
- Automated Content Extraction (ACE): 2000s, NIST
- Conference on Natural Language Learning (CoNLL): 2002-3
- Text Analysis Conference Knowledge Base Population (TAC KBP): 2009–, NIST

- BioCreative
- Genia
- BioNLP

BioNLP example



The low Mg2+ activation of the PmrA protein requires PhoQ, a protein
PhoQ's cognate regulator PhoP, and the PhoP-activated protein PmrD (
PmrD binds to the phosphorylated form of PmrA protecting it from depl

Transform problems into those we know how to solve

- measuring association
- regular expressions
- language modelling
- similarity in vector space
- classification
- sequence labelling
- later: tree structured labelling

- (almost?) all have been applied to information extraction tasks

Supervision

- Datasets often small and closely tied to interests of sponsors
- Supervised IE has been domain-specific
- Unsupervised has helped to filled the gap
- Recently, large collaborative resources available (e.g., Wikipedia)
  $\Longrightarrow$ serve as large-scale knowledge bases
  $\Longrightarrow$ sources of noisy training data

Part I

Entity coreference

Coreference Resolution

- NER only produces a list of mention strings
  $\Longrightarrow$ how track through/across documents?
- Coreference: when mentions refer to the same entity
  E.g.: `When` **David Murray** `visited` **Senegal**`,` **he** `recruited` **Amidu Berry** `from` **Positive Black Soul**`.` **David** `also recruited` **DJ Awadi**`.`
  `David Murray`**,** `he` **and** `David` **refer to the same person.**
- Coreference resolution: cluster entity mentions within document

Another example

The battered US Navy destroyer Cole has begun its journey home from Yemen … Flanked by other warships and guarded by aircraft, the ship was towed out of Aden Harbor to rendezvous with a huge Norwegian transport vessel.

`OntoNotes 5: bn/voa/00/voa_0068`

Another example

The battered US Navy destroyer Cole has begun its journey home from Yemen … Flanked by other warships and guarded by aircraft, the ship was towed out of Aden Harbor to rendezvous with a huge Norwegian transport vessel.

`OntoNotes 5: bn/voa/00/voa_0068`

Common approach. For each mention in turn:

1. identify candidate *antecedents* from a window of recent mentions
2. score each candidate with a **classifier trained on mention pairs**
3. select the best, perhaps updating entity knowledge
4. proceed to next mention

Haghighi and Klein (2009): Simple Coref w/ Rich Features

- separate components for syntactic, semantic, and discourse constraints
- syntactic: Various constraints based on phrase structure tree
- semantic: Compatibility
  (*spokesperson* can *announce*, *Microsoft* is a *company*)
  Extracted from large unlabelled text corpus
- Discourse: Salience (importance of entity/mention in context)
  E.g.: `Nintendo of America announced its new console.`
  `its` corefers with `Nintendo of America` not `America`
- requires deep linguistic preprocessing
- **outperforms almost all unsupervised and many supervised systems**

http://aclweb.org/anthology/D/D09/D09-1120.pdf

Rughunathan et al. (2010): Precise Multi-pass Sieve

- precision-ordered passes in sieve:
  1. exact match
  2. appositives (`Australia's Prime Minister, Malcolm Turnbull`), Predicate nominatives (`Malcolm Turnbull is Australia's PM.`)
  3. strict head matching (`Turnbull` matches antecedent cluster)
  4. three variants/relaxations of head matching
  5. pronouns (match number, gender, etc.)
- F-scores comparable to Haghighi and Klein
- **precision higher than Haghighi and Klein**

`http://www.aclweb.org/anthology/D/D10/D10-1048.pdf`

- Easy-first approach made statistical by Stoyanov and Eisner (2012)

Coreference resolution is only part of the solution

- Coreference resolution: cluster entity mentions within document **no cross-document tracking**
- Cross-document coreference resolution
- Other relations: discourse understanding involves recognising **near-identity** relationships, such as:
  - Zuckerberg ⇔ Facebook
  - Canberra ⇔ [Government of] Australia
  - Al Qaeda ⇔ the suicide bomber

Who is Michael Jordan?

wikipedia/Michael Jordan

Thanks to an enterprising thief at the Orlando Arena, Michael Jordan became the best athlete to ever wear number 12. . .

wikipedia/Michael I. Jordan

Michael Jordan: University of California, Berkeley
For contributions to the theory and application of machine learning

twitter/@AM_MJordan

Michael Jordan: West Coast Editor, Automobile Magazine Los Angeles, California

Why is this useful?

- Entity-oriented document access
- Integrating structured semantic information
- Reducing ambiguity

- like **named entity recognition**
- like **cross-document coreference resolution**
- like **word-sense disambiguation**
- but specialised for references to entities in a knowledge base

- other names: wikification, entity linking

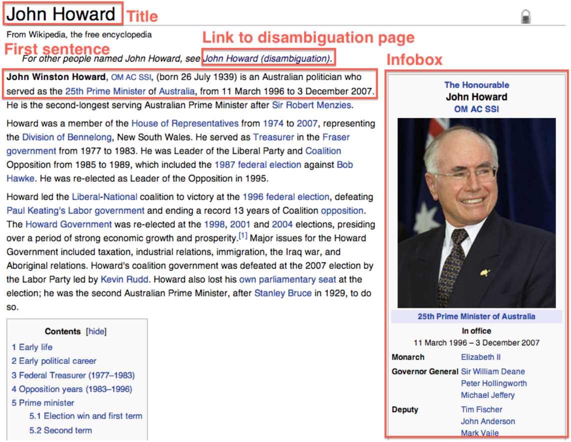
NEL: grounding entity mentions to a knowledge base

- Input, a Wikipedia-derived KB (n=800K), web/newswire docs and queries like:

```
<query id="EL\_00102">
  <name>Adams</name>
  <docid>eng-NG-31-142265-10040632</docid>
  <beg>4601</beg><end>4606</end>
</query>
```

- Return an id that clusters coreferent mentions:
  - ENT_001
  - NIL or NIL_001
- Variant: input plain text; systems also need to perform NER

Wikipedia is a rich multilingual KB



John Howard

From Wikipedia, the free encyclopedia

For other people named John Howard, see John Howard (disambiguation).

John Winston Howard, OM AC SSI, (born 26 July 1939) is an Australian politician who served as the 25th Prime Minister of Australia, from 11 March 1996 to 3 December 2007. He is the second-longest serving Australian Prime Minister after Sir Robert Menzies.

Howard was a member of the House of Representatives from 1974 to 2007, representing the Division of Bennelong, New South Wales. He served as Treasurer in the Fraser government from 1977 to 1983. He was Leader of the Liberal Party and Coalition Opposition from 1985 to 1989, which included the 1987 federal election against Bob Hawke. He was re-elected as Leader of the Opposition in 1995.

Howard led the Liberal-National coalition to victory at the 1996 federal election, defeating Paul Keating's Labor government and ending a record 13 years of Coalition opposition. The Howard Government was re-elected at the 1998, 2001 and 2004 elections, presiding over a period of strong economic growth and prosperity.[1] Major issues for the Howard Government included taxation, industrial relations, immigration, the Iraq war, and Aboriginal relations. Howard's coalition government was defeated at the 2007 election by the Labor Party led by Kevin Rudd. Howard also lost his own parliamentary seat at the election; he was the second Australian Prime Minister, after Stanley Bruce in 1929, to do so.

Contents [hide]

The Honourable
John Howard
OM AC SSI

25th Prime Minister of Australia

In office
11 March 1996 – 3 December 2007

Monarch	Elizabeth II
Governor General	Sir William Deane Peter Hollingworth Michael Jeffery
Deputy	Tim Fischer John Anderson Mark Vaile

- Categories
- Redirects
- Link graph

General approach

- Retrieve candidates for given name
- Score each candidate according to:
  - entity **popularity**, or prior likelihood of entity given name
    - e.g. number of incoming links to page
  - **compatibility** of context with what we know about entity
    - e.g. cosine between BOW of mention context and Wikipedia text

Issues matching mentions to Wikipedia content

- Names may not match exactly, e.g. Little Johnny Howard
- Context words may not match exactly (sparsity)
- *Mismatch* may be more important to eliminate a candidate
  - wrong age, wrong nationality, wrong spouse, wrong occupation
- Some context information may be especially informative:
  - weighting by distinctiveness: TF.IDF; learnt weights
  - authors provide most informative context when introducing an entity: Australian actor John Howard vs Former prime minister JH

- Best approaches represent KB and context in latent feature space
  - Compression of sparse vocabulary
  - Incorporating free text and structured knowledge
  - Enriched representations of KB entity

The long tail remains a challenge for named entity disambuguation

- For many entities, choosing the most popular candidate is overwhelmingly right
- Entities without much associated knowledge are hard
- Beyond Wikipedia: LinkedIn, IMDB, Facebook
- Identifying when a referenced entity is not in the KB
- Clustering entity references that are not in the KB (NIL clustering)
    - only pertains to some applications
    - if there's no popular referent for some name then mentions are almost always non-coreferent
    - for unusual names matching by name may be good enough
    - for any small collection of text

Events are fundamental to communication

Somali Gunmen Release Ship Carrying Tsunami Aid

The United Nations **says** Somali gunmen who **hijacked** a U.N.-**chartered** vessel **carrying** food **aid** for **tsunami** victims have **released** the ship after **holding** it for more than two months.

who, what, when, where, why, which

Events are fundamental to communication

Somali Gunmen Release Ship Carrying Tsunami Aid

The United Nations **says** Somali gunmen who **hijacked** a U.N.-**chartered** vessel **carrying** food **aid** for **tsunami** victims have **released** the ship after **holding** it for more than two months.

who, what, when, where, why, which

absolute and relative temporal references

Event coreference is particularly challenging

Coreferent arguments $\not\Rightarrow$ coreferent events

1. …Somali gunmen who hijacked a U.N.-[chartered]$_a$ vessel …
2. The World Food Program [hired]$_a$ the Kenyan vessel …

Changing frame of reference

1. Two have died after [an explosion at the Boston Marathon]$_b$.
2. Three have died after [a terror attack at the Boston Marathon]$_b$.
3. [the Boston Marathon bombings]$_b$

Scriptal events like hijacked have wide and narrow readings

Part II

Relation and Fact Extraction

Extracting relations from text

`In April 2011 , Prime Minister Mykola Azarov of Ukraine met with the President of Brazil , Dilma Rousseff , in Sanya .`

Binary relation triples:

- (`Mykola Azarov`, *president-of*, `Ukraine`)
- (`Mykola Azarov`, *title*, `President`)
- (`Dilma Rousseff`, *president-of*, `Brazil`)
- (`Dilma Rousseff`, *met*, `Mykola Azarov`)
- (`Dilma Rousseff`, *visited*, `Sanya`)

RE as Knowledge Base Population

The University of Sydney is an Australian public research university in Sydney, Australia. Founded in 1850, it is Australia's first university.

Wikipedia infobox fields for University of Sydney:

- *Type*: Public university
- *Established*: 1850
- *Location*: Sydney, Australia

Using RE

- Creating structured data from unstructured text
- Create or extend knowledge bases
- Support other tasks, e.g. search and question answering
- Useful types of relations are highly dependent on task!

Different schemas require different extractions

In April 2011 , Prime Minister Mykola Azarov of Ukraine met with the President of Brazil , Dilma Rousseff , in Sanya .

ACE:

- (Ukraine, *employee-executive*, Mykola Azarov)
- (Brazil, *employee-executive*, Dilma Rousseff)

TAC, for query Dilma Rousseff:

- (*per:employee of*, Brazil)
- (*per:title*, President)
- (*per:country of birth*, Brazil)

Relation schema: ACE

relation type	subtypes
physical	located, near, part-whole
personal-social	business, family, other
employment / membership / subsidiary	employ-executive, employ-staff, employ-undetermined, member-of-group, partner, subsidiary, other
agent-artifact	user-or-owner, inventor-or-manufacturer, other
person-org affiliation	ethnic, ideology, other
GPE affliation	citizen-or-resident, based-in, other
discourse	-

Relation schema: TAC KBP (per)

per:alternate names	per:date of birth
per:age	per:country of birth
per:state of birth	per:city of birth
per:origin	per:date of death
per:country of death	per:state of death
per:city of death	per:cause of death
per:countries of residence	per:states of residence
per:cities of residence	per:schools attended
per:title	per:member of
per:employee of	per:religion
per:spouse	per:children
per:parents	per:siblings
per:other family	per:charges

Relation schema: TAC KBP (org)

org:alternate names	org:political religious affiliation
org:top members employees	org:number of employees
org:members	org:member of
org:subsidiaries	org:parents
org:founded by	org:date founded
org:date dissolved	org:country of headquarters
org:state of headquarters	org:city of headquarters
org:shareholders	org:website

How to extract relations

1. Hand-coded rules
2. Supervised
3. Semi-supervised
    - Bootstrapping
    - Distant supervision
4. Unsupervised

Relations are often dependent on NE types. . .

- *location-of-birth*, PER-LOC
- *location-of-headquarters*, ORG-LOC
- *employee-of*, PER-ORG (or GPE)

. . . but this is not precise enough

For:

- *employee-of*, PER-ORG (or GPE)
- `Barack Obama, US`

Correct and incorrect extractions:

- `Barack Obama is an employee of the US.`
- `Barack Obama is the president of the US.`
- `*Barack Obama was born in the US.`
- `*Barack Obama returned to the US.`

Hand-coded patterns as a starting point

- `Barack Obama is an employee of the US.`
- *PER is an employee of the GPE*

- `Barack Obama is the president of the US.`
- *PER is (an|the) (employee|president) of the GPE*

- `US president Barack Obama...`
- *GPE president PER*
  PER is (an|the) (employee|president) of the GPE

Benefits, limitations of hand-coded patterns

Benefits:

1. High-precision
2. Interpretable
3. Can be a fast start on a new domain or task

Limitations:

1. VERY low recall
2. Huge amount of work to scale to many relations
3. Not perfect precision anyway
   `Barack Obama is the president of the US in the new hit TV political drama.`

Relation classification

Given a:

1. Set of relation types
2. Collection of text (sentences or documents)

Preprocess:

1. Label entities (NER)
2. Manually annotate relations
3. Split data into train, dev and test

Train:

1. Extract features for every pair of entities
2. Train binary+n-way classifier, or $n + 1$-way classifier

What features are used?

`In April 2011 , Prime Minister Mykola Azarov of Ukraine met with the President of Brazil , Dilma Rousseff , in Sanya .`

- **Entity headwords:** `E1:Azarov E2:Ukraine`
- **Entity BOW:** `E1:Mykola E1:Azarov Ukraine`
- **Entity context:** `E1-1:Minister E1+1:of E2-1:of E2+1:of`
- **Entity types:** `E1:PER E2:GPE PER-GPE`
- **BOW between:** `of`
- **Dependency path:** E1←of→E2
- Other parse features, gazetteers, word clusters and embeddings

Classification

- Best classifier depends on task: Naive Bayes, SVM, MaxEnt
- Precision/Recall/F-score evaluation
- Good performance with enough training data
- Brittle, domain specific, training data is still expensive and doesn't scale

Supervised Relation Extraction

- Document-level information extraction
- Supervised approaches do not achieve high performance on ACE
  - 45.8 F-score for SVM with subsequence kernel (Bunescu and Mooney NIPS05)
  - 52.8 F-score for SVM with dependency tree kernel (Bunescu and Mooney EMNLP05)
- Very small data/schema sets
  $\Longrightarrow$ **very limited coverage**

Bootstrapping relation extractors

- Task: given a small number of seeds for a given relation type, bootstrap a wide-coverage extractor
- E.g., authors:

Author	Book
Isaac Asimov	The Robots of Dawn
David Brin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickens	Great Expectations
William Shakespeare	The Comedy of Errors

- E.g., headquarters:

Organisation	Location
Microsoft	Redmond
Exxon	Irving
IBM	Armonk
Boeing	Seattle
Intel	Santa Clara

Bootstrapping process

Brin (1999). **Extracting patterns and relations from the world wide web.** In *Proceedings of the International Workshop on the World Wide Web and Databases*.

1. Initialise R with seeds
2. $O \leftarrow \mathit{FindOccurrences}(R, D)$
   find instances of relation pairs from R that occur together in D
3. $P \leftarrow \mathit{GeneratePatterns}(O)$
   generate extraction patterns $\langle \mathit{author}, \mathit{title}, \mathit{order}, \mathit{prefix}, \mathit{middle}, \mathit{suffix} \rangle$
4. $R \leftarrow \mathit{ExtractRelations}(P, D)$
   extract relations from D using the new patterns P
5. If R is large enough, return. Else go to step 2.

Agichtein and Gravano (2000)

Agichtein and Gravano (2000). **Snowball: extracting relations from large plain-text collections**. In *Proceedings of the 5th ACM Conference on Digital Libraries*.
http://www.mathcs.emory.edu/~eugene/papers/dl00.pdf

- Perform NER on documents in D
- E.g.: `The Irving-based Exxon Corporation`
  < {the}, *location*, {- based}, *organisation*, {}, >
- Generalise patterns by clustering
- Select patterns that are productive and reliable
  < {}, *location*, {- based}, *organisation*, {} >
  < {}, *company*, {'s headquarters in}, *location*, {} >

Performance: 90.0 precision (82.5 recall)
Precision can be improved by setting stricter thresholds
Still domain-specific/user-driven

Never-Ending Language Learning

Carlson el al (2010). **Toward an Architecture for Never-Ending Language Learning.**. In *Proceedings of AAAI 2010*
http://rtw.ml.cmu.edu/papers/carlson-aaai10.pdf

- Bootstrapped lexical and POS tag patterns.
- Heavy constraints: mutual exclusion; fine-grained type constraints; semi-structured features.
- Bootstrapping over ClueWeb09: non-stop since Jan 2010, 1 iteration/day.
- Precision is high ($> 90\%$), recall is 2 million high-confidence assertions from ClueWeb09.
- Uses human reinforcement (rtw.ml.cmu.edu)

Distant supervision

Mintz el al (2009). **Distant supervision for relation extraction without labeled data**. In *Proceedings of the 47th ACL*, among several other works
`http://web.stanford.edu/~jurafsky/mintz.pdf`

- Combine semi-supervised and supervised
- Instead of a small number of seeds, use a huge KB
- Align examples to text and use these as training data

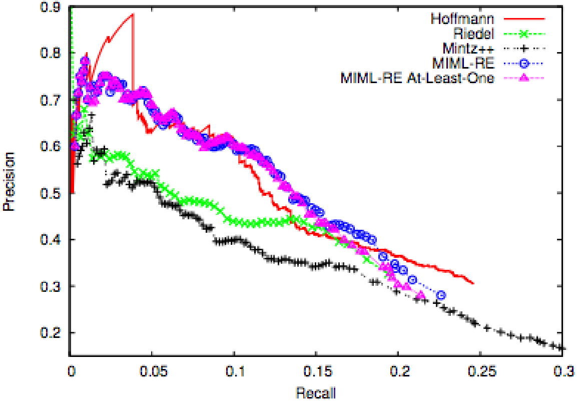
Distant supervision algorithm

1. For each relation type: `employee of`
2. For each tuple in KB: `Tim Cook, Apple`
3. Align to sentences in corpus that contain both entities:
   `Tim Cook is an American business executive, and is the Chief Executive Officer of Apple Inc.`
4. Use these instances as training data

- Large body of work in RE in recent years has focussed on improving how distant supervision is modelled

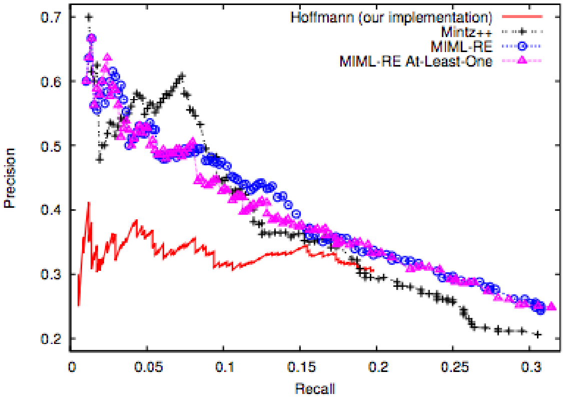
RE performance

Figure from Surdeanu el al (2012). **Multi-instance Multi-label Learning for Relation Extraction.**. In *Proceedings of EMNLP 2012*
http://nlp.stanford.edu/pubs/emnlp2012-mimlre.pdf



RE performance on KBP

Figure from Surdeanu el al (2012). **Multi-instance Multi-label Learning for Relation Extraction.**. In *Proceedings of EMNLP 2012* `http://nlp.stanford.edu/pubs/emnlp2012-mimlre.pdf`



RE remains a difficult problem

- Major pipelined error (NER + disambiguation + parsing + coreference)
- Training data is still very limited
- Differences between correct and incorrect extractions can be very subtle, and schemas are brittle
    - PER, president of ORG
    - PER, vice president of ORG
    - PER, executive vice president of ORG
    - PER, former president of ORG
    - PER, newest president of ORG
    - PER, most successful president of ORG
    - PER, wealthiest vice president of ORG
    - PER, not president of ORG

RE remains an even more difficult problem

- Complex inference
  `Simmons' father, Feri Witz, also Hungarian-born, remained in Israel, where he had one other son and three daughters.`
  (`Simmons`, *resided-in*, `Israel`)
  `https://en.wikipedia.org/wiki/Gene_Simmons`article
- Complex discourse
  ``` ``Who's that?'' Negroponte asked... ```
  ``` A young woman peeked into the living room. ``Where's George?'' asked Alejandra, 23. George, 17, appeared. Then Sophia, 13, and John, 19. Four of the five Negroponte children were at home. ```
  (`John Negroponte`, *parent-of*, `Sophia`)
  `https://www.washingtonpost.com/archive/politics/2007/01/29/for-negroponte-move-to-state-dept-is-a-homecoming/f7a692fb-a6b7-4bcf-bf60-60725e6be8d0/`article

Evaluation of OpenIE and related techniques

- Systems with large numbers of extractions are costly to fully annotate
- No measure of recall!
- Measure precision by top-k manual precision
- e.g. for the top-1000 most confidence instances, manually annotate, measure precision
- Tying OpenIE relations to a defined schema remains an open problem: rules still perform relatively well

Scenario Templates (events) in MUC-7 (NIST 1997)

http://www-nlpir.nist.gov/related_projects/muc/proceedings/walkthru_ie_text.html **(document)**
http://www-nlpir.nist.gov/related_projects/muc/proceedings/walkthru_st_key.html **(events)**

Output:

```
                                       VEHICLE_INFO:       <VEHICLE_INFO-9602140509-1>
                                       PAYLOAD_INFO:       <PAYLOAD_INFO-9602140509-1>
                                       LAUNCH_DATE:        <TIME-9602140509-1>
                                                           (15021996 local time)
<LAUNCH_EVENT-9602140509-1>      :=    LAUNCH_SITE:        <LOCATION-9602140509-1>
                                                           ('Xichang', 'China')
                                       MISSION_TYPE:       CIVILIAN
                                       MISSION_FUNCTION:   DEPLOY
                                       MISSION_STATUS:     FAILED
```

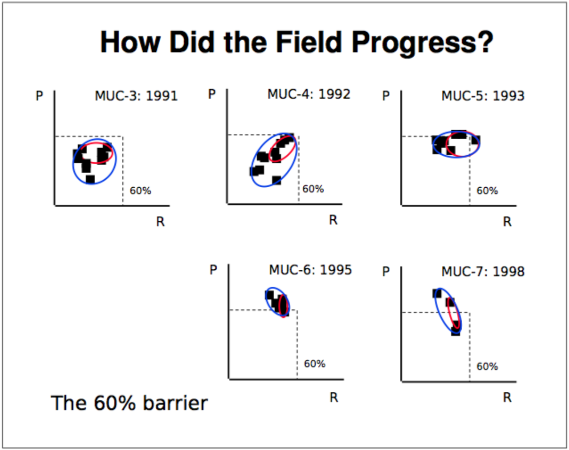
where:

```
<VEHICLE_INFO-9602140509-1> :=
VEHICLE:          <ENTITY-9602140509-34>
                  ('Long March 3B')
VEHICLE_TYPE:     ROCKET
VEHICLE_OWNER:    <ENTITY-9602140509-6>
                  ('Great Wall Industry Corp.')
```

```
<PAYLOAD_INFO-9602140509-1> :=
PAYLOAD:          <ENTITY-9602140509-35>
                  ('satellite built by Loral Corp.')
PAYLOAD_TYPE:     SATELLITE
PAYLOAD_FUNC:     TV
PAYLOAD_OWNER:    <ENTITY-9602140509-3>
                  ('Intelsat')
```

Supervised Scenario Template Extraction

Hobbs and Riloff (2010). **Information Extraction.** In: *Handbook of Natural Language Processing, 2nd edition*.



How Did the Field Progress?

The 60% barrier

Scenario Template Extraction is Hard

Hobbs and Riloff (2010). **Information Extraction.** In: *Handbook of Natural Language Processing, 2nd edition*.

- Biggest source of mistakes in entity and event coreference
  $\Longrightarrow$ coreference needs to improve!
- Only 60% of events are expressed in explicit language?
  $\Longrightarrow$ >60% requires inference and access to world knowledge!
- Long tail of extraction patterns?
  $\Longrightarrow$ active learning to identify difficult examples in unlabelled data!
- State-of-the-art NER performance around 91%
  Events typically require 4 **entities:** $0.91^4 \approx 0.69$
  $\Longrightarrow$ improved NER, joint extraction of entities and events?

Take away

- What is information extraction?
- Divided into numerous sub-tasks
- What is coreference resolution? relation extraction?
- Each one needs a schema to define it and annotation to evaluate it
- Solved with variants of the things we've seen befofre
- E.g. entity disambiguation as retrieval
- E.g. relation extraction as pair classification
- Sometimes good performance comes with innovative construction of training data