Document Level Models (1)

Coreference Resolution

Yoav Goldberg
(with additions by Ido Dagan)
Bar Ilan University

Credits for slides by Mihai Surdenau and Marta Recasens

Beyond sentences

Until now

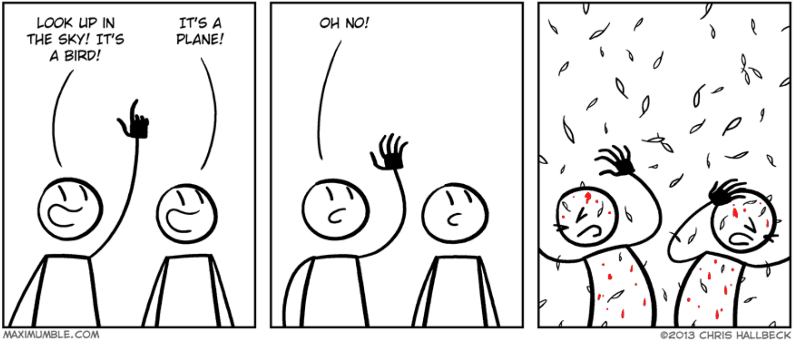
- Mostly low-level components (building blocks)
- Working at sentence level – analyzing each sentence individually

Today (++)

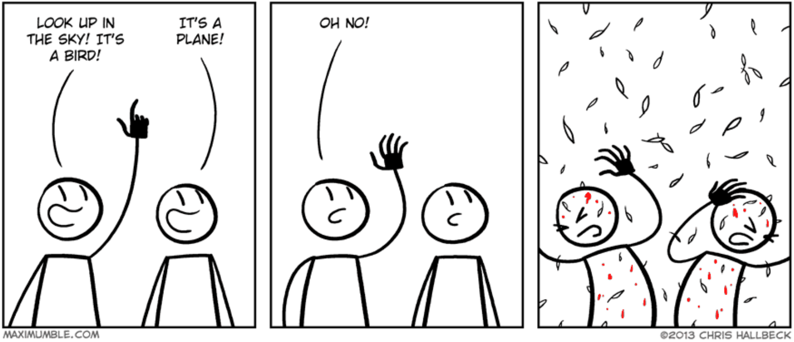
- Looking at the document and the corpus level.
- Still focusing on building-blocks.

Coreference Resolution

Coreference Resolution



Coreference Resolution



(Which entities are the same?)

Coreference Resolution Revisited

Mihai Surdeanu
msurdeanu@email.arizona.edu
www.surdeanu.info/mihai

(the Stanford coreference system)

April 19th, 2013

What is it?

Coreference resolution = the task of clustering together of expressions that refer to the same entity/concept.

Michelle LaVaughn Robinson Obama is an American lawyer and writer. She is the wife of the 44th and current President of the United States, Barack Obama, and the first African-American First Lady of the United States.

Why is it important?

- Question answering
  - “Who is Barack Obama’s spouse?”
- Information extraction
  - “Find all per:spouse relations between all named entities in this large corpus.”
- News aggregation
  - “What are recent events involving Michelle Obama?”
  - *Requires cross-document coreference resolution. More on this soon.*

Why is it important?

- Performance doubles for these applications when coreference resolution is used.
- See: R. Gabbard, M. Freedman, and R. Weischedel, "Coreference for Learning to Extract Relations: Yes, Virginia, Coreference Matters," ACL 2011.

Extensions:

event (predicate) coreference, cross document coreference

Real-world example

Chengdu, China (CNN) – The researcher dressed in blue plastic smock, slippers and gloves is having a tough time getting his work done.
Every time Zhang Zhen sets up his camera on a tripod in an effort to document the behavior of one of [[the panda cubs] scattered on [a grassy hillside]], one particularly frisky baby panda comes wobbling towards him, interrupting his shoot.
“Mumu!” he yells in frustration, as the four-month old cub rears up on [her] hind legs, lunging towards him. He picks Mumu up and deposits her at [the opposite end] of [the enclosure].
“I’m not sure why she’s been all over me like this. I think she’s excited today,” Zhang says.
Mumu is the oldest of fourteen baby pandas that were born last summer here at the Research Base of [[Giant Panda] Breeding] in [Chengdu, China].
`CNN.com, Dec 2013`

Stages towards understanding

1. (Pre-processing – sentence boundary, tagging, parsing, ...)
2. Entity Extraction (broader sense of entity, not just named)
3. Coreference Resolution
4. Entity Linking (link to an ontology / database / wikipedia)

Stages towards understanding

1. (Pre-processing – sentence boundary, tagging, parsing, . . .)
2. **Entity Extraction**
3. Coreference Resolution
4. Entity Linking (link to an ontology / database / wikipedia)

Chengdu, China (CNN) – The researcher dressed in blue plastic smock, slippers and gloves is having a tough time getting his work done.
Every time Zhang Zhen sets up his camera on a tripod in an effort to document the behavior of one of [[the panda cubs] scattered on [a grassy hillside]], one particularly frisky baby panda comes wobbling towards him, interrupting his shoot.
“Mumu!” he yells in frustration, as the four-month old cub rears up on [her] hind legs, lunging towards him. He picks Mumu up and deposits her at [the opposite end] of [the enclosure].
“I’m not sure why she’s been all over me like this. I think she’s excited today,” Zhang says.
Mumu is the oldest of fourteen baby pandas that were born last summer here at the Research Base of [[Giant Panda] Breeding] in [Chengdu, China].
`CNN.com, Dec 2013`

Chengdu, China (CNN) – The researcher dressed in blue plastic smock, slippers and gloves is having a tough time getting his work done.

Every time Zhang Zhen sets up his camera on a tripod in an effort to document the behavior of one of [[the panda cubs] scattered on [a grassy hillside]], one particularly frisky baby panda comes wobbling towards him, interrupting his shoot.

"Mumu!" he yells in frustration, as the four-month old cub rears up on [her] hind legs, lunging towards him. He picks Mumu up and deposits her at [the opposite end] of [the enclosure].

"I'm not sure why she's been all over me like this. I think she's excited today," Zhang says.

Mumu is the oldest of fourteen baby pandas that were born last summer here at the Research Base of [[Giant Panda] Breeding] in [Chengdu, China].

`CNN.com, Dec 2013`

- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

Stages towards understanding

1. (Pre-processing – sentence boundary, tagging, parsing, . . .)
2. **Entity Extraction**
3. Coreference Resolution
4. Entity Linking (link to an ontology / database / wikipedia)

Stages towards understanding

1. (Pre-processing – sentence boundary, tagging, parsing, . . .)
2. Entity Extraction
3. **Coreference Resolution**
4. Entity Linking (link to an ontology / database / wikipedia)

- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

Coreferene is a clustering task

- Decide on number of clusters
- Assign entity mentions to clusters

- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

Evaluation

How to evaluate?

Evaluation

How to evaluate?

- Types of mistakes:
  - Splitting a cluster
  - Merging of cluster
  - Incorrect assignment

Evaluation

How to evaluate?

- Types of mistakes:
  - Splitting a cluster
  - Merging of cluster
  - Incorrect assignment
- Which are more important?
- How do we design a metric to capture these?

- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

Evaluation

The B^3 F-Score Metric

$$P = \frac{1}{|docs|} \sum_{doc \in docs} \sum_{m \in doc} \frac{|g_m \cap p_m|}{|p_m|}$$

$$R = \frac{1}{|docs|} \sum_{doc \in docs} \sum_{m \in doc} \frac{|g_m \cap p_m|}{|g_m|}$$

$$F_1 = 2\frac{P\dot{R}}{P+R}$$

g_m gold cluster containing m

p_m predicted cluster containing m

$|x|$ size of cluster x

|docs| total number of mentions in all docs

Evaluation

The B^3 F-Score Metric

$$P = \frac{1}{|docs|} \sum_{doc \in docs} \sum_{m \in doc} \frac{|g_m \cap p_m|}{|p_m|}$$

$$R = \frac{1}{|docs|} \sum_{doc \in docs} \sum_{m \in doc} \frac{|g_m \cap p_m|}{|g_m|}$$

$$F_1 = 2\frac{P\dot{R}}{P + R}$$

g_m gold cluster containing m

p_m predicted cluster containing m

$|x|$ size of cluster x

Eval is open for debate

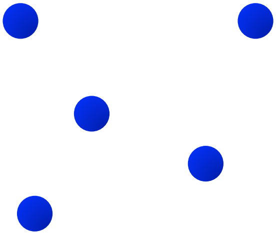
- B^3 is good, but not perfect.
- Other variants exist.

How to solve?

- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

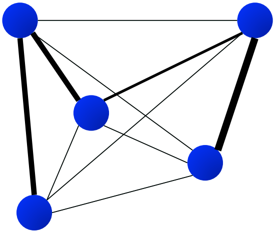
- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

A typical algorithm



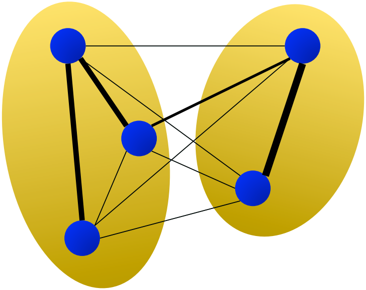
Identify all mentions

A typical algorithm



Compute link scores between all pairs

A typical algorithm



Partition this graph into entity clusters

Pairwise Approaches

- How to score each pair?
  - Classifier.
  - But how to train? (what are the examples? what are the features?)
- How to choose best partition given pair scores?

Choosing a partition

- Choosing a globally optimal clustering under reasonable objectives is NP-hard.
- Resort to heuristics or approximations.

Choosing a partition

- Choosing a globally optimal clustering under reasonable objectives is NP-hard.
- Resort to heuristics or approximations.

A possible algorithm 1

- For each mention in order of appearance
  - Compute scores to previous mentions
  - Decide if starting a new cluster or linking to existing cluster

Choosing a partition

- Choosing a globally optimal clustering under reasonable objectives is NP-hard.
- Resort to heuristics or approximations.

A possible algorithm 1

- For each mention in order of appearance
  - Compute scores to previous mentions
  - Decide if starting a new cluster or linking to existing cluster

A possible algorithm 2

- Assume that each mention has at most one antecedent
- Mentions with 0 antecedents start a new cluster
- Now, search for trees instead of clusters
  - (trees are easy..)

A possible algorithm 1

- For each mention in order of appearance
  - Compute scores to previous mentions
  - Decide if starting a new cluster or linking to existing cluster

Discuss features.

Discuss training examples.

Read!! - sections 2.2.1 & 3:
http://u.cs.biu.ac.il/~89-680/coref-features.pdf

Discuss potential problems with the pairwise approach.

- Chengdu, China
- CNN
- The researcher
- blue plastic smock, slippers and gloves
- though time
- [his] work
- Zhang Zhen
- [his] camera
- a tripod
- an effort
- the behavior
- one of the panda cubs scattered on a grassy hillside
- the panda cubs scattered on . . .
- a grassy hillside
- One particularly frisky baby panda
- him
- [his] shoot
- Mumu
- he
- the four-month old cub
- [her] hind legs
- him
- He
- Mumu
- her
- the opposite end of [the enclosure]
- I
- she
- me
- I
- she
- today
- Zhang
- Mumu
- fourteen baby pandas
- last summer
- here
- the Research Base of [[Giant Panda] Breeding] in [Changdu, China]

Some insights learned

- Most algorithms focus on step 2: computing mention-pair scores using machine learning, which is a *local* operation
  - Poor representation of context: only two mentions considered
- Recent work showed that it is important to address coreference resolution as a *global* task, where all mentions are modeled jointly
  - This is hard to model using machine learning
- ML models generalize poorly to new words, domains, and languages
  - Annotating coreference is expensive

Idea I

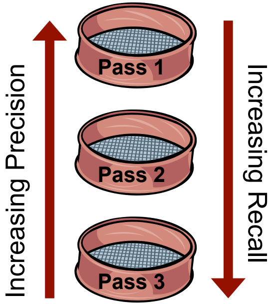
- ✖ ~~machine learning~~
- ✔ deterministic, rule-based model
- ✔ “baby steps” approach
- ✔ global model

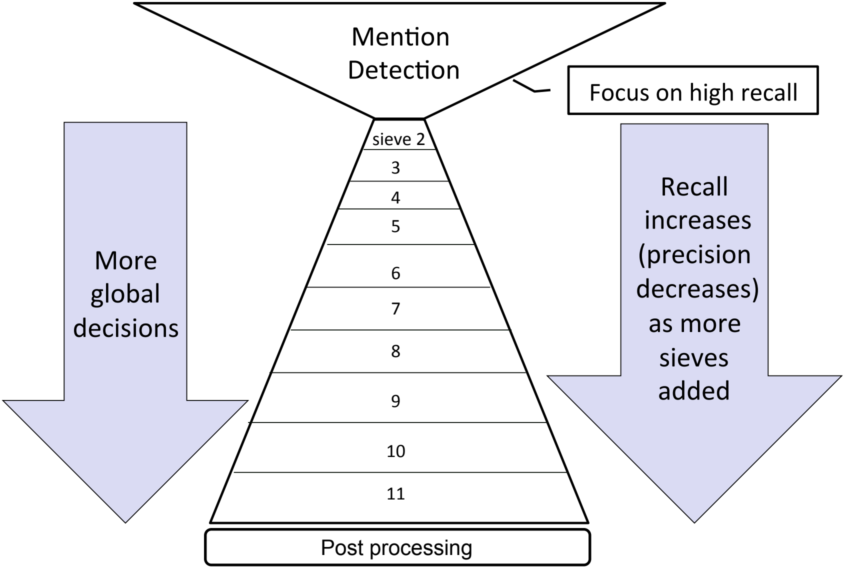
Entity coreference resolution model

- Novel architecture for coreference resolution:
  - **"Baby steps"** – accurate things first
  - **Global** – attribute sharing in clusters
  - **Deterministic** – rule-based model
- Top ranked system at CoNLL-2011 Shared Task:
  - 58.3% (open), 57.8% (closed)

Baby-steps approach

- Multiple passes (or “sieves”) over text
- Precision of each pass is smaller than preceding ones
- Recall keeps increasing with each pass
- Decisions once made cannot be modified by later passes
- Modular architecture




Mention
Detection
Focus on high recall
sieve 2
3
4
5
6
7
8
9
10
11
Post processing
More
global
decisions
Recall
increases
(precision
decreases)
as more
sieves
added

Why multiple sieves? (a new pass for each sieve)

```
number: plural
animacy: animate
```

```
number: plural
animacy: unknown
```

The second attack occurred after some rocket firings aimed, apparently, toward **the israelis**, apparently in retaliation. **we**’re checking our facts on that one. ...
the strike will undermine efforts by palestinian authorities to bring an end to terrorist attacks and does not contribute to the security of **israel**.

```
number: singular
animacy: inanimate
```

Why multiple sieves?

number: plural
animacy: animate

number: plural
animacy: unknown

The second attack occurred after some rocket firings aimed, apparently, toward **the israelis**, apparently in retaliation. **we**'re checking our facts on that one. ...
the strike will undermine efforts by palestinian authorities to bring an end to terrorist attacks and does not contribute to the security of **israel**.

number: singular
animacy: inanimate

Why multiple sieves?

number: plural
animacy: animate

number: plural
animacy: unknown

The second attack occurred after some rocket firings aimed, apparently, toward **the israelis**, apparently in retaliation. **we**’re checking our facts on that one. ...
the strike will undermine efforts by palestinian authorities to bring an end to terrorist attacks and does not contribute to the security of **israel**.

number: singular
animacy: inanimate

Why multiple sieves?

number: plural
animacy: animate

number: plural, singular
animacy: inanimate

The second attack occurred after some rocket firings aimed, apparently, toward **the israelis**, apparently in retaliation. **we**'re checking our facts on that one. ...
the strike will undermine efforts by palestinian authorities to bring an end to terrorist attacks and does not contribute to the security of **israel**.

number: plural, singular
animacy: inanimate

Pass 1 – Mention detection

- Extract all noun phrases (NP) plus pronouns and named entities even in modifier position

- Remove non-referring expressions, e.g., generic “it”, with manually written patterns
  - *E.g., It is possible that...*

Pass 2 – Speaker identification

- Extract speakers and use the info for resolution
  - *"....", she said.*

- Positive and negative constraints for following sieves:

  "I voted for Nader because he was most aligned with my values," she said.

Pass 3 – Exact string match

Exactly the same text:

…**TWA** 's bid for **USAir** skeptically , seeing it as a ploy to pressure **USAir** into buying **TWA.**

The Shahab 3 ground-ground missile: the new addition to Iran's military capabilities … developed **the Shahab 3 ground-ground missile** for defense purposes with capabilities ranging from …

Pass 4 – Relaxed string match

String match after dropping the text following the head word:

…**Clinton**… **Clinton, whose term ends in January**…

Pass 5 – Precise constructs

Appositives:

… but **Bob Gerson**, **video editor of This Week in Consumer Electronics**, says Sony conceives …

Predicate nominatives:

Started three years ago, **Chemical's interest-rate options group** was **a leading force** in the field.

Role appositives:

… **[[actress] Rebecca Schaeffer]** …
… **[[painter] Pablo Picasso]** …

Pass 5 – Precise constructs

Relative pronouns:

… **[the finance street [which] has already formed in the Waitan district]** …

Acronyms:

Agence France Presse … **AFP**

Demonyms/Gentilics:

Israel… **Israeli**

Passes 6 – 9: Strict head match

The Japanese company already has 12% of the total camcorder market, ranking it third behind the RCA and Panasonic brands … The company also plans to aggressively start marketing … The electronics ~~company~~…

- Coupled with various constraints:
  - No new information in mentions to be resolved
  - No location mismatch, “Lebanon” != “southern Lebanon”
  - No numeric mismatch, “people” != “around 200 people”
  - No i-within-i, e.g., [[Sony Corporation] of America]

Pass 10 – Relaxed head match

- Same constraints as above but anaphora head can match any word in the candidate cluster

"Sanders"

is compatible with the cluster:

{Sauls, the judge, Circuit Judge N. Sanders Sauls}

Pass 11 – Pronoun resolution

- Attributes must agree
  - Number
  - Gender
  - Person
  - Animacy
- Assigned using POS tags, NER labels, static list of assignments for pronouns
- Improved further using gender and animacy dictionaries of Bergsma and Lin (2006), and Ji and Lin (2009)

Post processing

- Discard singleton clusters
  - This is why we could maximize recall in mention detection!
- Discard shorter mentions in appositive patterns
- Discard mentions that appear later in copulative relations

- Implemented to comply with OntoNotes annotations

A run-through example

John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.

A run-through example

John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.

Mention detection

A run-through example

John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.

Speaker identification

A run-through example

John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.

String match

A run-through example

John is a musician. He played a new song. A girl was listening to the song. "It is my favorite," John said to her.

Precise constructs

A run-through example

John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.

Strict head match

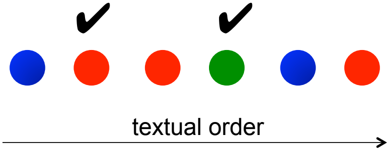
A run-through example

John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.

Pronoun resolution

Mention selection in a given sieve

- In each sieve, we consider for resolution only mentions that are currently first in textual order in their cluster. (in order to decide whether to merge it with an antecedent)
- Most informative!



Features are shared within clusters

- Within a cluster:
  - Union of all modifiers
  - Union of all head words
  - Union of all attributes: number, gender, animacy
- Robustness to missing/incorrect attributes

"a group of students"

`number: singular`

"five students"

`number: plural`

`number: singular, plural`

EXPERIMENTS

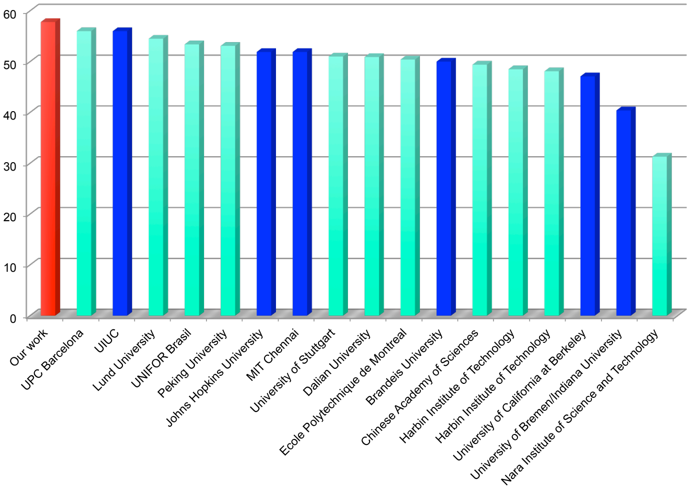
Results on older corpora

UNSUPERVISED	ACE 2004 Test	ACE NWIRE	MUC6
This work	**81**	**80.2**	74.4
Haghighi and Klein (2009)	79.0	76.9	**75.0**

SUPERVISED	Ace 2004 Test	ACE NWIRE	MUC6
Culotta et al. (2007)	79.3	-	-
Bengston and Roth (2008)	80.8	-	-
Finkel and Manning (2008) +G	-	74.5	64.3

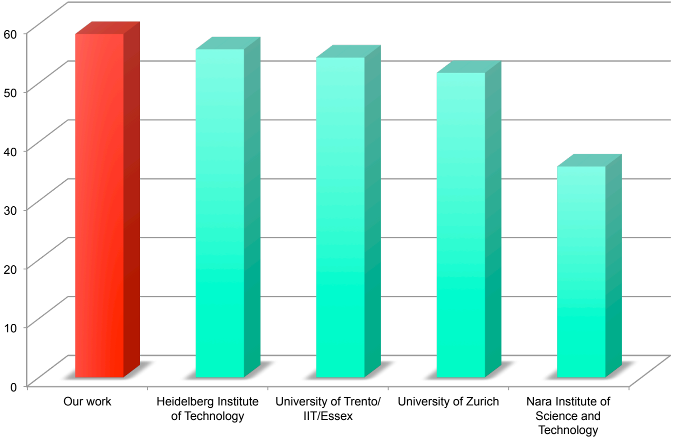
B^3 F1 scores of different systems on standard corpora

Results: CoNLL-2011closed track



CoNLL score = (MUC F1 + B^3 F1 + CEAF F1) / 3

Results: CoNLL-2011 open track



CoNLL score = (MUC F1 + B^3 F1 + CEAF F1) / 3

CoNLL-2012 shared task

- Multilingual unrestricted coreference resolution in OntoNotes
  - English, Chinese, Arabic
- Higher barrier of entry
  - 16 submissions vs. 23 submissions in 2011
- But there was significant progress
  - Best score for English increased from 58.3 to 63.4

CoNLL-2012 shared task

- Two out of the top three systems used our system
- Fernandes et al., PUC/IBM Brazil
  - Adapted our system to Chinese and Arabic
  - Reranked the output of our system
  - Best system overall
- Chen and Ng, UT Dallas
  - Adapted our system to Chinese and Arabic
  - Added two ML-based sieves to our system
  - Best for Chinese, top 3 overall
- Proof that our approach is multilingual

Analysis: Importance of sharing features

Entity-centric model	**59.3**
Mention-pair model	55.9

CoNLL F1 in OntoNotes Dev

Analysis: Importance of multiple sieves

Multi-pass model	**59.3**
Single-pass model	53

CoNLL F1 in OntoNotes Dev

Analysis: Importance of features

Complete	59.3	
wo/ Number	56.7	*- 2.6*
wo/ Gender	58.9	*- 0.4*
wo/ Animacy	58.3	*- 1.0*
wo/ NE	58.8	*- 0.5*

CoNLL F1 in OntoNotes Dev

Idea I: Conclusions

- Novel architecture for coreference resolution
  - “Baby steps”
  - Global
  - Deterministic
- State of the art results (in multiple languages)
  - Best at CoNLL-2011
  - Two of the top 3 systems at CoNLL-2012 used it

Big-picture conclusions

- Understanding the problem is more important than machine learning
- Model things jointly when you can

Recent Improvements

Taking Coreference Resolution beyond the 60% Performance Barrier

Marta Recasens
Google Research

(Joint work with Matthew Can, Marie-Catherine de Marneffe, Chris Potts, Dan Jurafsky, Eduard Hovy, and M. Antònia Martí)

April 26, 2013 · Carnegie Mellon University

* Same Referent, Different Words: Unsupervised Mining of Opaque Coreferent Mentions. NAACL 2013.

Life and Death of DEs

Nestle USA issued a voluntary recall of its Nesquik chocolate powder after being tipped off by an ingredient supplier of possible salmonella contamination.

The Glendale-based company said it was calling back canisters of the product, which is mixed with milk to create a sweet drink, that were made in October and sold nationwide.

Consumers should look for containers bearing an expiration date of October 2014.

Nestle decided to recall the power

Life and Death of DEs

Nestle USA issued a voluntary recall of its Nesquik chocolate powder after being tipped off by an ingredient supplier of possible salmonella contamination.

The Glendale-based company said it was calling back canisters of the product, which is mixed with milk to create a sweet drink, that were made in October and sold nationwide.

Consumers should look for containers bearing an expiration date of October 2014.

Nestle decided to recall the powder

Life and Death of DEs

Nestle USA issued a voluntary recall of its Nesquik chocolate powder after being tipped off by **an ingredient supplier** of possible salmonella contamination.

The Glendale-based company said it was calling back **canisters** of the product, which is mixed with **milk** to create **a sweet drink**, that were made in **October** and sold nationwide.

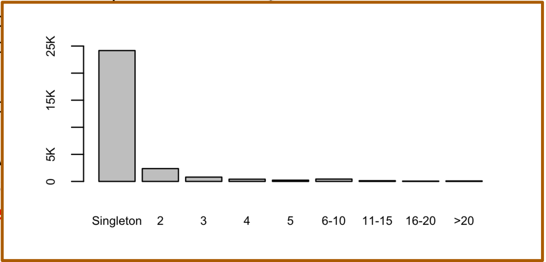
Consumers should look for **containers bearing an expiration date of October 2014**.

Nestle decided to recall the powder

Life and Death of DEs

Nestle USA issued a voluntary recall of its Nesquik chocolate powder after being tipp
supplier c
contamin

The Glend
was callin
product, v
create a s
in Octobe



Consume
bearing an expiration date of October 2014.

Nestle decided to recall the powder

Singletons

- Singleton mentions are hard and common.
- Design a classifier specifically for predicting them.
- Use a different set of linguistically motivated features.

Then, predicted singletons are filtered and not considered by the coreference algorithm, reducing its errors.

What’s hard? Why only $\sim$60% accuracy?

The unsolved problem of coreference resolution

The flaw was first reported by a security researcher David Emery, who posted his findings to the Cryptome mailing list. [...] **The bug** has not been corrected by any subsequent updates .

The software is used to turn 2D photos into 3D models; in reality, a person uploads photos taken or stored on an iPad to the Autodesk Cloud, where the actual conversion happens. [...] **The app** is free, but requires an iPad 2 or better running iOS 5.x.

The unsolved problem of coreference resolution

Autodesk's had its 123D Catch **iPad** application in the works for quite some time now, but starting today, you'll finally be able to use **that Cupertino slate** to turn those beautiful snaps into three-dee creations.

Now you can keep up with all of the people you follow with a "best-of" weekly email from **Twitter**. [...]
The micro-blogging service will now be sending out weekly email digests that will feature a summary of your Twitter stream.

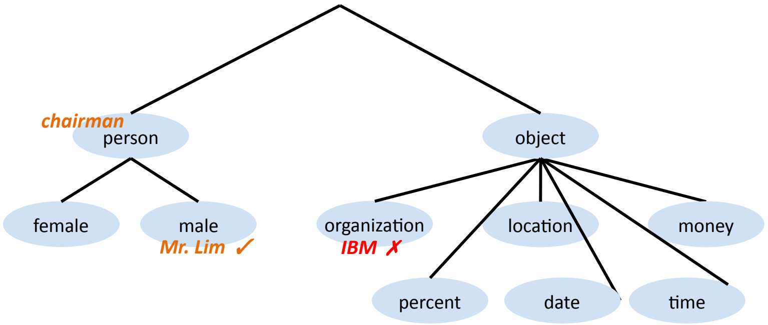
Features

(Soon et al. 2001, Ng & Cardie 2002)

- Surface features
  - String/head match
  - Sentence/token distance
- Morphological features
  - Mention is a pronoun/definite/demonstrative/proper noun
- Syntactic features
  - Gender/number agreement
  - Grammatical role
- Semantic features
  - NE type
  - WordNet
  - Wikipedia
  - Others: Yago, lexico-semantic patterns, etc.

WordNet

Semantic class match
(Soon et al. 01)



WordNet

WordNet paths
(Harabagiu et al. 01, Ng & Cardie 02, Poesio et al. 04, Ponzetto & Strube 06)

IN-GLOSS

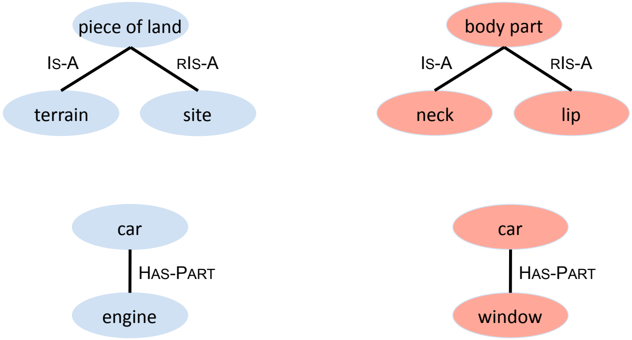
S: (n) manufacturer, **maker**, manufacturing business (a business engaged in manufacturing some product)

SYNONYM

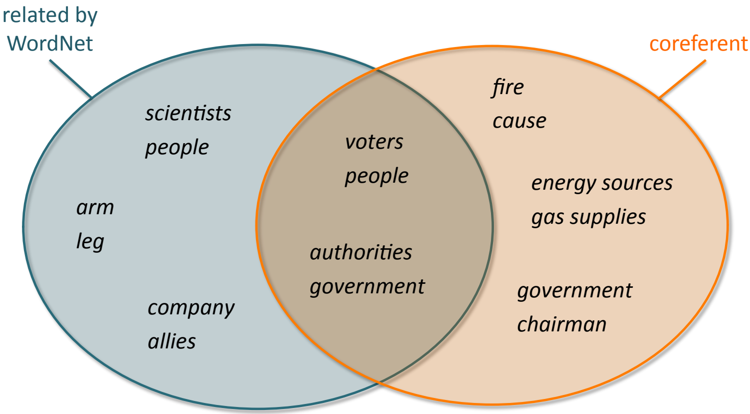
S: (n) manufacturer, **maker**, manufacturing business (a business engaged in manufacturing some product)

WordNet

WordNet paths
(Harabagiu et al. 01, Ng & Cardie 02, Poesio et al. 04, Ponzetto & Strube 06)



Semantic similarity is not coreference



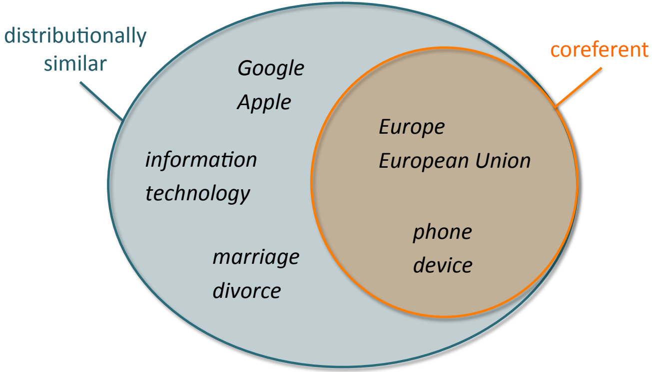
Distributional similarity

Distributional hypothesis (Harris 1954): words that occur in the same contexts tend to have similar meanings.

	aardvark	computer	data	pinch	result	sugar	…
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	

	subj-of, absorb	subj-of, adapt	subj-of, behave	…	pobj-of, inside	pobj-of, into	…	nmod-of, abnormality	nmod-of, anemia	nmod-of, architecture	…	obj-of, attack	obj-of, call	obj-of, come from	obj-of, decorate	…	nmod, bacteria	nmod, body	nmod, bone marrow
cell	1	1	1		16	30		3	8	1		6	11	3	2		3	2	2

Distributional similarity is still not coreference

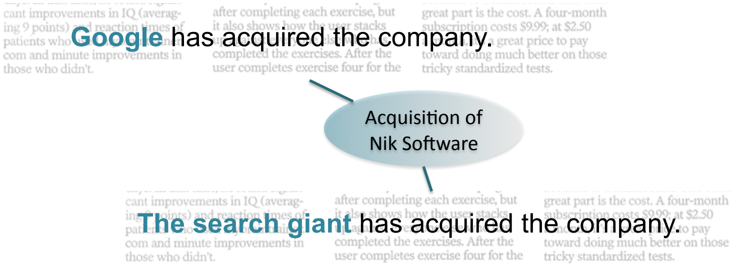


Intuition of our solution:

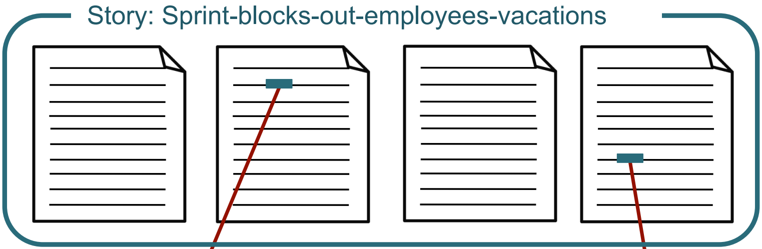
Microsoft has released a new feature.

The search giant has released a new feature.

Intuition of our solution: Restricted distributional similarity



Restricted distributional semantics



block out
release
launch
prevent
…

Sprint blocks out vacation days for a major phone announcement.

According to SprintFeed, **the carrier** is blocking out vacation days for employees.

Comparable corpus

Techmeme (www.techmeme.com)

Kunur Patel / AdAge:

Zynga's New Ad Pitch for Draw Something: 'Draw This Brand' — NHL Among First to Buy Paid Terms in Hugely Popular Social game — Zynga's latest big-ticket acquisition has already figured out how to draw in users, but now Draw Something has an ad model that brings brands into the picture.

More: TechCrunch, Betabeat, The Next Web, Simply Zesty, memeburn, WebProNews, Tecca, The Verge and VG247. **Thanks:** @kunur

---

Jay Yarow / Business Insider:

Apple Is Beating Android In The U.S., Despite Reports To The Contrary — Apple appears to have taken control of 50% of the smartphone market in the first quarter of 2012, despite a report to the contrary by NPD this morning. — NPD put out a press release saying Android …

More: Appolicious Advisor, parislemon, NPD Group, WebProNews, TechCrunch, GigaOM, SlashGear, I4U News, Macgasm and Daring Fireball

---

Forbes:

Comparable corpus

Kunur Patel / AdAge:

Zynga's New Ad Pitch for Draw Something: 'Draw This Brand' — NHL Among First to Buy Paid Terms in Hugely Popular Social game — Zynga's latest big-ticket acquisition has already figured out how to draw in users, but now Draw Something has an ad model that brings brands into the picture.

More: TechCrunch, Betabeat, The Next Web, Simply Zesty, memeburn, WebProNews, Tecca, The Verge and VC247. **Thanks:** @kunur

With the ink still drying following the acquisition of OMGPOP and its hit app Draw Something, Zynga is putting its new mobile property to work by engaging advertisers and encouraging them to pay for words that allow users to literally draw their brands.

AdAge reports that Zynga is pushing the new form of advertising (on top of its mobile ad banners and paid upgrades), now inserting words connected with brands, encouraging app users to draw logos or a product that they sell.

The National Hockey League is one of the first advertisers to have paid for ice

Zynga's stock price is dropping, so now the San Francisco-based company is attempting to devise creative ways to monetize some of its more popular games. Yesterday we reported that Draw Something is hemorrhaging users, but that hasn't stopped its new parent company from thrusting more ads upon the once wildly-popular game.

Zynga's latest big-ticket acquisition has already figured out how to draw in users, but now Draw Something has an ad model that brings brands into the picture.

Until recently, the Pictionary-like game had only run spammy banner ads in its free mobile app that, including the paid no-ads version, has amassed a staggering 50 million downloads in five months. Now, with a direct-sales force that's been on the ground for a whole eight weeks, Draw Something is inserting advertisers' paid terms into the game for players to literally draw brands.

Here's how the game works: Pick a word from a list of three, then create a drawing so a Facebook

Comparable corpus

Kunur Patel / AdAge:

Zynga's New Ad Pitch for Draw Something: 'Draw This Brand' — NHL Among First to Buy Paid Terms in Hugely Popular Social game — Zynga's latest big-ticket acquisition has already figured out how to draw in users, but now Draw Something has an ad model that brings brands into the picture.

More: TechCrunch, Betabeat, The Next Web, Simply Zesty, memeburn, WebProNews, Tecca, The Verge and VG247. **Thanks:** @kunur

With the ink still drying following the acquisition of OMGPOP and its hit app Draw Something, Zynga is putting its new mobile property to work by engaging advertisers and encouraging them to pay for words that allow users to literally draw

the new form of advertising (on top of its
), now inserting words connected with brands,
or a product that they sell.

f the first advertisers to have paid for ice

Zynga's stock price i
based company is at
monetize some of its
reported that Draw
that hasn't stopped i
more ads upon the c

2 years worth of Techmeme
160 million words
374,547 documents
24,612 stories

Zynga's latest big-ticket ac w
Something has an ad mod

Until recently, the Pictionary-like game had only run spammy banner ads in its free mobile app that, including the paid no-ads version, has amassed a staggering 50 million downloads in five months. Now, with a direct-sales force that's been on the ground for a whole eight weeks, Draw Something is inserting advertisers' paid terms into the game for players to literally draw brands.

Here's how the game works: Pick a word from a list of three, then create a drawing so a Facebook

Extraction

- Stanford sentence splitter, tagger, NER
- MaltParser (linear time)

- Top 10 tf*idf ranked verbs for each story

$$\mathrm{tf}(v, s) * \mathrm{idf}(v, S) = \mathrm{tf}(v, s) * \log \frac{|S|}{|\{s \in S : v \in s\}|}$$

  - Phrasal verbs (*give up* vs. *give away*)
  - Excluding: light verbs (*do*, *have*, *give*...)
    report verbs (*say*, *tell*...)
    copular verbs (*seem*, *become*...)
  - WordNet synonyms are included (*release, publish...*)

Extraction

- Assumption: In a story, the same verb refers to the same event
- Subjects and objects are clustered, repectively
  - Passive constructions (*X compromised Y* → *Y has been compromised*)
  - Ergative verbs (*X scattered Y* → *Y scattered*)
  - Nominalizations from NomBank (*acquire* → *Google's **acquisition** of Sparrow*)
- Exclude same-head NPs and pronouns

Google
the Internet giant
the search giant
the company

crawl

locate

their missing phones
lost or stolen smartphones
your device
your lost iPhone

Extraction

Coreference relations

Android phones *products*

pictures *shots*

Mark Zuckerberg *the hoodie-wearing Facebook co-founder*

Bad relations

① Parsing errors

[attacks against Chrome]$_S$ exploit ... *[the full details on the]$_S$ exploit*

② Algorithm violations (one verb ≠ one event)

Remove [spam from the emails]$_O$ ... *Remove [the test accounts]$_O$*

③ Text extraction errors

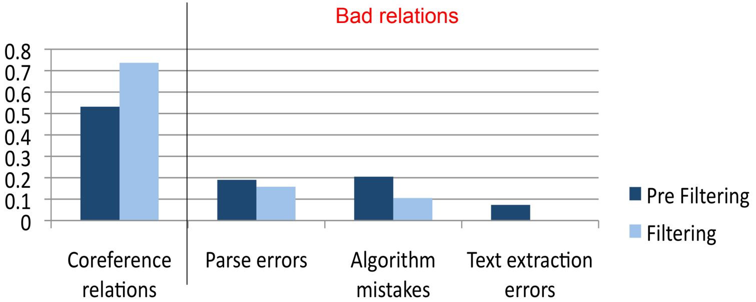
* <li id="gadgets"> Networking *

Filtering

Filters for:

- Parsing errors
  - Non-nominal head *shopping (ah*
- Algorithm violations
  - NE – NE *Yahoo, Google*
  - Negation *[But the operators aren't mandating] plans*
  - Enumeration *[1. Remove] spam from the emails*
  - Numbers *40,000 per year*
  - Temporals *6:00 PM Pacific time*
- Text extraction errors
  - Mention length *charges that Google unfairly ranks competitors in its search results, penalizing them with lower rankings*
  - Sentence length
  - Ill-formed sentence * <li id="gadgets">*

Filtering


Bad relations
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Coreference relations
Parse errors
Algorithm mistakes
Text extraction errors
Pre Filtering
Filtering

Generalization

128,492 coreferent pairs

- Remove
  - determiners

    the promotion >> promotion
  - relative, -ing, -ed clauses

    the device available online from Google >> device
- Keep adjectives and prepositional modifiers

  ***online** piracy*

  *distribution **of pirated material***
- Generalize NE to types

  Cook's departure >> PERSON's departure
- Lemmas

  data >> datum

Generalization

- Frequency counts

(rule, limitation)	5	(company, HP)	35
(phone, experience)	1	(company, price)	12
(FBI, agent)	20		

- Normalized PMI (Bouma 2009) [–1, 1]

$$\mathrm{PMI}(x,y) = \ln \frac{p(x,y)}{p(x)\,p(y)} \qquad \mathrm{NPMI}(x,y) = \frac{\mathrm{PMI}}{-\ln p(x,y)}$$

(rule, limitation)	0.417	(company, HP)	0.203
(phone, experience)	–0.152	(company, price)	–0.053
(FBI, agent)	0.566		

Dictionary snapshot

offering, IPO
password, login information
user, consumer
firm, company
phone, device
Apple, company
iPad, slate
Android, platform
site, company
app, software
agreement, wording
platform, code
filing, complaint

search, search result
update, change
bug, issue
Google, search giant
search algorithm, search engine
hardware key, digital lock
content, photo
rule, limitation
coupon, sale
medical record, medical file
device, developer
version, handset
Groupon, company

Dictionary snapshot

- Synonymy

  user, consumer
- Hypernymy

  Google, company
- Metonymy

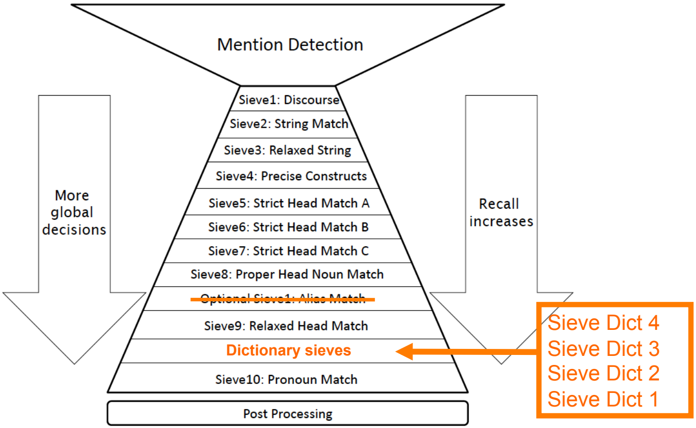
  cloud, users
- General nouns

  bug, issue
- World knowledge

  Google, search giant

Stanford coreference system

(Lee et al. 2011)



Existing Tools

Two good tools (available for download) are:

- Stanford Coreference System
  - Sieve + singletons
- Illinois Coreference Package
  - Pairwise classification with strong features

Summary

The coreference resolution task

- Definition
- Evaluation
- Pairwise / machine learning approach
  - Features
  - Constructing training examples
- Rule-based systems (Sieve, baby-steps)
  - Many smart decisions
  - Global constraints
- Targetting specific problems
  - A method for detecting singletons
  - “semantic” knowledge acquisition from large corpus

Extensions: event (predicate) coreference, cross document coreference