CS 4650/7650
Anaphora and coreference

Jacob Eisenstein

October 30, 2014

At a press conference Monday, Wikileaks founder Julian Assange said he would leave the embassy where he been hiding in plain sight "soon," the Guardian reported.
He did not say when "soon" meant, however.
Assange, whose organization facilitated the publication of materials leaked by Bradley Manning (now Chelsea Manning), has been in exile at the Ecuadorian embassy in London for more than two years. He is wanted in Sweden, where he allegedly sexually assaulted two women.
"It has been two years since I have been granted political asylum in this embassy," said Assange, sporting a white beard. He added: "I have not been charged with an offense... How can it be that ... a person is held and his freedom of movement restricted?"

At a press conference Monday, Wikileaks founder Julian Assange said he would leave the embassy where he been hiding in plain sight "soon," the Guardian reported.
He did not say when "soon" meant, however.
Assange, whose organization facilitated the publication of materials leaked by Bradley Manning (now Chelsea Manning), has been in exile at the Ecuadorian embassy in London for more than two years. He is wanted in Sweden, where he allegedly sexually assaulted two women.
"It has been two years since I have been granted political asylum in this embassy," said Assange, sporting a white beard. He added: "I have not been charged with an offense... How can it be that ... a person is held and his freedom of movement restricted?"

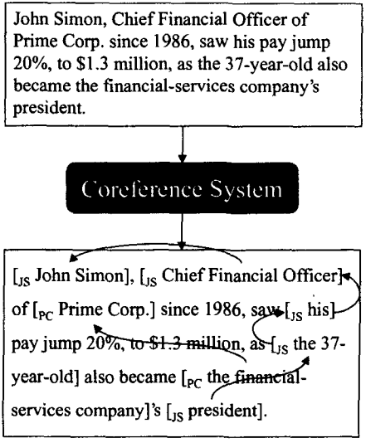
- Who has been hiding in plain sight?
- In what Embassy has he been hiding?
- What organization facilitated the publication of the leaks?
- Whose freedom of movement is restricted?

Some terminology

Assange said he is able to get only one hour of exercise per day. The Ecuadorian foreign minister, Ricardo Patino, said **his$_1$ country$_2$** will "continue to offer **him$_3$ our$_4$** protection," according to the Guardian.

- **Referring expressions**: Assange, he, our, his, his country ...
- **Referents** are (often) entities, like JULIAN-ASSANGE, ECUADOR, ECUADORIAN-EMBASSY
- **Coreference** is when referring expressions have the same referent. The Ecuadorian foreign minister, Ricardo Patino, and he are all coreferent.
- Assange **evokes** the referent JULIAN-ASSANGE.
  He **accesses** it.

Coreference resolution



(Cardie and Wagstaff 1999)

Choosing referring expressions

There are many possibilities for describing a referent.

- Indefinite NPs: a person, two years
- Definite NPs: the Ecuadorian embassy
- Pronouns: he, his, I, you
- Demonstratives: this chainsaw, that abandoned mall
- Names: Bradley Manning, London

Choosing referring expressions

How do you know which type of referring expression to use?

- **Language generation** requires getting this right.
- **Language understanding** requires figuring out which referent is intended by ambiguous referring expressions.
  - Anaphora resolution deals with pronouns like it, this, her
  - Coreference resolution also includes
    - **Names**: Barack Obama, Obama, President Obama, Barry O, Nobama, Obummer
    - **Nominals**: the 44th president, the former senator from Illinois, our first African-American president

Are all referents entities?

- They told me that I was too ugly, but I didn't believe **it**.
- Alice saw Bob get angry, and I saw **it** too.
- They told me that I was too ugly, but **that** was a lie.
- Jess said she worked in security.
  I suppose **that**'s one way to put it.

Are all pronouns referential?

Generic referents

- A good father takes care of **his** kids.
- I want to buy a Porsche, **they** are so fast.
- On the moon, **you** have to carry **your** own oxygen.

Are all pronouns referential?

Generic referents

- A good father takes care of **his** kids.
- I want to buy a Porsche, **they** are so fast.
- On the moon, **you** have to carry **your** own oxygen.

Nonreferential pronouns

- Pleonastic: **It**'s raining. **It**'s crazy out there.
- Cleft: **It**'s money that she's really after.
- Extraposition: **It** sucks that we have to work so hard.
- Other languages:
  - S'**il** vous plaît (literally: if it pleases you)
  - Wie geht es Ihnen (literally: how's it going to you)

Detecting ambiguous references

Bergsma, Lin, and Goebel (2008) propose a substitutability text.

- You can make it in advance → You can make **them** in advance
- You can make it in Hollywood → You can make **them** in Hollywood

Substitutability test

... said here Thursday that **it** is unnecessary to continue

said	here	Thursday	that	*					
	here	Thursday	that	*	is				
		Thursday	that	*	is	unnecessary			
			that	*	is	unnecessary	to		
				*	is	unnecessary	to	continue	

Substitutability test

- For each pattern, compute counts of five **pattern fillers**:
  1. it/its
  2. they/them/their
  3. other pronouns she/her/...
  4. rare words (almost always nouns)
  5. all other tokens (usually nouns)
- Convert these counts into a feature vector
- Train a supervised classifier
- Data: `http://webdocs.cs.ualberta.ca/~bergsma/ItBank/`
- Nice combination of big unlabeled data & small labeled data.

Ambiguous references

Assange said he is able to get only one hour of exercise per day. Ecuadorian foreign minister, Ricardo Patino, said **his**$_1$ **country**$_2$ will "continue to offer **him**$_3$ **our**$_4$ protection," according to the Guardian.

- **his** $\stackrel{?}{=}$ Patino, Ecuador, Assange, ...
- **him** $\stackrel{?}{=}$ Patino, Assange, Ecuador, ...
- **our** $\stackrel{?}{=}$ Patino, Assange, Ecuador, the Guardian, ...
- **his country** $\stackrel{?}{=}$ Ecuador, United States, London, Sweden, ...

How can we resolve these references?

Constraints

Constraints

Agreement, e.g. number

- Number(he) = singular
- Number($\mathrm{Assange}$) = singular
- Number($\mathrm{officials}$) = plural

Constraints

Agreement, e.g. number

- Number(he) $=$ singular
- Number(Assange) $=$ singular
- Number(officials) $=$ plural

Syntax

- Uma spoke with her
- Can Uma and her be coreferential?
- Why not?

Preferences

Preferences

- **Recency**:
  Drew made **a nice pie**, but Liz made **an even better pie**.
  It had apples and bacon.
- **Repeated mention**
- **Grammatical role**: subj > direct obj > indirect obj
  Elmo went to the bar with **Grover**.
  He ordered two vodka tonics.
- **Parallelism**: **Ellen** went with **Linda** to Providence.
  Jim went with **her** to Boston.
- **Selectional preference**:
  They took **the dishes** from **the guests**, and washed **them**.
- **Semantics**:
  - Elmo telephoned Grover. He had broken the laptop.
  - Elmo yelled at Grover. He had broken the laptop.

Combining constraints and preferences

- Several constraints and preferences on anaphora resolution, some of which conflict. How to combine them?
  - **Hobbs**: tree search plus constraints
  - **Centering**: ordered preferences plus constraints
  - **Lappin & Lease**: numerical preferences + constraints
  - **Ge, Hale, & Charniak**: multiply probabilities

Combining constraints and preferences

- Several constraints and preferences on anaphora resolution, some of which conflict. How to combine them?
  - **Hobbs**: tree search plus constraints
  - **Centering**: ordered preferences plus constraints
  - **Lappin & Lease**: numerical preferences + constraints
  - **Ge, Hale, & Charniak**: multiply probabilities
- Modern NLP: encode preferences and constraints as features in a supervised classifier.
- For each ambiguous referring expression, predict the correct **antecedent**.

$$\hat{y}_i = \arg\max_{j<i \vee j=\varnothing} \boldsymbol{\theta}^\top \mathbf{f}(i, j, \mathbf{x}) \tag{1}$$

This is called the **mention-pair** model.

The multipass sieve

Pass	Type	Features
1	N	exact extent match
2	N,P	appositive | predicate nominative | role appositive | relative pronoun | acronym | demonym
3	N	cluster head match & word inclusion & compatible modifiers only & not i-within-i
4	N	cluster head match & word inclusion & not i-within-i
5	N	cluster head match & compatible modifiers only & not i-within-i
6	N	relaxed cluster head match & word inclusion & not i-within-i
7	P	pronoun match

Progressively relaxed matching rules:

- Appositive, e.g. Obama, the 44th president, ...
- i-within-i: child and parent NPs cannot corefer
- Pronouns must pass agreement criteria

Features for classification

(1) Separately, Clinton transition officials said that *Frank Newman*, 50, *vice chairman* and chief financial officer of BankAmerica Corp., is expected to be nominated as assistant Treasury secretary for domestic finance.

Table 1
Feature vector of the markable pair (i = *Frank Newman*, j = *vice chairman*).

Feature	**Value**	**Comments**
DIST	0	i and j are in the same sentence
I_PRONOUN	–	i is not a pronoun
J_PRONOUN	–	j is not a pronoun
STR_MATCH	–	i and j do not match
DEF_NP	–	j is not a definite noun phrase
DEM_NP	–	j is not a demonstrative noun phrase
NUMBER	+	i and j are both singular
SEMCLASS	1	i and j are both persons (This feature has three values: false(0), true(1), unknown(2).)
GENDER	1	i and j are both males (This feature has three values: false(0), true(1), unknown(2).)
PROPER_NAME	–	Only i is a proper name
ALIAS	–	j is not an alias of i
APPOSITIVE	+	j is in apposition to i

Notes on features

Although the overall approach is statistical, a substantial amount of knowledge is encoded in the features.

- **STR_MATCH** tests string equality after stripping determiners: the data matches that data
- **SEM_CLASS** is based on WordNet, and considers FEMALE, MALE, PERSON, ORGANIZATION, LOCATION, DATE, TIME, MONEY, PERCENT, OBJECT.
- **GENDER** is based on titles and a database of names
- **ALIAS** is based on a database on known aliases, e.g. IBM/International Business Machines

Soon et al (2001) classifier

```
STR_MATCH = +: +
STR_MATCH = -:
:...J_PRONOUN = -:
    :...APPOSITIVE = +: +
    :   APPOSITIVE = -:
    :   :...ALIAS = +: +
    :       ALIAS = -: -
    J_PRONOUN = +:
    :...GENDER = 0: -
        GENDER = 2: -
        GENDER = 1:
        :...I_PRONOUN = +: +
            I_PRONOUN = -:
            :...DIST > 0: -
                DIST <= 0:
                :...NUMBER = +: +
                    NUMBER = -: -
```

Coreference as structured prediction

Coreference decisions are interdependent:

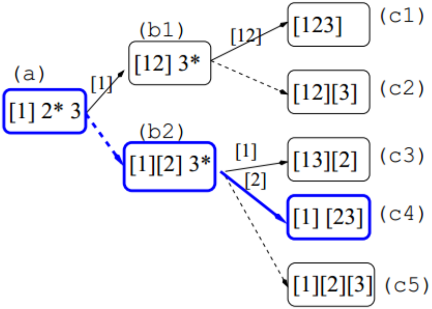
- Example:
  - Obama congratulated the Brazilian President.
  - She wished him a happy birthday.
- If she refers to the Brazilian President,
  then him likely refers to Obama
- If she refers to Obama,
  then him likely refers to the Brazilian President.

Can we make these coreference decisions jointly?

$$\mathbf{y} = \arg\max_{\mathbf{y}} \boldsymbol{\theta}^{\top} \mathbf{f}(\mathbf{y}, \mathbf{x}) \tag{2}$$

Bell Tree clustering for Coreference

We can incrementally build a coreference clustering using the **Bell Tree** representation (Luo et al, 2004)



- Can search over many clusterings, but size of Bell Tree (Bell Number) grows very rapidly, so pruning is needed.
- Approximate $P(y|x_i, e_j) \approx \max_{x' \in e_j} P(y|x_i, x')$

Coreference as a Markov random field

Basic idea is to compute transitive closure of all coreference pairs (Wellner et al, 2003)

- x_i, x_j: mentions
- y_{ij}: coreference label for x_i and x_j
- $\mathbf{f}(x_i, x_j, y_{ij})$: features of the mention pair
- $f_*(y_{ij}, y_{jk}, y_{ij})$: agreement feature for the label triple

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z_x} \exp\left(\sum_{i,j} \boldsymbol{\theta}^\top \mathbf{f}(x_i, x_j, y_{ij}) + \sum_{i,j,k} \theta_* f_*(y_{ij}, y_{jk}, y_{ij}) \right)$$

Can learn weights $\boldsymbol{\theta}$ as a conditional random field or structured perceptron, but inference is NP-Hard.

ILP formulation

- $y_{ij} = 1$ if mentions i and j are coreferent
- $\psi_{ij} = \boldsymbol{\theta}^{\mathsf{T}}\mathbf{f}(x_i, x_j)$
- Can you see how to formulate joint inference as an ILP?

ILP formulation

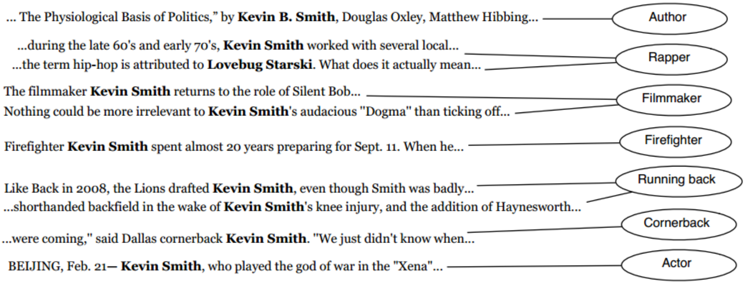
- $y_{ij} = 1$ if mentions i and j are coreferent
- $\psi_{ij} = \boldsymbol{\theta}^{\mathsf{T}}\mathbf{f}(x_i, x_j)$
- Can you see how to formulate joint inference as an ILP?
- Denis and Baldridge (2007) go further, adding **anaphoricity** as a variable z_i, with constraints

$$y_{i,j} \leq z_i, \forall i, j$$

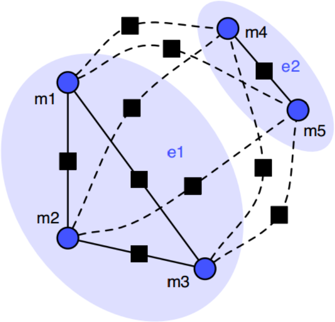
$$z_i \leq \sum_j y_{i,j}$$

Multi-document coreference resolution

Broaden the task: find all mentions of an entity across a big set of documents.



A pairwise model



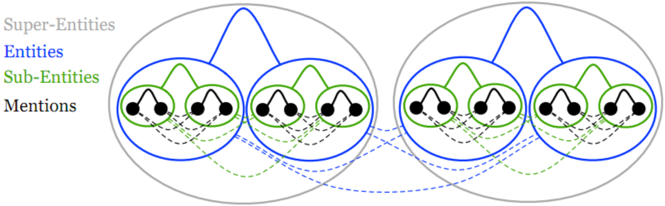
$$p(\mathbf{e}) \propto \exp \sum_{e \in \mathbf{e}} \left\{ \sum_{m,n \in e, n \neq m} \psi_a(m,n) + \sum_{m \in e, n \notin e} \psi_r(m,n) \right\}$$

Singh *et al.*(ACL 2011) estimate $p(\mathbf{e})$ using Metropolis-Hastings, with moves to swap mentions between entities.

Parallelizing cross-document coreference

A four-level hierarchy:

- Mentions
- Sub-entities (sets of mentions that likely corefer)
- Entities (sets of mentions thought to corefer)
- Super-entities (sets of entities which might corefer)



- Key idea: parallelize by keeping super-entities on separate machines, occasionally shuffling them around.
- Using a very simple mention similarity function, they can scale up to 1.5M mention strings.