Coreference Resolution

Part 2

CS224n
Christopher Manning
(borrows slides from Roger Levy, Altaf Rahman, Vincent Ng)

Knowledge-based Pronominal Coreference

- [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 (1972)

- See: Hector J. Levesque "On our best behaviour" IJCAI 2013. http://www.cs.toronto.edu/~hector/Papers/ijcai-13-paper.pdf
  - Winograd Schema Challenge @ CommonSense 2015
    - http://commonsensereasoning.org/winograd.html

Hobbs' algorithm: commentary

"… *the naïve approach is quite good. Computationally speaking, it will be a long time before a semantically based algorithm is sophisticated enough to perform as well, and these results set a very high standard for any other approach to aim for.*

"Yet there is every reason to pursue a semantically based approach. The naïve algorithm does not work. Any one can think of examples where it fails. In these cases it not only fails; it gives no indication that it has failed and offers no help in finding the real antecedent."

— Hobbs (1978), *Lingua*, p. 345

Plan

1. **Evaluation of Coreference** [5+5 mins]
2. Introduction to machine learning approaches to coreference [15 mins]
3. Feature-based discriminative classifiers [15 mins]
4. Feature-based softmax/maxent linear classifiers [20 mins]
5. Different conceptualizations of coreference as a machine learning task [15 mins]

Coreference Evaluation

- B-CUBED algorithm for evaluation
  - Shading = gold standard, circles = system clustering

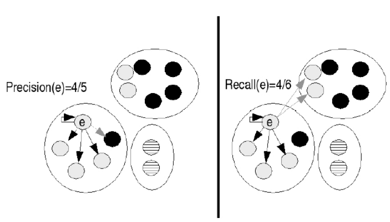


Figure from Amigo et al 2009

Evaluation

- B^3 (B-CUBED) algorithm for evaluation
  - Precision & recall for *entities in a reference chain*
  - Precision (P): % of elements in a hypothesized reference chain that are in the true reference chain
  - Recall (R): % of elements in a true reference chain that are in the hypothesized reference chain
  - Overall precision & recall are the (perhaps weighted) average of per-chain precision & recall
  - Optimizing chain-chain pairings is a hard problem
    - In the computational NP-hard sense
  - Greedy matching is done in practice for evaluation
  - F1 measure is harmonic mean of P and R

Evaluation metrics

- MUC Score (Vilain et al., 1995)
  - Link based: Counts the number of common links and computes f-measure
- CEAF (Luo 2005); entity based, two variants
- BLANC (Recasens and Hovy 2011): Cluster RAND-index
- …

- All of them are sort of evaluating getting coreference links/clusters right and wrong, but the differences can be important
  - Look at it in PA3

Plan

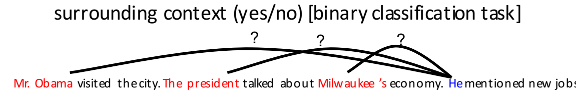
1. Evaluation of Coreference [5+5 mins]
2. **Introduction to machine learning approaches to coreference** [15 mins]
3. Feature-based discriminative classifiers [15 mins]
4. Feature-based softmax/maxent linear classifiers [20 mins]
5. Different conceptualizations of coreference as a machine learning task [15 mins]

Machine learning models of coref

- Start with supervised data
  - positive examples that corefer
  - negative examples that don't corefer
  - Note that it's very skewed
    - The vast majority of mention pairs *don't* corefer
- Usually learn some sort of discriminative classifier for phrases/clusters coreferring
  - Predict 1 for coreference, 0 for not coreferent
- But there is also work that builds clusters of coreferring expressions
  - E.g., generative models of clusters in (Haghighi & Klein 2007)

Supervised Machine Learning Pronominal Anaphora Resolution

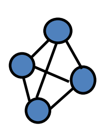
- Given a pronoun and an entity mentioned earlier, classify whether the pronoun refers to that entity or not given the surrounding context (yes/no) [binary classification task]



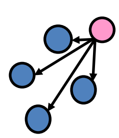
- Usually first filter out pleonastic pronouns like "It is raining." (perhaps using hand-written rules)
- Use any classifier, get "yes" examples from training data, make "no" examples by pairing pronoun with other (wrong) entities
- Decision rule: take nearest mention classified "yes", if any.

Kinds of Coref Models

- Mention Pair models
  - Treat coreference chains as a collection of pairwise links
  - Make independent pairwise decisions and reconcile them in some way (e.g. clustering or greedy partitioning)



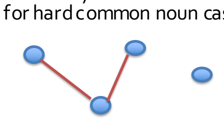
- Mention ranking models
  - Explicitly rank all candidate antecedents for a mention



- Entity-Mention models
  - A cleaner, but less studied, approach
  - Posit single underlying entities
  - Each mention links to a discourse entity [Pasula et al. 03], [Luo et al. 04]

Mention Pair Models

- Most common machine learning approach
- Build a binary classifier over pairs of mentions
  - For each mention, pick a preceding mention or NEW
  - Or, for each antecedent candidate, choose link or no-link
- Clean up non-transitivity with clustering or graph partitioning algorithms
  - E.g.: [Soon et al. 01], [Ng and Cardie 02]
  - Some work has done the classification and clustering jointly [McCallum and Wellner 03]
- Failures are mostly because of insufficient knowledge or features for hard common noun cases



Features: Grammatical Constraints

Are the two mentions in a coreference grammatical relationship?

- Apposition
  - Nefertiti, Amenomfis the IVth's wife, was born in …
- Predicatives/equatives
  - Sue is the best student in the class
  - It's questionable whether predicative cases should be counted, but they generally are.

Features: Soft Discourse Constraints

- Recency
- Salience
- Focus
- Centering Theory [Grosz et al. 86]
- Coherence Relations

Other coreference features

- Additional features to incorporate aliases, variations in names etc., e.g. Mr. Obama, Barack Obama; Megabucks, Megabucks Inc.
- Semantic Compatibility
  - Smith had bought *a used car* that morning.
    - *The dealership* assured him it was in good condition.
    - *The machine* needed a little love, but the engine was in good condition.

But it's complicated … so weight features

- Common nouns can differ in number but be coreferent:
  - a patrol … the soldiers
- Common nouns can refer to proper nouns
  - George Bush … the leader of the free world
- Gendered pronouns can refer to inanimate things
  - India withdrew her ambassador from the Commonwealth
- Split antecedence
  - John waited for Sasha. And then they went out.

Pairwise Features

1. **strict gender [true or false]**. True if there is a strict match in gender (e.g. male pronoun Pro_i with male antecedent NP_j).
2. **compatible gender [true or false]**. True if Pro_i and NP_j are merely compatible (e.g. male pronoun Pro_i with antecedent NP_j of unknown gender).
3. **strict number [true or false]** True if there is a strict match in number (e.g. singular pronoun with singular antecedent)
4. **compatible number [true or false]**. True if Pro_i and NP_j are merely compatible (e.g. singular pronoun Pro_i with antecedent NP_j of unknown number).
5. **sentence distance [0, 1, 2, 3,...]**. The number of sentences between pronoun and potential antecedent.
6. **Hobbs distance [0, 1, 2, 3,...]**. The number of noun groups that the Hobbs algorithm has to skip, starting backwards from the pronoun Pro_i, before the potential antecedent NP_j is found.
7. **grammatical role [subject, object, PP]**. Whether the potential antecedent is a syntactic subject, direct object, or is embedded in a PP.
8. **linguistic form [proper, definite, indefinite, pronoun]**. Whether the potential antecedent NP_j is a proper name, definite description, indefinite NP, or a pronoun.

Pairwise Features

Category	Features	Remark
Lexical	exact_strm	1 if two mentions have the same spelling; 0 otherwise
	left_subsm	1 if one mention is a left substring of the other; 0 otherwise
	right_subsm	1 if one mention is a right substring of the other; 0 otherwise
	acronym	1 if one mention is an acronym of the other; 0 otherwise
	edit_dist	quantized editing distance between two mention strings
	spell	pair of actual mention strings
	ncd	number of different capitalized words in two mentions
Distance	token_dist	how many tokens two mentions are apart (quantized)
	sent_dist	how many sentences two mentions are apart (quantized)
	gap_dist	how many mentions in between the two mentions in question (quantized)
Syntax	POS_pair	POS-pair of two mention heads
	apposition	1 if two mentions are appositive; 0 otherwise
Count	count	pair of (quantized) numbers, each counting how many times a mention string is seen
Pronoun	gender	pair of attributes of {female, male, neutral, unknown }
	number	pair of attributes of {singular, plural, unknown}
	possessive	1 if a pronoun is possessive; 0 otherwise
	reflexive	1 if a pronoun is reflexive; 0 otherwise

[Luo et al. 04]

Plan

1. Evaluation of Coreference [5+5 mins]
2. Introduction to machine learning approaches to coreference [15 mins]
3. Feature-based discriminative classifiers [15 mins]
4. Feature-based softmax/maxent linear classifiers [20 mins]
5. **Different conceptualizations of coreference as a machine learning task** [15 mins]

Mention-Pair (MP) Model

- Soon et al. 2001 ; Ng and Cardie 2002
- Classifies whether **two mentions** are coreferent or not.
- Weaknesses
  – Insufficient information to make an informed coreference decision.

20

Mention-Pair (MP) Model

- Soon et al. 2001 ; Ng and Cardie 2002
- Classifies whether **two mentions** are coreferent or not.
- Weaknesses
  – Insufficient information to make an informed coreferenced decision.

Barack Obama ……..……..Hillary Rodham Clinton …….his

………. **secretary of state** ……..…..……..…..He …….…...her

21

Mention-Pair (MP) Model

- Soon et al. 2001 ; Ng and Cardie 2002
- Classifies whether **two mentions** are coreferent or not.
- Weaknesses
  – Insufficient information to make an informed coreference decision.
  – Each candidate antecedent is considered independently of the others.

22

Mention-Pair (MP) Model

- Soon et al. 2001 ; Ng and Cardie 2002
- Classifies whether **two mentions** are coreferent or not.
- Weaknesses
  – Insufficient information to make an informed coreferenced decision.
  – Each candidate antecedent is considered independently of the others.

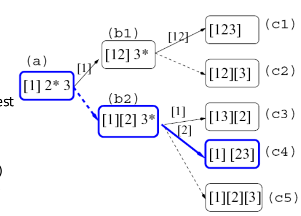
Barack Obama ………Hillary Rodham Clinton …….his ………..

secretary of state ………..the President……..He …..……..**her**

23

An Entity Mention Model

- Example: [Luo et al. 04]
- Bell Tree (link vs. start decision list)
- Entity centroids, or not?
  – Not for [Luo et al. 04], see [Pasula et al. 03]
  – Some features work on nearest mention (e.g. recency and distance)
  – Others work on "canonical" mention (e.g. spelling match)
  – Lots of pruning, model highly approximate
  – (Actually ends up being like a greedy-link system in the end)



Entity-Mention (EM) Model

- Pasula et al. 2003 ; Luo et al. 2004 ; Yang et al. 2004
- Classifies whether **a mention** and **a preceding, possibly partially formed cluster** are coreferent or not.
- Strength
  - – Improved expressiveness.
    - – Allows the computation of cluster level features
- Weakness
  - – Each candidate cluster is considered independently of the others.

Barack Obama ………………Hillary Rodham Clinton …….his
……….. secretary of state ……………………..He ………...her

25

Mention-Ranking (MR) Model

- Denis & Baldridge 2007, 2008
- Imposes a **ranking** on a set of candidate antecedents
- Strength
  - – Considers all the candidate antecedents simultaneously
- Weakness
  - – Insufficient information to make an informed coreference decision.

Barack Obama ………………Hillary Rodham Clinton …….his
……….. secretary of state ……………………..He ………...her

26

First Ranking Mention Model

- Actually, we don't need a ranking on all candidate antecedents
- We can just find the **highest** ranking antecedent
- This is equivalent to multiclass classification:
  - Choose the antecedent
  - But without a fixed set of classes
    - **structured prediction**
- Used in recent (high-performing) paper of Durrett and Klein (EMNLP 2013)
  - They use a maxent/softmax model just as we have been discussing

27