Supervised Models for Coreference Resolution

Altaf Rahman and Vincent Ng

Human Language Technology Research Institute

University of Texas at Dallas

What is Coreference Resolution ?

– Identify all noun phrases (**mentions**) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affair experience as a former First Lady.

What is Coreference Resolution ?

– Identify all noun phrases (**mentions**) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affair experience as a former First Lady.

What is Coreference Resolution ?

– Identify all noun phrases (**mentions**) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his

secretary of state on Monday

had foreign affair experience

What is Coreference Resolution ?

– Identify all noun phrases (**mentions**) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affair experience as a former First Lady.

What is Coreference Resolution ?

– Identify all noun phrases (**mentions**) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chos
had foreign affair experience as a form

Plan for the talk

- Existing learning based coreference models
  - Overview
  - Implementation details
- Our cluster ranking model
- Evaluation
- Summary

Plan for the talk

- Existing learning based coreference models
  - Overview
  - Implementation details
- Our cluster ranking model
- Evaluation
- Summary

Existing learning based coreference models

- Mention-Pair (MP) model
- Entity-Mention (EM) model
- Mention-Ranking (MR) model

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.

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

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 antecedents is considered independently of the others.

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 antecedents is considered independently of the others.

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

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

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

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

Goal

- Propose a cluster ranking (CR) model
  - **ranks** all the preceding **clusters** for a mention
  - combines the strengths of EM and MR models
    - Improve expressiveness by using cluster level features.
    - Considers all the candidate clusters simultaneously

Plan for the talk

- Existing learning based coreference models
  - Overview
  - Implementation details
- Our cluster ranking model
- Evaluation
- Summary

Mention-Pair (MP) Model

- Classifies whether two mentions are coreferent or not
- Training
  - Each instance is created between m_j and m_k
  - 39 features

Mention-Pair (MP) Model

- Classifies whether two mentions are coreferent or not
- Training
  - Each instance is created between m_j and m_k
  - 39 features
    - Features describing m_j a candidate antecedent
      - **Pronoun ? Subject ? Nested ?**

Mention-Pair (MP) Model

- Classifies whether two mentions are coreferent or not
- Training
  - Each instance is created between m_j and m_k
  - 39 features
    - Features describing m_j a candidate antecedent
      - **Pronoun ? Subject ? Nested ?**
    - Features describing m_k, the mention to be resolved
      - **Number ? Gender ? Pronoun2 ? Semantic Class ? Animacy ?**

Mention-Pair (MP) Model

- Classifies whether two mentions are coreferent or not
- Training
  - Each instance is created between m_j and m_k
  - 39 features
    - Features describing m_j a candidate antecedent
      - **Pronoun ? Subject ? Nested ?**
    - Features describing m_k, the mention to be resolved
      - **Number ? Gender ? Pronoun2 ? Semantic Class ? Animacy ?**
    - Features describing the relation between m_j a candidate antecedent m_k the mention to be resolved
      - **Head match ? String match ? Gender match ? Span ? Appositive ? Alias ? Distance ?**

Mention-Pair (MP) Model

- Training instance creation (Soon et al.)
  - create
    - positive instance for each anaphoric mention, m_j and its closest preceding antecedent mention, m_i
    - negative instance for m_j and each intervening mention, m_{i+1}, m_{i+2} , … , m_{j-1}
    - No instance for non-anaphors.

Mention-Pair (MP) Model

- Training instance creation (Soon et al.)
  - create
    - positive instance for each anaphoric mention, m_j, and its closest preceding antecedent mention, m_i
    - negative instance for m_j and each intervening mention, m_{i+1}, m_{i+2} , … , m_{j-1}
    - No instance for non-anaphors.
- Testing (Soon et al.)
  - For each m_j
    - Select as the antecedent of m_j the closest preceding mention that is classified as the coreferent with m_j
    - if no such mention exist m_j is considered non-anaphoric

Entity-Mention (EM) Model

- Classifies whether a mention and a preceding cluster are coreferent or not.
- Training
  - Each instance is between a mention and a preceding partially formed cluster.
  - Cluster level features

Entity-Mention (EM) Model

- Training instance creation:
  - Positive Instance
    - For each anaphoric mention m_k and preceding cluster c_j to which it belongs

Entity-Mention (EM) Model

- Training instance creation:
  - Positive Instance
    - For each anaphoric mention m_k and preceding cluster c_j to which it belongs

  - No instance for non-anaphors.

Entity-Mention (EM) Model

- Training instance creation:
  - Positive Instance
    - For each anaphoric mention m_k and preceding cluster c_j to which it belongs
  - No instance for non-anaphors.
  - Negative Instance
    - For each anaphoric mention m_k and partial cluster whose *last mention* appears between m_k and its closest antecedent in c_j to which it belongs

Entity-Mention (EM) Model continued

………m_1……… m_2…….. m_3…… m_4………m_5…………m_6

For mention m_6

- Positive instance :
  - features between m_6 and the cluster {m_2, m_4}

Entity-Mention (EM) Model continued

.........m_1......... m_2........ m_3...... m_4.........m_5............m_6

For mention m_6

- Positive instance :
  - features between m_6 and the cluster $\{m_2, m_4\}$
- Negative instance :
  - features between m_6 and the cluster $\{m_1, m_5\}$

Entity-Mention (EM) Model continued

………m_1……… m_2…….. m_3…… m_4………m_5…………m_6

For mention m_6

- Positive instance :
  - features between m_6 and the cluster $\{m_2, m_4\}$
- Negative instance :
  - features between m_6 and the cluster $\{m_1, m_5\}$
  - No negative instance created between m_6 and $\{m_3\}$

Entity-Mention (EM) Model

- Testing
  - Like MP model except now we resolve the mention to the closest **preceding cluster** that is classified as coreferent.

Entity-Mention(EM) Model continued

- For each relational feature used by MP model we create a set of **cluster level** features
  - Example : Gender = Compatible ? Incompatible ? Not Applicable?

Entity-Mention(EM) Model continued

- For each relational feature used by MP model we create a set of **cluster level** features
  - Example : Gender = Compatible ? Incompatible ? Not Applicable?

Hillary Clinton….	.whose…………..…	she…………	him…….
(female)	*(neutral)*	*(female)*	*(male)*

For the mention "him"

– #C = 0   Normalized_C = 0/3 = 0.00

#I = 2   Normalized_I = 2/3 = 0.66

#NA = 1   Normalized_NA = 1/3 = 0.33

Entity-Mention(EM) Model continued

- For each relational feature used by MP model we create a set of **cluster level** features
  - Example : Gender = Compatible ? Incompatible ? Not Applicable?

Hillary Clinton….	.whose………..….she………….		him…….
(female)	*(neutral)*	*(female)*	*(male)*

For the mention "him"

- #C = 0 Normalized_C = 0/3 = 0.00
  #I = 2 Normalized_I = 2/3 = 0.66
  #NA = 1 Normalized_NA = 1/3 = 0.33

0.0 = *NONE*
<0.5 = *MOST-FALSE*
>0.5 = *MOST-TRUE*
1.0 = *ALL*

Entity-Mention(EM) Model continued

- For each relational feature used by MP model we create a set of **cluster level** features
  - Example : Gender = Compatible ? Incompatible ? Not Applicable?

Hillary Clinton....	.whose…………..….	she…………..	him…….
(female)	*(neutral)*	*(female)*	*(male)*

For the mention "him"

- #C = 0 Normalized_C = 0/3 = 0.00
  #I = 2 Normalized_I = 2/3 = 0.66
  #NA = 1 Normalized_NA = 1/3 = 0.33

  0.0 = *NONE*
  <0.5 = *MOST-FALSE*
  >0.5 = *MOST-TRUE*
  1.0 = *ALL*

- GenderC=*NONE* GenderI=*MOST-TRUE* GenderNA=*MOST-FALSE*

Mention-Ranking (MR) Model

- Ranks a set of candidate antecedents for each mention
- Training
  - Each instance represents 2 mentions (m_j, m_k)
  - Same 39 features as in Mention-pair (MP) model
  - Used the SVM ranker learning algorithm (Joachims 2002).

Mention-Ranking (MR) Model continued

- Training instance creation
  - According to Soon et al's method
    - Rank value is 2 if positive
    - Otherwise rank 1

Mention-Ranking (MR) Model continued

- Training instance creation
  - According to Soon et al's method
    - Rank value is 2 if positive
    - Otherwise rank 1
- Testing
  - First check anaphoricity of m_j using separate anaphoricity classifier.
    - If m_j is non-anaphoric then create a new cluster.
    - Otherwise, resolve m_j to the highest ranked m_k among ALL the candidate antecedents.

Plan for the talk

- Existing learning based coreference models
  - Overview
  - Implementation details
- Our cluster ranking model
- Evaluation
- Summary

Cluster Ranking(CR) model

- Combines the strength of MP and EM model
- Ranks all the preceding clusters for each mention
- Training
  - Each instance is comprised of features between a mention m_k and its preceding cluster c_j
  - Instances are created like the EM model.
  - Rank values are assigned like the MR model.

Cluster Ranking (CR) Model

- Since no instances were created from the non-anaphors, we need to rely on a separate classifier is to determine whether a mention is anaphoric
- Problem
  - Errors in anaphoricity determination can be propagated to coreference resolution
- Hypothesis
  - A model for jointly determining anaphoricity and coreference resolution can overcome this problem with the pipeline approach

Joint anaphoricity determination and coreference resolution

- Idea
  - Create additional training instances from non-anaphors
  - If m_k is non-anaphoric, assign rank value
    - 1 to each instance formed between m_k and each preceding cluster
    - 2 to all instances formed between m_k and **a (hypothetical) null** cluster
      - Use only features describing m_k
- Same idea can be applied to create joint version of MR model

Plan for the talk

- Existing learning based coreference models
  - Overview
  - Implementation details
- Our cluster ranking model
- **Evaluation**
- Summary

Evaluation

- Experimental setup
  - ACE2005 corpus
  - 599 documents of 7 sources -BC, BN, CTS, NW, UN, WL
  - 80% for Training and 20% for Testing.
  - True mentions
  - System mentions (extracted by a learned mention extractor that is trained on train docs)
  - Scoring programs
    - MUC (Vilain et al. 1995)
    - CEAF (Lu et al. 2005)
    - B^3 (Bagga & Baldwin 1998)
  - Recall, precision and f-measure

System Mention Results

- Baseline Systems
  - MP model
  - EM model
  - MR model

All models are trained using SVM-light. All learning parameters are set to default values

System Mention Results(MP baseline)

Coreference Models	CEAF			B^3		
	Recall	Precision	F	Precision	Recall	F
MP model	56.1	51.0	53.4	50.8	57.9	54.1

- CEAF F score is 53.4
- B^3 F score is 54.1

System Mention Results (EM baseline)

Coreference Models	CEAF			B^3		
	Recall	Precision	F	Precision	Recall	F
MP model	56.1	51.0	53.4	50.8	57.9	54.1
EM model	56.3	50.2	53.1	51.2	57.8	54.3

- F score change is insignificant despite the improved expressiveness of EM model.
- Similar trends have been reported by Luo et al 2004.

System Mention Results(MR baseline)

Coreference Models	CEAF			B^3		
	Recall	Precision	F	Precision	Recall	F
MP model	56.1	51.0	53.4	50.8	57.9	54.1
EM model	56.3	50.2	53.1	51.2	57.8	54.3
MR model (pipeline)	51.6	56.7	54.1	52.3	61.8	56.6
MR model (joint)	53.0	58.5	55.6	50.4	65.5	56.9

- 2 architectures for using anaphoricity information
  - Pipeline
  - Joint

System Mention Results(MR baseline)

Coreference Models	CEAF			B^3		
	Recall	Precision	F	Precision	Recall	F
MP model	56.1	51.0	**53.4**	50.8	57.9	**54.1**
EM model	56.3	50.2	53.1	51.2	57.8	54.3
MR model (pipeline)	51.6	56.7	**54.1**	52.3	61.8	**56.6**
MR model (joint)	53.0	58.5	**55.6**	50.4	65.5	**56.9**

- 2 architectures for using anaphoricity information
  - Pipeline
  - Joint
- Both show significant improvements over MP baseline

System Mention Results(MR baseline)

Coreference Models	CEAF			B^3		
	Recall	Precision	F	Precision	Recall	F
MP model	56.1	51.0	53.4	50.8	57.9	54.1
EM model	56.3	50.2	53.1	51.2	57.8	54.3
MR model (pipeline)	51.6	56.7	**54.1**	52.3	61.8	**56.6**
MR model (joint)	53.0	58.5	**55.6**	50.4	65.5	**56.9**

- 2 architectures for using anaphoricity information
  - Pipeline
  - Joint
- Both show significant improvements over MP baseline
- Joint architecture outperforms pipeline architecture.

System Mention Results (CR model)

Coreference Models	CEAF			B^3		
	Recall	Precision	F	Precision	Recall	F
MP model	56.1	51.0	53.4	50.8	57.9	54.1
EM model	56.3	50.2	53.1	51.2	57.8	54.3
MR model (pipeline)	51.6	56.7	54.1	52.3	61.8	56.6
MR model (joint)	53.0	58.5	55.6	50.4	65.5	56.9
CR model (pipeline)	54.1	59.3	**56.6**	55.3	63.7	**59.2**
CR model (joint)	56.7	62.6	**59.5**	54.4	70.5	**61.4**

System Mention Results (CR model)

Coreference Models	CEAF			B^3		
	Recall	Precision	F	Precision	Recall	F
MP model	56.1	51.0	53.4	50.8	57.9	54.1
EM model	56.3	50.2	53.1	51.2	57.8	54.3
MR model (pipeline)	51.6	56.7	**54.1**	52.3	61.8	**56.6**
MR model (joint)	53.0	58.5	55.6	50.4	65.5	56.9
CR model (pipeline)	54.1	59.3	**56.6**	55.3	63.7	**59.2**
CR model (joint)	56.7	62.6	59.5	54.4	70.5	61.4

- Cluster-ranking model outperforms Mention-ranking model
  - Due to simultaneous gains in recall and precision

System Mention Results (CR model)

Coreference Models	CEAF			B^3		
	Recall	Precision	F	Precision	Recall	F
MP model	56.1	51.0	53.4	50.8	57.9	54.1
EM model	56.3	50.2	53.1	51.2	57.8	54.3
MR model (pipeline)	51.6	56.7	54.1	52.3	61.8	56.6
MR model (joint)	53.0	58.5	**55.6**	50.4	65.5	**56.9**
CR model (pipeline)	54.1	59.3	56.6	55.3	63.7	59.2
CR model (joint)	56.7	62.6	**59.5**	54.4	70.5	**61.4**

- Cluster-ranking model outperforms Mention-ranking model
  - Due to simultaneous gains in recall and precision

System Mention Results (CR model)

Coreference Models	**CEAF**			**B^3**		
	Recall	Precision	F	Precision	Recall	F
MP model	56.1	51.0	53.4	50.8	57.9	54.1
EM model	56.3	50.2	53.1	51.2	57.8	54.3
MR model (pipeline)	51.6	56.7	54.1	52.3	61.8	56.6
MR model (joint)	53.0	58.5	55.6	50.4	65.5	56.9
CR model (pipeline)	54.1	59.3	**56.6**	55.3	63.7	**59.2**
CR model (joint)	56.7	62.6	**59.5**	54.4	70.5	**61.4**

- Joint architecture outperforms pipeline architecture.
  - Due to simultaneous gains in recall and precision

Plan for the talk

- Existing learning based coreference models
  - Overview
  - Implementation details
- Our cluster ranking model
- Evaluation
- Summary

Summary

- Proposed a cluster ranking approach
  - Combines the strengths of EM and MR models.
  - Jointly learns coreference resolution and anaphoricity determination
  - Significantly outperforms three commonly-used learning-based coreference models