807 - TEXT ANALYTICS

Massimo Poesio

Lecture 7: Anaphora resolution (Coreference)

Anaphora resolution: the problem

Example

Sophia Loren says she will always be grateful to Bono. The actress revealed that the U2 singer helped her calm down when she became scared by a thunderstorm while travelling on a plane.

- *she* ⇒ *Sophia Loren*
- *the actress* ⇒ *Sophia Loren*
- *the U2 singer* ⇒ *Bono*
- *her* ⇒ *Sophia Loren*
- *she* ⇒ *Sophia Loren*

Anaphora resolution: coreference chains

Example

Sophia Loren says she will always be grateful to Bono. The actress revealed that the U2 singer helped her calm down when she became scared by a thunderstorm while travelling on a plane.

Coreference Chains:

- {Sophia Loren, she, the actress, her, she}
- {Bono, the U2 singer }
- {a thunderstorm}
- {a plane}

Anaphora resolution as Structure Learning

- So far we have only seen examples of text analytics applications in which the task was to label a SINGLE OBJECT
- In the case of anaphora resolution/coreference, the task is to label a STRUCTURE
  - In its simplest form, the antecedent / anaphor pair (MENTION PAIR)
- This is an example of so-called STRUCTURED LEARNING

Factors that affect the interpretation of anaphoric expressions

- Factors:
  - Morphological features (agreement)
  - Syntactic information
  - Salience
  - Lexical and commonsense knowledge
- Distinction often made between CONSTRAINTS and PREFERENCES

Agreement

- GENDER strong CONSTRAINT for pronouns (in other languages: for other anaphors as well)
  - [Jane] blamed [Bill] because HE spilt the coffee (Ehrlich, Garnham e.a, Arnold e.a)
- NUMBER also strong constraint
  - [[Union] representatives] told [the CEO] that THEY couldn't be reached

Lexical and commonsense knowledge

[The city council] refused [the women] a permit because they feared violence.
[The city council] refused [the women] a permit because they advocated violence.
Winograd (1974), Sidner (1979)

BRISBANE – a terrific right rip from [Hector Thompson] dropped [Ross Eadie] at Sandgate on Friday night and won him the Australian welterweight boxing title. (Hirst, 1981)

Problems to be resolved by an AR system: mention identification

- Effect: recall
- Typical problems:
  - Nested NPs (possessives)
    - [a city] 's [computer system] → [[a city]'s computer system]
  - Appositions:
    - [Madras], [India] → [Madras, [India]]
  - Attachments

Problems for AR: agreement extraction

- The committee are meeting / is meeting
- The Union sent a representative. They ….
- The doctor came to visit my father. SHE told him …

Problems to be solved: anaphoricity determination

- Expletives:
  - IT's not easy to find a solution
  - Is THERE any reason to be optimistic at all?
- Non-anaphoric definites

Problems for AR: Complex attachments

- [The quality that's coming out of [software from [India]]
  - The quality that's coming out of software from India is now exceeding the quality of software that's coming out from the United States
- scanning through millions of lines of computer code
  - ACE/bnews/devel/ABC19981001.1830.1257

Early systems

- Hobbs 1976 Naïve Algorithm
  - Pronouns only
  - Syntax based
  - Still very competitive
- Sidner 1979
- Carter 1986

MODERN WORK IN ANAPHORA RESOLUTION

- Availability of the first anaphorically annotated corpora circa 1993 (MUC6) made statistical methods possible
- Most current anaphora resolution systems are based on machine learning, but there is one notable exception, the Stanford Coreference system

MUC

- First big initiative in Information Extraction
- Produced first sizeable annotated data for coreference
- Developed first methods for evaluating systems

MUC terminology:

- MENTION: any markable
- COREFERENCE CHAIN: a set of mentions referring to an entity
- KEY: the (annotated) solution (a partition of the mentions into coreference chains)
- RESPONSE: the coreference chains produced by a system

The Stanford Deterministic Coreference Resolution System

- Part of the Stanford CORE Pipeline
- The best-performing system at CONLL 2011, and used as a component by two of the top three systems at CONLL 2012
- Key to its performance are
  - A very high quality mention detection component based on the Stanford CORE pipeline
  - A PRECISION-FIRST coreference resolution component based on 10 filters called SIEVES that implement many of the restrictions on anaphora resolution discussed in previous slides

The Sieves

1. *Speaker Identification*: This sieve first identifies speakers, then matches first and second pronouns to these speakers.
2. *ExactMatch*:Thissievelinkstogethertwomentionsonlyiftheycontainexactly the same text, including both determiners and modifiers.
3. *Relaxed String Match*: This sieve links together two mentions only if they contain exactly the same text after dropping the postmodifiers.
4. *Precise Constructs*: This sieve links together two mentions if they occur in one of a series of high precision constructs: e.g., if they are in an appositive construction (*[the speaker of the House], [Mr. Smith]* . . .), or if both mentions are tagged as NNP and one of them is an acronym of the other.
5. *Strict Head Match*: This sieve links together a mention with a candidate antecedent entity if *all* of a number of constraints are satisfied: (a) the head of the mention matches any of the heads of the candidate antecedent; (b) all non- stop words of the mention are included in the non-stop words of the candidate antecedent; (c) all mention modifiers are included among the modifiers of the candidate antecedent; and (d) the two mentions are not in an i-within-i situation, i.e., one is not a child in the other.

The Sieves

6. *Variants of Strict Head Match*: Sieve 6 relaxes the 'compatible modifiers only' constraint in the previous sieve, whereas Sieve 7 relaxes the 'word inclusion' constraint.
7. *Proper Head Match*: This sieve links two proper noun mentions if their head words match and a few other constraints apply.
8. *Relaxed Head Match*: This sieve relaxes the requirement that the head word of the mention must match a head word of the candidate antecedent entity.
9. *Pronounresolution*:Finally,pronounsareresolved,byfindingcandidatesmatch- ing the pronoun in number, gender, person, animacy, and NER label, and at most 3 sentences distant.

STATISTICAL APPROACHES TO ANAPHORA RESOLUTION

- UNSUPERVISED approaches
  - Eg Cardie & Wagstaff 1999, Ng 2008
- SUPERVISED approaches
  - Early (NP type specific)
  - Soon et al: general classifier + modern architecture

Soon et al 2001

- First 'modern' ML approach to anaphora resolution
  - Resolves ALL anaphors
  - Fully automatic mention identification
- Developed instance generation & decoding methods used in a lot of work since

ANAPHORA RESOLUTION AS A CLASSIFICATION PROBLEM

1. Classify MENTION PAIR <NP1,NP2> as coreferential or not
2. Build a complete coreferential chain

Soon et al: MENTION PAIRS

<ANAPHOR (j), ANTECEDENT (i)>

SOME KEY DECISIONS

- ENCODING
  - I.e., what positive and negative instances to generate from the annotated corpus
  - Eg treat all elements of the coref chain as positive instances, everything else as negative:
- DECODING
  - How to use the classifier to choose an antecedent
  - Some options: 'sequential' (stop at the first positive), 'parallel' (compare several options)

Soon et al: preprocessing

- POS tagger: HMM-based
  - 96% accuracy
- Noun phrase identification module
  - HMM-based
  - Can identify correctly around 85% of mentions (?? 90% ??)
- NER: reimplementation of Bikel Schwartz and Weischedel 1999
  - HMM based
  - 88.9% accuracy

Soon et al 2001: Features

- NP type
- Distance
- Agreement
- Semantic class

Soon et al: NP type and distance

NP type of anaphor j (3)
`j-pronoun, def-np, dem-np (bool)`

NP type of antecedent i
`i-pronoun (bool)`

Types of both
`both-proper-name (bool)`

DIST
`0, 1, ….`

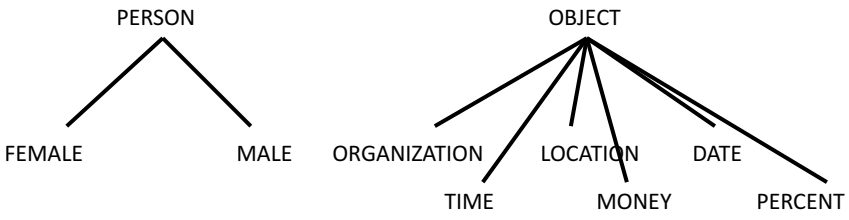
Soon et al features: string match, agreement, syntactic position

STR_MATCH
ALIAS
dates (1/8 – January 8)
person (Bent Simpson / Mr. Simpson)
organizations: acronym match
(Hewlett Packard / HP)

AGREEMENT FEATURES
number agreement
gender agreement

SYNTACTIC PROPERTIES OF ANAPHOR
occurs in appositive contruction

Soon et al: semantic class agreement

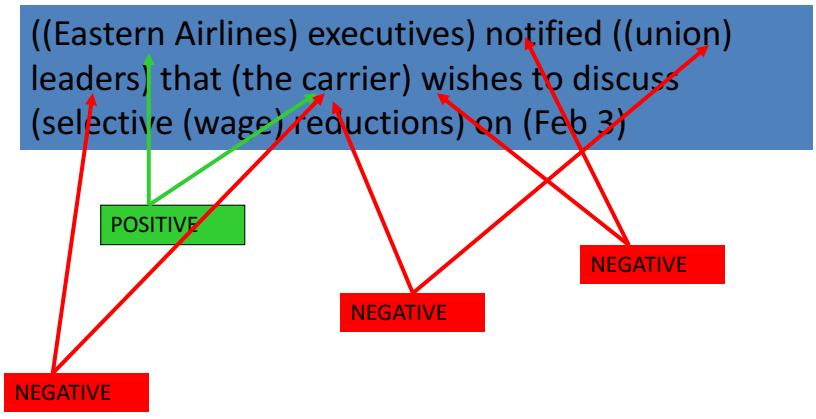


SEMCLASS = true iff semclass(i) <= semclass(j) or viceversa

Soon et al: generating training instances

- Marked antecedent used to create positive instance
- All mentions between anaphor and marked antecedent used to create negative instances

Generating training instances



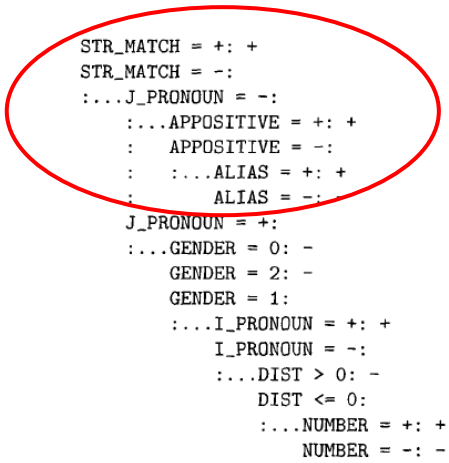
Soon et al: decoding

- Right to left, consider each antecedent until classifier returns true

Soon et al: evaluation

- MUC-6:
  - P=67.3, R=58.6, F=62.6
- MUC-7:
  - P=65.5, R=56.1, F=60.4

Soon et al: evaluation



Evaluation of coreference resolution systems

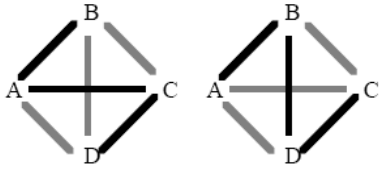
- Lots of different measures proposed
- ACCURACY:
  - Consider a mention correctly resolved if
    - Correctly classified as anaphoric or not anaphoric
    - 'Right' antecedent picked up
- Measures developed for the competitions:
  - Automatic way of doing the evaluation
- More realistic measures (Byron, Mitkov)
  - Accuracy on 'hard' cases (e.g., ambiguous pronouns)

Vilain et al 1995

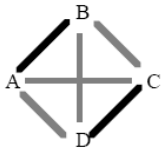
- The official MUC scorer
- Based on precision and recall of links

Vilain et al: the goal

The problem: given that A,B,C and D are part of a coreference chain in the KEY, treat as equivalent the two responses:



And as superior to:



Vilain et al: RECALL

- To measure RECALL, look at how each coreference chain S_i in the KEY is partitioned in the RESPONSE, and count how many links would be required to recreate the original, then average across all coreference chains.

$$R_T = \frac{\sum(|S_i| - |p(S_i)|)}{\sum(|S_i| - 1)}$$

Vilain et al: Example recall

- In the example above, we have one coreference chain of size 4 (|S| = 4)
- The incorrect response partitions it in two sets (|p(S)| = 2)
- R = 4-2 / 4-1 = 2/3

Vilain et al: precision

- Count links that would have to be (incorrectly) added to the key to produce the response
- I.e., 'switch around' key and response in the equation before

Beyond Vilain et al

- Problems:
  - Only gain points for links. No points gained for correctly recognizing that a particular mention is not anaphoric
  - All errors are equal
- Proposals:
  - Bagga & Baldwin's B-CUBED algorithm
  - Luo recent proposal

After Soon et al 2001

- Different models of the task
- Different preprocessing techniques
- Using lexical / commonsense knowledge (particularly semantic role labelling)
- Salience
- Anaphoricity detection
- Development of AR toolkits (GATE, LingPipe, GUITAR)

Error analysis (Soon et al)

- Errors most affecting precision:
  - Prenominal modifiers identified as mentions and other errors in mention identification
  - String match but noun phrases refer to different entities
- Errors most affecting recall:
  - Errors in mention identification (11%)
  - Errors in SEMCLASS determination (10%)
  - Need more features (63.3%)

Soon et al examples of errors:

- Tarnoff, a former Carter administration official and president of the Council on foreign relations, is expected to be named **[undersecretary]** for political affairs … Former. Sen Tim Wirth is expected to get a newly created **[undersecretary]** post for global affairs
- [Ms Washington and Mr. Dingell] have been considered [allies] of **[the Securities exchanges],** while [banks] and **[future exchanges]** often have fought with THEM

Mention detection errors in GUITAR (Kabadjov, 2007)

[The bow] (see detail, below right) is decorated with a complicated arrangement of horses and lions' heads.

Above the lions' heads are four sphinxes.

Three pairs of lions clamber up the section from the point where [the sheath and bow] are joined.

More recent models

- Cardie & Wagstaff: coreference as (unsupervised) clustering
  - Much lower performance
- Ranking models:
  - Ng and Cardie 2002
  - Yang 'twin-candidate' model
- Entity-mention models
- Joint entity detection & tracking

Ng and Cardie 2002

- 2002:
  - Changes to the model:
    - Positive: first NON PRONOMINAL
    - Decoding: choose MOST HIGH PROBABILITY
  - Many more features:
    - Many more string features
    - Linguistic features (binding, etc)
- Subsequently:
  - Discourse new detection

Ranking models

- Idea: train a model that imposes a **ranking** on the candidate antecedents for an NP to be resolved so that it assigns the highest rank to the correct antecedent
- A ranker allows all candidate antecedents to be considered simultaneously and captures competition among them
  - Allows us find the best candidate antecedent for an NP
- There is a natural resolution strategy for a ranking model
  - An NP is resolved to the highest-ranked candidate antecedent

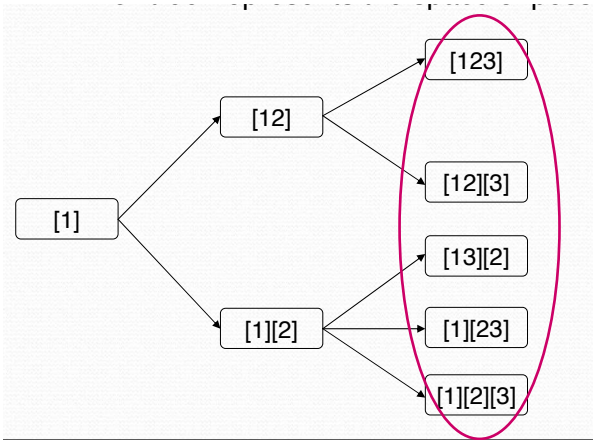
How to train a ranking model

- Convert the problem of ranking *m* NPs into the a set of pairwise ranking problems
  - Each pairwise ranking problem involves determining which of two candidate antecedents is better for an NP to be resolved
    - Each one is essentially a classification problem
- Ranking rediscovered independently by
  - Yang et al. (2003) (twin-candidate model)
  - Iida et al. (2003) (tournament model)
- Denis & Baldridge (2007, 2008): train the ranker using maximum entropy
  - model outputs a rank value for each candidate antecedent

Entity-mention models

- Classifiers that determine whether (or how likely) an NP belongs to a preceding COREFERENCE CLUSTER

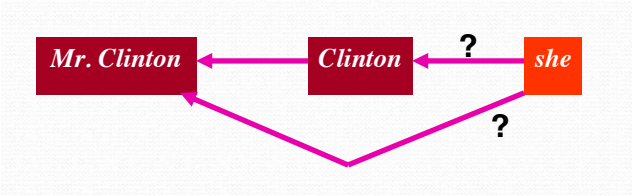
Luo et al's Bell Tree model



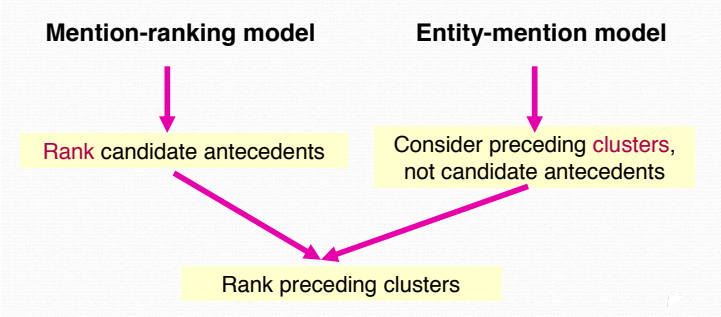
Entity-mention models

- Classifiers that determine whether (or how likely) an NP belongs to a preceding coreference cluster
- more **expressive** than the mention-pair model
  - can employ cluster-level features defined over any subset of NPs in a preceding cluster

Cluster-level features



Rahman and Ng's cluster-ranking model



Joint Entity Detection and Tracking

- Daume and Marcu 2005: Mention identification, classification, and linking take place at the same time
- Denis and Balridge 2007: ILP

The state of the art in coreference: the 2012 CONLL Shared Task

- Data: OntoNotes
  - 1.6M words English, 900K words Chinese, 300K words Arabic
  - Annotated with: syntactic information, wordsenses, propositional information
- Tracks:
  - Closed
  - Open
- Metrics: MELA
  - (a combination of MUC / B3 / CEAF)

CONLL 2012 ST: RESULTS

Participant	Open			Closed			Official	Final model	
	English	Chinese	Arabic	English	Chinese	Arabic	Score	Train	Dev
fernandes				**63.37**	58.49	**54.22**	**58.69**	√	√
björkelund				61.24	59.97	53.55	58.25	√	√
chen		**63.53**		59.69	**62.24**	47.13	56.35	√	×
stamborg				59.36	56.85	49.43	55.21	√	√
uryupina				56.12	53.87	50.41	53.47	√	√
zhekova				48.70	44.53	40.57	44.60	√	√
li				45.85	46.27	33.53	41.88	√	√
yuan		61.02		58.68	60.69		39.79	√	√
xu				57.49	59.22		38.90	√	×
martschat				61.31	53.15		38.15	√	×
chunyang				59.24	51.83		37.02	–	–
yang				55.29			18.43	√	×
chang				60.18	45.71		35.30	√	×
xinxin				48.77	51.76		33.51	√	√
shou				58.25			19.42	√	×
xiong	59.23	44.35	44.37				0.00	√	√

ANAPHORA / COREFERENCE DATASETS

- MUC6/MUC7 (small, old)
- ACE 2002/2005
- ONTONOTES
- ARRAU (locally developed)

Tools for AR

- Java-RAP (pronouns)
- GUITAR (Kabadjov, 2007)
- BART (Versley et al, 2008)
- Stanford Deterministic Coreference Resolver (Lee et al 2013)
  - See labs
- CORT (Martschat & Strube 2015)
  - See in labs

Readings

- W. M. Soon, H. T. Ng, and D. C. Y. Lim, 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521—544
- Vincent Ng, 2010. Supervised Coreference Resolution: the first fifteen years. *Proc. Of the ACL*.