Unsupervised Models for Coreference Resolution

Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas

Coreference

- Identify the noun phrases (or *mentions*) that refer to the same real-world entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist, was summoned to help the King overcome his speech impediment...

Coreference

- Identify the noun phrases (or *mentions*) that refer to the same real-world entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist, was summoned to help the King overcome his speech impediment...

Coreference

- Identify the noun phrases (or *mentions*) that refer to the same real-world entity

> Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist, was summoned to help the King overcome his speech impediment...

Coreference

- Identify the noun phrases (or *mentions*) that refer to the same real-world entity

> Queen Elizabeth set about transforming her husband, King George VI, into <u>a viable monarch</u>. <u>A renowned speech therapist</u>, was summoned to help the King overcome his <u>speech impediment</u>...

Coreference

- Identify the noun phrases (or *mentions*) that refer to the same real-world entity

> Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist, was summoned to help the King overcome his speech impediment...

Coreference

- Identify the noun phrases (or *mentions*) that refer to the same real-world entity

Queen Elizabeth set about transforming her husband, King George VI, into <u>a viable monarch</u>. <u>A renowned speech therapist</u>, was summoned to help the King overcome his <u>speech impediment</u>...

- Lots of prior work on supervised coreference resolution
  - Soon et al. (2001), Strube et al. (2002), Yang et al. (2003), Luo et al. (2004), Denis and Baldridge (2007), …

Unsupervised Coreference Resolution

Perform coreference resolution using little or no annotated data

Previous Work

- Apply a weakly supervised or unsupervised learning algorithm to pronoun resolution
  - co-training (Müller et al., 2002)
  - self-training (Kehler et al., 2004)
  - EM (Cherry and Bergsma, 2005)

Previous Work

- Apply a weakly supervised or unsupervised learning algorithm to pronoun resolution
  - co-training (Müller et al., 2002)
  - self-training (Kehler et al., 2004)
  - EM (Cherry and Bergsma, 2005)
- A nonparametric fully-Bayesian approach to unsupervised coreference resolution (Haghighi and Klein, 2007)

Goals

- Design a new model for unsupervised coreference resolution
- Improve Haghighi and Klein's model with three modifications

Unsupervised Coreference as EM Clustering

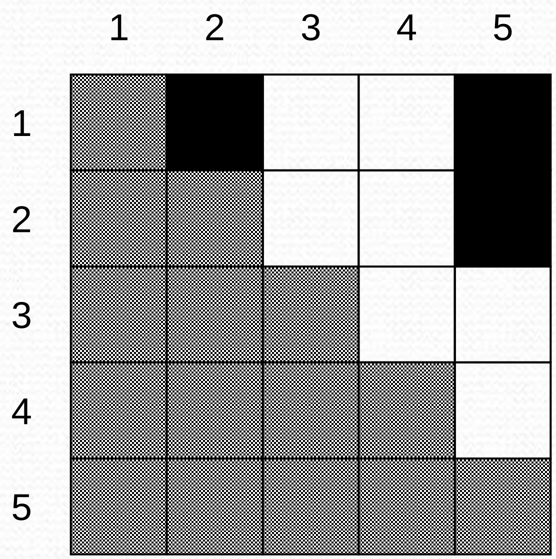
- Design a generative model that can be used to induce a clustering of the mentions in a given document

Representing a Clustering

- A clustering C of n mentions is an n x n Boolean matrix, where $C_{ij} = 1$ iff mentions i and j are coreferent

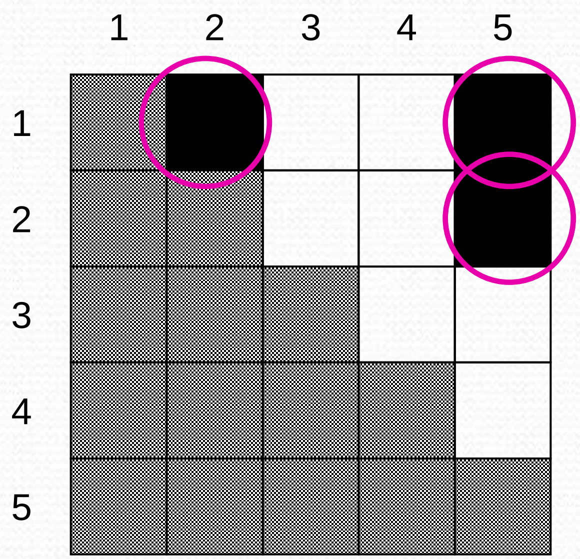
Representing a Clustering

- A clustering C of n mentions is an n x n Boolean matrix, where $C_{ij} = 1$ iff mentions i and j are coreferent



Representing a Clustering

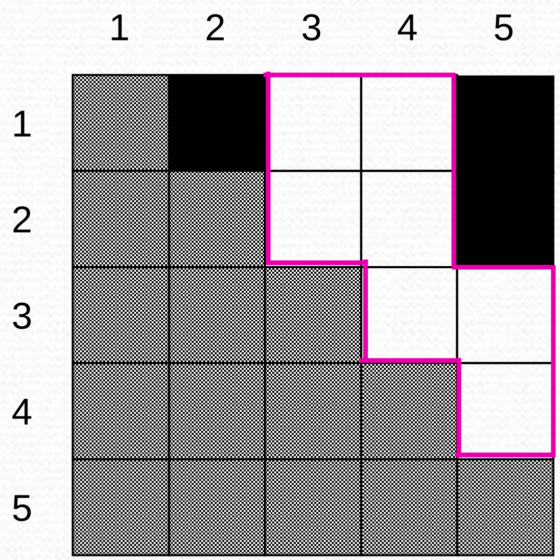
- A clustering C of n mentions is an n x n Boolean matrix, where $C_{ij} = 1$ iff mentions i and j are coreferent



Coreferent

Representing a Clustering

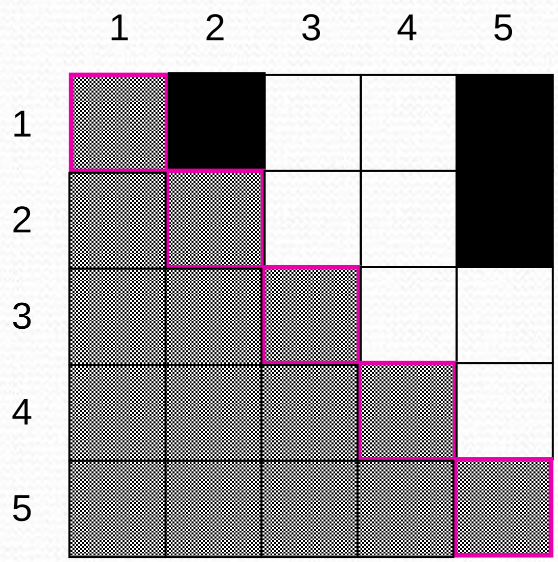
- A clustering C of n mentions is an n x n Boolean matrix, where $C_{ij} = 1$ iff mentions i and j are coreferent



Not Coreferent

Representing a Clustering

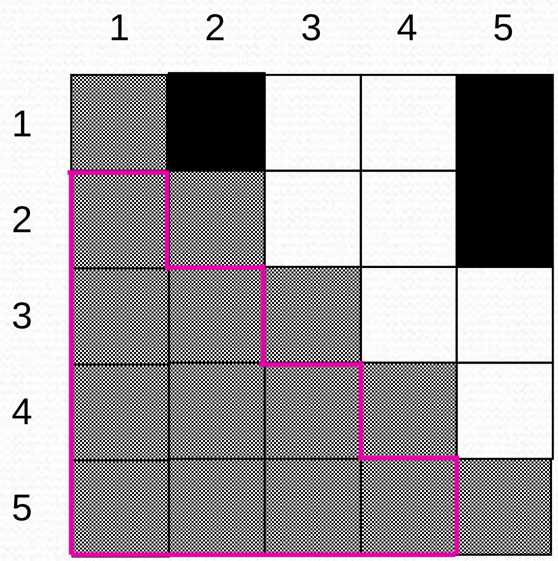
- A clustering C of n mentions is an n x n Boolean matrix, where $C_{ij} = 1$ iff mentions i and j are coreferent



Don't care about diagonal entries

Representing a Clustering

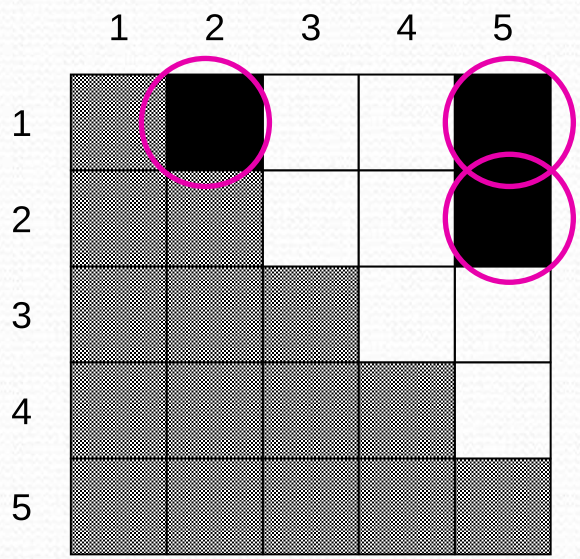
- A clustering C of n mentions is an n x n Boolean matrix, where $C_{ij} = 1$ iff mentions i and j are coreferent



Don't care about entries below the diagonal

Representing a Clustering

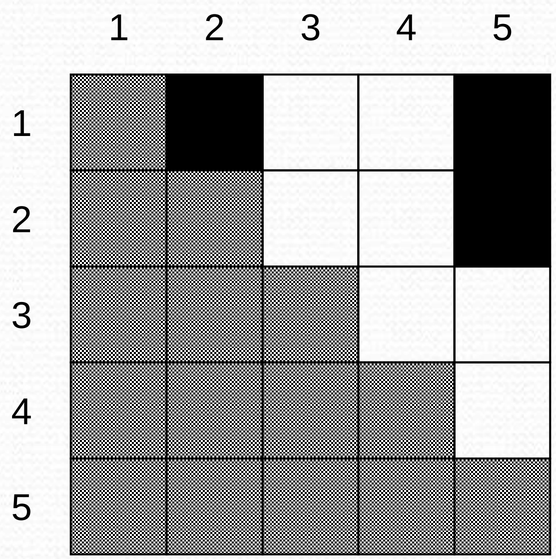
- A clustering C of n mentions is an n x n Boolean matrix, where C_{ij} = 1 iff mentions i and j are coreferent



Transitive

Representing a Clustering

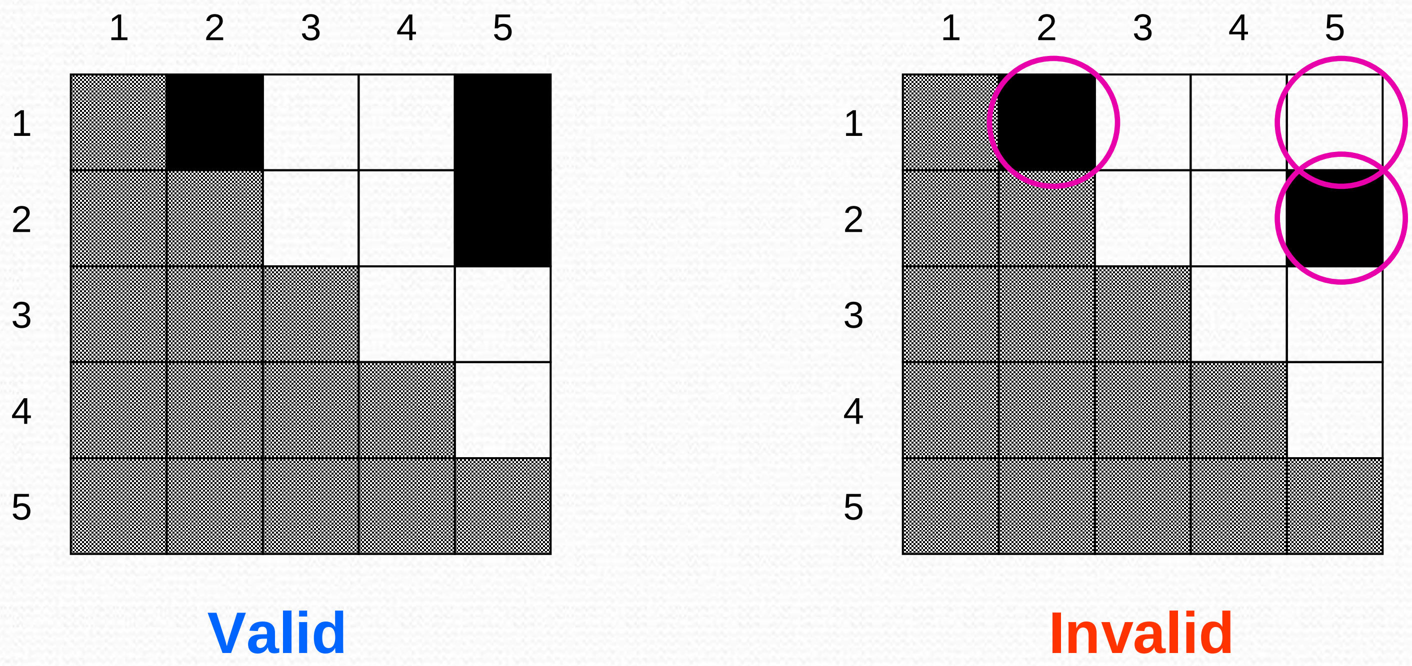
- A clustering C of n mentions is an n x n Boolean matrix, where $C_{ij} = 1$ iff mentions i and j are coreferent



Valid

Representing a Clustering

- A clustering C of n mentions is an n x n Boolean matrix, where $C_{ij} = 1$ iff mentions i and j are coreferent



The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$P(D,C)=P(C)\,P(D|C)$$

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$P(D,C)=P(C)P(D|C)$$

How to generate D given C?

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$P(D,C) = P(C)\,P(D|C)$$

How to generate D given C?

- Assume that D is represented by its mention pairs

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$P(D,C)=P(C)P(D|C)$$

How to generate D given C?

- Assume that D is represented by its mention pairs
- To generate D, generate all pairs of mentions in D
  - (Queen Elizabeth, her), (Queen Elizabeth, husband), (Queen Elizabeth, King George VI), …

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$
\begin{aligned}
P(D,C) &= P(C)\,P(D|C) \\
&= P(C)\,P(mp_{12}, mp_{13}, mp_{14}, \ldots |C)
\end{aligned}
$$

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

mp$_{ij}$ is the pair formed from mention i and mention j

$$P(D,C) = P(C)\,P(D|C)$$
$$= P(C)\,P(mp_{12}, mp_{13}, mp_{14}, \ldots |C)$$

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$P(D,C) = P(C)\,P(D|C)$$
$$= P(C)\,P(mp_{12}, mp_{13}, mp_{14}, \ldots |C)$$

Let's simplify this term

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$P(D,C) = P(C)\,P(D|C)$$
$$= P(C)\,P(mp_{12}, mp_{13}, mp_{14}, \ldots |C)$$

Let's simplify this term

- assume that each mention pair mp_{ij} is generated conditionally independently given C_{ij}

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$
\begin{aligned}
P(D,C) &= P(C)\,P(D|C) \\
&= P(C)\,P(mp_{12}, mp_{13}, mp_{14}, \ldots |C) \\
&= P(C) \prod_{Pairs(D)} P(mp_{ij}|C_{ij})
\end{aligned}
$$

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$
\begin{aligned}
P(D,C) &= P(C)\,P(D|C) \\
&= P(C)\,P(mp_{12}, mp_{13}, mp_{14}, \dots |C) \\
&= P(C)\prod_{Pairs(D)} P(mp_{ij}|C_{ij})
\end{aligned}
$$

How to represent a mention pair mp_{ij}?

Features

- Use 7 linguistic features divided into 3 groups

Strong Coreference Indicators	String match Appositive Alias (one is an acronym or abbreviation of the other)
Linguistic Constraints	Gender agreement Number agreement Semantic compatibility
Mention Type Pairs	(ti, tj), where ti, tj $\in$ { Pronoun, Name, Nominal }

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$
\begin{aligned}
P(D,C) &= P(C)\,P(D|C) \\
&= P(C)\,P(mp_{12}, mp_{13}, mp_{14}, \ldots |C) \\
&= P(C)\prod_{Pairs(D)} P(mp_{ij}|C_{ij})
\end{aligned}
$$

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$
\begin{aligned}
P(D,C) &= P(C)\,P(D|C) \\
&= P(C)\,P(mp_{12}, mp_{13}, mp_{14},\ldots|C) \\
&= P(C)\prod_{Pairs(D)} P(mp_{ij}|C_{ij}) \\
&= P(C)\prod_{Pairs(D)} P(mp_{ij}^{1}, mp_{ij}^{2},\ldots, mp_{ij}^{7}|C_{ij})
\end{aligned}
$$

7 feature values

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$
\begin{aligned}
P(D,C) &= P(C)\,P(D|C) \\
&= P(C)\,P(mp_{12}, mp_{13}, mp_{14}, \ldots |C) \\
&= P(C)\prod_{Pairs(D)} P(mp_{ij}|C_{ij}) \\
&= P(C)\prod_{Pairs(D)} P(mp_{ij}^{1}, mp_{ij}^{2}, \ldots, mp_{ij}^{7} |C_{ij})
\end{aligned}
$$

Let's simplify this term

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$P(D,C) = P(C)\,P(D|C)$$

$$= P(C)\,P(mp_{12}, mp_{13}, mp_{14}, \ldots |C)$$

$$= P(C)\prod_{Pairs(D)} P(mp_{ij}|C_{ij})$$

$$= P(C)\prod_{Pairs(D)} P(mp_{ij}^{1}, mp_{ij}^{2}, \ldots, mp_{ij}^{7} |C_{ij})$$

Let's simplify this term

- assume that feature values from different groups are conditionally independent of each other

The Generative Model

- Given a document D,
  - generate a clustering C according to P(C)
  - generate D given C

$$
\begin{aligned}
P(D,C) &= P(C)\,P(D|C) \\
&= P(C)\,P(mp_{12}, mp_{13}, mp_{14},\ldots |C) \\
&= P(C)\prod_{Pairs(D)} P(mp_{ij}|C_{ij}) \\
&= P(C)\prod_{Pairs(D)} P(mp_{ij}^{1}, mp_{ij}^{2},\ldots, mp_{ij}^{7}|C_{ij}) \\
&= P(C)\,P(mp_{ij}^{1}, mp_{ij}^{2}, mp_{ij}^{3}|C_{ij})\,P(mp_{ij}^{4}, mp_{ij}^{5}, mp_{ij}^{6}|C_{ij}) \\
&\quad P(mp_{ij}^{7}|C_{ij})
\end{aligned}
$$

Model Parameters

$$P(mp^1, mp^2, mp^3 \mid c)$$

$$P(mp^4, mp^5, mp^6 \mid c)$$

$$P(mp^7 \mid c)$$

mp^i are the feature values

$c \in$ { Coref, Not Coref }

Model Parameters

$$P(mp^1, mp^2, mp^3 | c)$$

$$P(mp^4, mp^5, mp^6 | c)$$

$$P(mp^7 | c)$$

mp^i are the feature values

$c \in$ { Coref, Not Coref }

Model Parameters

$$P(mp^1, mp^2, mp^3 | c)$$

$$P(mp^4, mp^5, mp^6 | c)$$

$$P(mp^7 | c)$$

mp^i are the feature values

$c \in$ { Coref, Not Coref }

Model Parameters

$$P(mp^1, mp^2, mp^3 | c)$$

$$P(mp^4, mp^5, mp^6 | c)$$

$$P(mp^7 | c)$$

mp^i are the feature values

$c \in$ { Coref, Not Coref }

Next step: use EM to iteratively

- estimate the model parameters
- probabilistically induce a clustering for a document

The Induction Algorithm

- Given a set of unlabeled documents

The Induction Algorithm

- Given a set of unlabeled documents
  - guess a clustering for each document according to P(C)

The Induction Algorithm

- Given a set of unlabeled documents
  - guess a clustering for each document according to P(C)

Initial labelings are
presumably noisy

The Induction Algorithm

- Given a set of unlabeled documents
  - guess a clustering for each document according to P(C)

  - estimate the model parameters based on the automatically labeled documents **(M-step)**
    - maximum likelihood estimation

The Induction Algorithm

- Given a set of unlabeled documents
  - guess a clustering for each document according to P(C)

  - estimate the model parameters based on the automatically labeled documents **(M-step)**
    - maximum likelihood estimation
  - assign a probability to each possible clustering of the mentions for each document **(E-step)**

The Induction Algorithm

- Given a set of unlabeled documents
  - guess a clustering for each document according to P(C)

  - estimate the model parameters based on the automatically labeled documents **(M-step)**
    - maximum likelihood estimation
  - assign a probability to each possible clustering of the mentions for each document **(E-step)**

3 mentions: 1, 2, 3

The Induction Algorithm

- Given a set of unlabeled documents
  - guess a clustering for each document according to P(C)
  - estimate the model parameters based on the automatically labeled documents **(M-step)**
    - maximum likelihood estimation
  - assign a probability to each possible clustering of the mentions for each document **(E-step)**

3 mentions: 1, 2, 3

[123]	[1][2][3]	[13][2]	[12][3]	[1][23]	+ invalid clusterings

The Induction Algorithm

- Given a set of unlabeled documents
  - guess a clustering for each document according to P(C)
  - estimate the model parameters based on the automatically labeled documents **(M-step)**
    - maximum likelihood estimation
  - assign a probability to each possible clustering of the mentions for each document **(E-step)**

3 mentions: 1, 2, 3

[123]	[1][2][3]	[13][2]	[12][3]	[1][23]	+ invalid clusterings
0.23	0.21	0.11	0.29	0.05	…

The Induction Algorithm

- Given a set of unlabeled documents
  - guess a clustering for each document according to P(C)

Iterate till convergence

- estimate the model parameters based on the automatically labeled documents **(M-step)**
  - maximum likelihood estimation
- assign a probability to each possible clustering of the mentions for each document **(E-step)**

3 mentions: 1, 2, 3

[123]	[1][2][3]	[13][2]	[12][3]	[1][23]	+ invalid clusterings
0.23	0.21	0.11	0.29	0.05	...

The Induction Algorithm

How to cope with the computational complexity of the E-step?

- Given a set of unlabeled documents
  - guess a clustering for each document according to P(C)

Iterate till convergence

- estimate the model parameters based on the automatically labeled documents **(M-step)**
  - maximum likelihood estimation
- assign a probability to each possible clustering of the mentions for each document **(E-step)**

3 mentions: 1, 2, 3

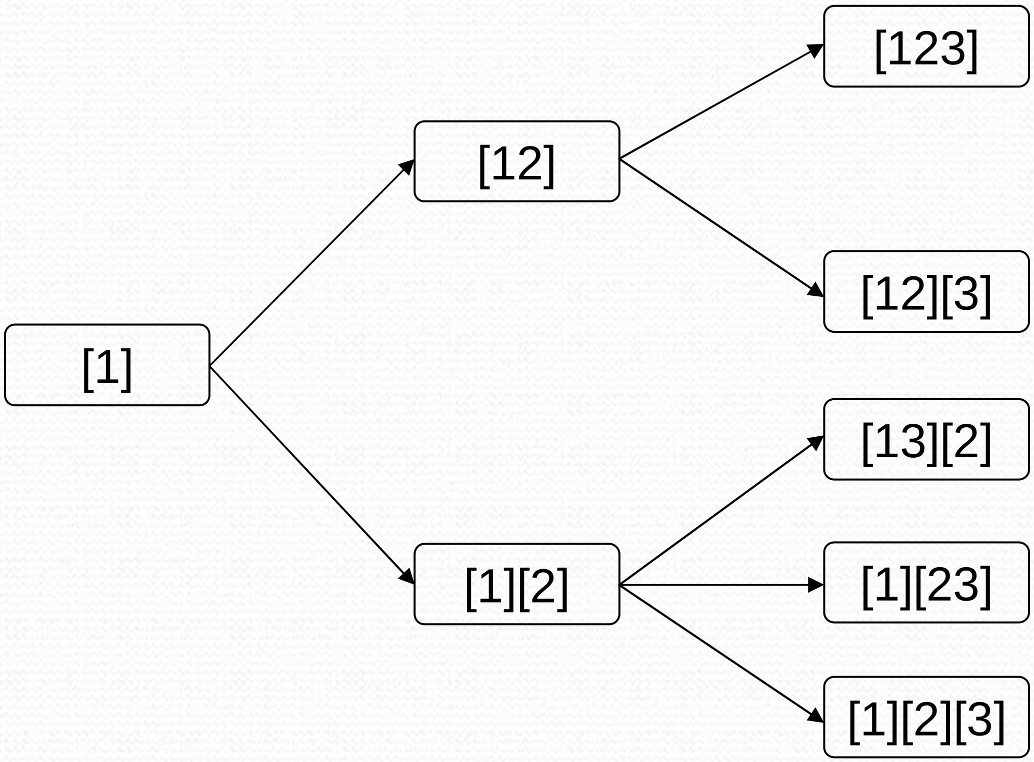
[123]	[1][2][3]	[13][2]	[12][3]	[1][23]	+ invalid clusterings
0.23	0.21	0.11	0.29	0.05	…

Approximating the E-step

- Search for the N most probable clusterings only
  - using Luo et al.'s (2004) search algorithm

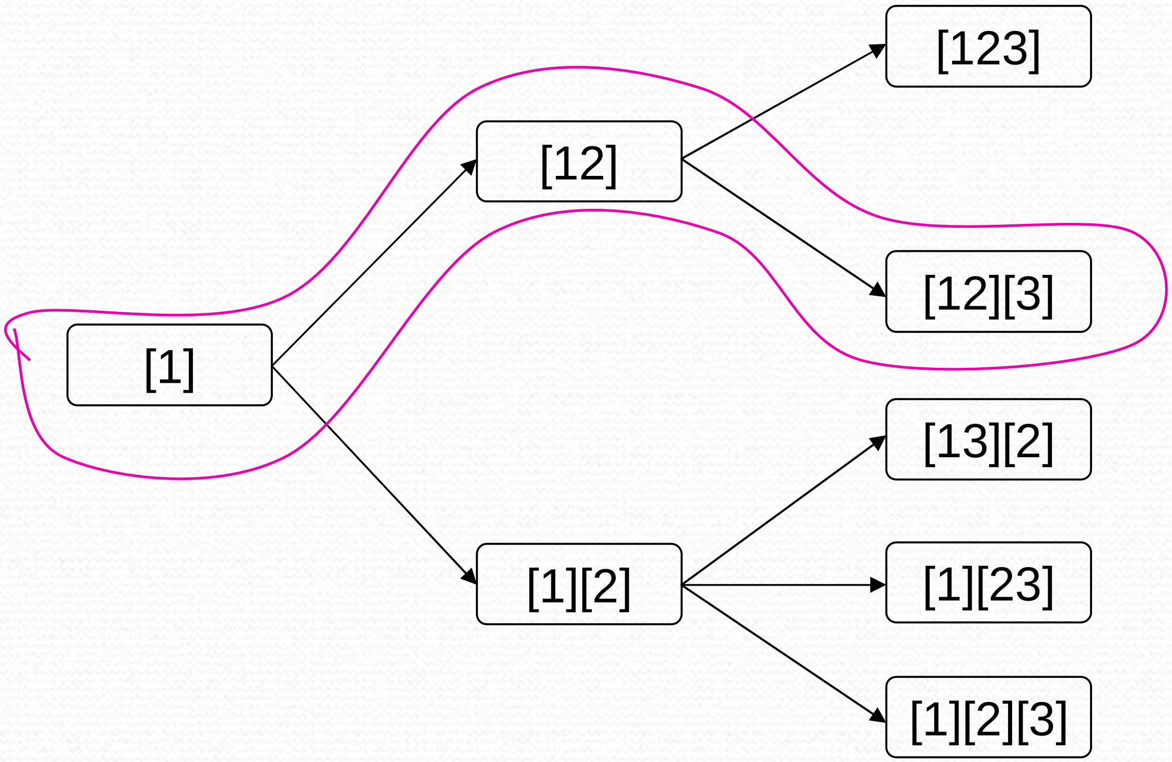
Approximating the E-step

- Search for the N most probable clusterings only
  - using Luo et al.'s (2004) search algorithm



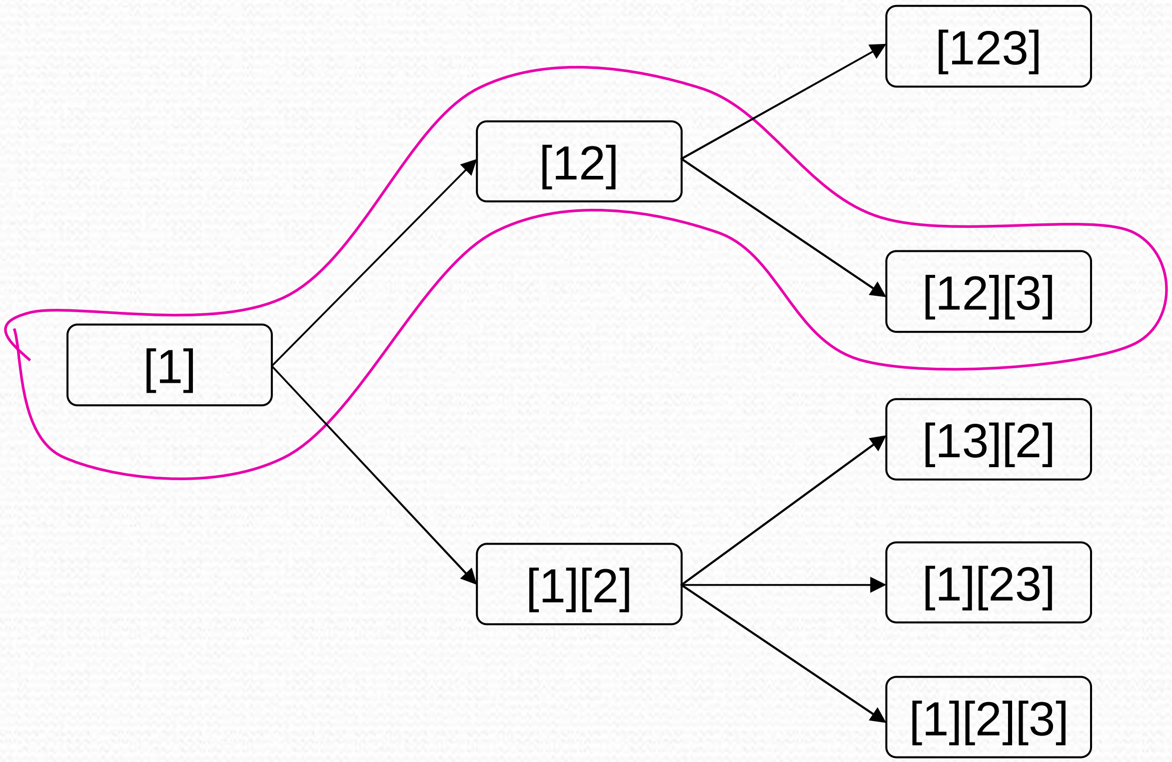
Approximating the E-step

- Search for the N most probable clusterings only
  - using Luo et al.'s (2004) search algorithm



Approximating the E-step

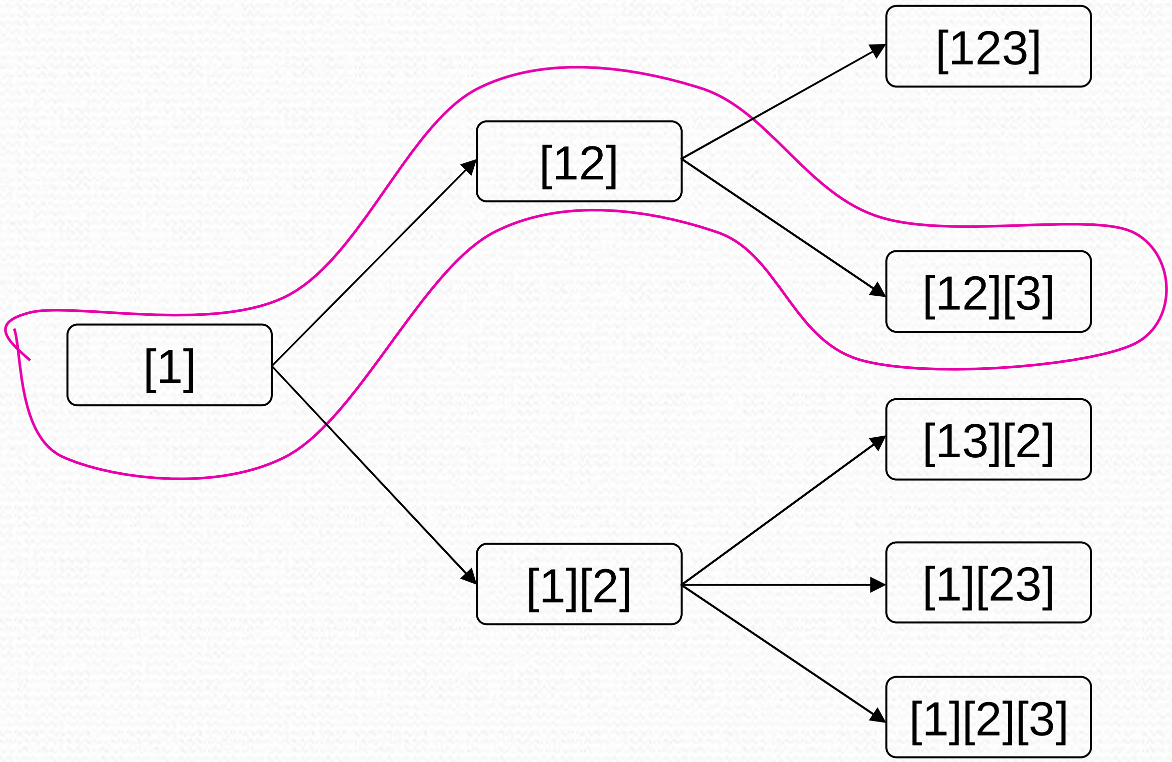
- Search for the N most probable clusterings only
  - using Luo et al.'s (2004) search algorithm



performs a beam search, expanding the most promising paths

Approximating the E-step

- Search for the N most probable clusterings only
  - using Luo et al.'s (2004) search algorithm



performs a beam search, expanding the most promising paths

scores a path based on pairwise coreference probabilities

The Induction Algorithm

- Given a set of unlabeled documents
  - guess a clustering for each document according to P(C)

Iterate till convergence

  - estimate the model parameters based on the automatically labeled documents **(M-step)**
    - maximum likelihood estimation
  - assign a probability to each possible clustering of the mentions of each document **(E-step)**
    - use the normalized scores of the 50-best clusterings

Goals

- Design a new model for unsupervised coreference resolution
- Improve Haghighi and Klein's model with three modifications

Haghighi and Klein's Model

- Cluster-level model
  - assigns a cluster id to each mention

Haghighi and Klein's Model

- Cluster-level model
  - assigns a cluster id to each mention

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist, was summoned to help the King overcome his speech impediment...

Haghighi and Klein's Model

- Cluster-level model
  - assigns a cluster id to each mention

1

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist, was summoned to help the King overcome his speech impediment...

Haghighi and Klein's Model

- Cluster-level model
  - assigns a cluster id to each mention

1 1

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist, was summoned to help the King overcome his speech impediment...

Haghighi and Klein's Model

- Cluster-level model
  - assigns a cluster id to each mention

1 1 2

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist, was summoned to help the King overcome his speech impediment...

Haghighi and Klein's Model

- Cluster-level model
  - assigns a cluster id to each mention

Queen Elizabeth[1] set about transforming her[1] husband[2], King George[2] VI, into a viable monarch[3]. A renowned speech therapist[4], was summoned to help the King[2] overcome his[2] speech impediment[5]...

Haghighi and Klein's Model

- Cluster-level model
  - assigns a cluster id to each mention
  - ensures transitivity automatically

1 1 2
Queen Elizabeth set about transforming her husband,
2 3
King George VI, into a viable monarch. A renowned
4 2
speech therapist, was summoned to help the King
2 5
overcome his speech impediment...

Haghighi and Klein's Generative Story

Haghighi and Klein's Generative Story

- For each mention encountered in a document,
  - generate a cluster id for the mention (according to some cluster id distribution)
  - generate the head noun of the mention (according to some cluster-specific head distribution)

Haghighi and Klein's Generative Story

- For each mention encountered in a document,
  - generate a cluster id for the mention (according to some cluster id distribution)
  - generate the head noun of the mention (according to some cluster-specific head distribution)
- **Inference:** Gibbs sampling

Haghighi and Klein's Generative Story

- For each mention encountered in a document,
  - generate a cluster id for the mention (according to some cluster id distribution)
  - generate the head noun of the mention (according to some cluster-specific head distribution)
- **Inference:** Gibbs sampling
- **Problem with the model: Too simplistic!**
  - mentions with the same head likely to get the same cluster id

Haghighi and Klein's Generative Story

- For each mention encountered in a document,
  - generate a cluster id for the mention (according to some cluster id distribution)
  - generate the head noun of the mention (according to some cluster-specific head distribution)
- **Inference:** Gibbs sampling
- **Problem with the model: Too simplistic!**
  - mentions with the same head likely to get the same cluster id
    - two occurrences of "she" will likely be posited as coreferent
    - particularly inappropriate for generating pronouns

Haghighi and Klein's Generative Story

- For each mention encountered in a document,
  - generate a cluster id for the mention (according to some cluster id distribution)
  - generate the head noun of the mention (according to some cluster-specific head distribution)
- **Inference:** Gibbs sampling
- **Problem with the model: Too simplistic!**
  - mentions with the same head likely to get the same cluster id
- **Extensions:**
  - use a separate "pronoun head model" to generate pronouns
  - incorporate salience

Improving Haghighi and Klein’s Model

- 3 modifications
  - relaxed head generation
  - agreement constraints
  - pronoun-only salience

Modification 1: Relaxed Head Generation

- Motivation
  - H&K's model is linguistically impoverished
    - does not exploit useful knowledge: alias, appositives, …

Modification 1: Relaxed Head Generation

- Motivation
  - H&K's model is linguistically impoverished
    - does not exploit useful knowledge: alias, appositives, …
- Goal
  - simple method for incorporating such knowledge sources

Modification 1: Relaxed Head Generation

- pre-process a document by assigning a "head id" to each mention, such that two mentions have the same head id iff
  - they are the same string
  - or they are aliases
  - or they are in an appositive relation

Modification 1: Relaxed Head Generation

- pre-process a document by assigning a "head id" to each mention, such that two mentions have the same head id iff
  - they are the same string
  - or they are aliases
  - or they are in an appositive relation

International Business Corporation →	1
IBM →	1
Charniak →	2
…	…

Modification 1: Relaxed Head Generation

- pre-process a document by assigning a "head id" to each mention, such that two mentions have the same head id iff
  - they are the same string
  - or they are aliases
  - or they are in an appositive relation
- instead of generating the head noun, generate the head id

International Business Corporation	→	1
IBM	→	1
Charniak	→	2
…		…

Modification 1: Relaxed Head Generation

- pre-process a document by assigning a "head id" to each mention, such that two mentions have the same head id iff
  - they are the same string
  - or they are aliases
  - or they are in an appositive relation

International Business Corporation	→	**1**
IBM	→	**1**
Charniak	→	**2**
…		**…**

- instead of generating the head noun, generate the head id
  - the model views "International Business Corporation" and "IBM" as two mentions having the same head

Modification 1: Relaxed Head Generation

- pre-process a document by assigning a "head id" to each mention, such that two mentions have the same head id iff
  - they are the same string
  - or they are aliases
  - or they are in an appositive relation

International Business Corporation	→	1
IBM	→	1
Charniak	→	2
…		…

- instead of generating the head noun, generate the head id
  - the model views "International Business Corporation" and "IBM" as two mentions having the same head
  - encourages the model to put the two into the same cluster

Modification 2: Agreement Constraints

- Motivation
  - gender and number agreement is implemented as a preference, not as a constraint, in H&K's model

Modification 2: Agreement Constraints

- Motivation
  - gender and number agreement is implemented as a preference, not as a constraint, in H&K's model
    - while the model favors the assignment of a pronoun to a gender- and number-compatible cluster
    - it also favors the assignment of a pronoun to a large cluster

Modification 2: Agreement Constraints

- Motivation
  - gender and number agreement is implemented as a preference, not as a constraint, in H&K's model
    - while the model favors the assignment of a pronoun to a gender- and number-compatible cluster
    - it also favors the assignment of a pronoun to a large cluster
    - if a cluster is large enough, the model may assign the pronoun to the cluster even if the two are not compatible

Modification 2: Agreement Constraints

- Motivation
  - gender and number agreement is implemented as a preference, not as a constraint, in H&K's model
    - while the model favors the assignment of a pronoun to a gender- and number-compatible cluster
    - it also favors the assignment of a pronoun to a large cluster
    - if a cluster is large enough, the model may assign the pronoun to the cluster even if the two are not compatible
- Goal
  - implement gender and number agreement as a constraint

Modification 2: Agreement Constraints

- disallow the generation of a mention by any cluster where the two are incompatible in number or gender

Modification 3: Pronoun-Only Salience

- In H&K’s model, salience is applied to all types of mentions (pronouns, names and nominals) during cluster assignment

- Our hypothesis
  - since names and nominals are less sensitive to salience, the net benefit of applying salience to names and nominals could be negative as a result of inaccurate modeling of salience

- We restrict the application of salience to pronouns only

Improving Haghighi and Klein's Model

- 3 modifications
  - relaxed head generation
  - agreement constraints
  - pronoun-only salience

Evaluation

- EM-based model
- Haghighi and Klein’s model
  - with and without the 3 modifications

Experimental Setup

- The ACE 2003 coreference corpus
  - 3 data sets (Broadcast News, Newswire, Newspaper)
  - each has a training set and a test set; evaluate on test set only

Experimental Setup

- The ACE 2003 coreference corpus
  - 3 data sets (Broadcast News, Newswire, Newspaper)
  - each has a training set and a test set; evaluate on test set only
- Mentions
  - system mentions (mentions extracted by an NP chunker)
  - perfect mentions (mentions extracted from answer key)

Experimental Setup

- The ACE 2003 coreference corpus
  - 3 data sets (Broadcast News, Newswire, Newspaper)
  - each has a training set and a test set; evaluate on test set only
- Mentions
  - system mentions (mentions extracted by an NP chunker)
  - perfect mentions (mentions extracted from answer key)
- Scoring programs: recall, precision, F-measure

Experimental Setup

- The ACE 2003 coreference corpus
  - 3 data sets (Broadcast News, Newswire, Newspaper)
  - each has a training set and a test set; evaluate on test set only
- Mentions
  - system mentions (mentions extracted by an NP chunker)
  - perfect mentions (mentions extracted from answer key)
- Scoring programs: recall, precision, F-measure
  - MUC scoring program (Vilain et al., 1995)

Experimental Setup

- The ACE 2003 coreference corpus
  - 3 data sets (Broadcast News, Newswire, Newspaper)
  - each has a training set and a test set; evaluate on test set only
- Mentions
  - system mentions (mentions extracted by an NP chunker)
  - perfect mentions (mentions extracted from answer key)
- Scoring programs: recall, precision, F-measure
  - MUC scoring program (Vilain et al., 1995)
    - under-penalizes partitions where mentions are over-clustered
    - does not reward successful identification of singleton clusters

Experimental Setup

- The ACE 2003 coreference corpus
  - 3 data sets (Broadcast News, Newswire, Newspaper)
  - each has a training set and a test set; evaluate on test set only
- Mentions
  - system mentions (mentions extracted by an NP chunker)
  - perfect mentions (mentions extracted from answer key)
- Scoring programs: recall, precision, F-measure
  - MUC scoring program (Vilain et al., 1995)
  - CEAF scoring program (Luo, 2005)
    - addresses both weaknesses of the MUC scoring program

Experimental Setup

- The ACE 2003 coreference corpus
  - 3 data sets (Broadcast News, Newswire, Newspaper)
  - each has a training set and a test set; evaluate on test set only
- Mentions
  - system mentions (mentions extracted by an NP chunker)
  - perfect mentions (mentions extracted from answer key)
- Scoring programs: recall, precision, F-measure
  - MUC scoring program (Vilain et al., 1995)
  - CEAF scoring program (Luo, 2005)
  - CEAF variant
    - same as CEAF, but ignores singleton clusters

Heuristic Baseline

- Simple rule-based system

- Posits two mentions as coreferent if and only if they are
  - the same string
  - aliases
  - in an appositive relation

Heuristic Baseline: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**

Heuristic Baseline: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**

Heuristic Baseline: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**

Heuristic Baseline: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**

Heuristic Baseline: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**

EM-Based Model

- Initialize the parameters using one (labeled) document
  - rather than using randomly guessed clusterings

EM-Based Model: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**
Our EM-based Model	42.4	66.0	**51.6**	55.2	60.6	**57.8**

EM-Based Model: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**
Our EM-based Model	42.4	66.0	**51.6**	55.2	60.6	**57.8**

- gains in both recall and precision
- F-measure increases by 15%

Duplicated Haghighi and Klein's Model

- The version that incorporates both salience and the separate model for generating pronouns
- Use the same labeled document as in the EM-based model to learn one of the concentration parameters, α

Duplicated H&K's Model: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**
Our EM-based Model	42.4	66.0	**51.6**	55.2	60.6	**57.8**
Duplicated Haghighi and Klein	50.8	40.7	**45.2**	43.0	40.9	**41.9**

Duplicated H&K's Model: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**
Our EM-based Model	42.4	66.0	**51.6**	55.2	60.6	**57.8**
Duplicated Haghighi and Klein	50.8	40.7	**45.2**	43.0	40.9	**41.9**

- In comparison to EM-based model
  - precision drops substantially
  - F-measure decreases by 6-16%

Adding 3 Modifications: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**
Our EM-based Model	42.4	66.0	**51.6**	55.2	60.6	**57.8**
Duplicated Haghighi and Klein	50.8	40.7	**45.2**	43.0	40.9	**41.9**
+ Relaxed Head Generation	48.3	45.7	**47.0**	40.9	50.0	**45.0**
+ Agreement Constraints	50.4	47.5	**48.9**	41.7	51.2	**46.0**
+ Pronoun-only Salience	52.2	53.0	**52.6**	44.3	57.3	**50.0**

Adding 3 Modifications: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**
Our EM-based Model	42.4	66.0	**51.6**	55.2	60.6	**57.8**
Duplicated Haghighi and Klein	50.8	40.7	**45.2**	43.0	40.9	**41.9**
+ Relaxed Head Generation	48.3	45.7	**47.0**	40.9	50.0	**45.0**
+ Agreement Constraints	50.4	47.5	**48.9**	41.7	51.2	**46.0**
+ Pronoun-only Salience	52.2	53.0	**52.6**	44.3	57.3	**50.0**

- In comparison to Duplicated Haghighi and Klein
  - F-measure improves after the addition of each modification

Adding 3 Modifications: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**
Our EM-based Model	42.4	66.0	**51.6**	55.2	60.6	**57.8**
Duplicated Haghighi and Klein	50.8	40.7	**45.2**	43.0	40.9	**41.9**
+ Relaxed Head Generation	48.3	45.7	**47.0**	40.9	50.0	**45.0**
+ Agreement Constraints	50.4	47.5	**48.9**	41.7	51.2	**46.0**
+ Pronoun-only Salience	52.2	53.0	**52.6**	44.3	57.3	**50.0**

- In comparison to Duplicated Haghighi and Klein
  - F-measure improves after the addition of each modification
  - modest gain in recall and substantial gain in precision when all modifications are applied (7-9% gain in F-measure)

Supervised Resolver: MUC Results

Experiments on System Mentions	Broadcast News			Newswire		
	R	P	F	R	P	F
Heuristic Baseline	30.9	44.3	**36.4**	36.3	53.4	**43.2**
Our EM-based Model	42.4	66.0	**51.6**	55.2	60.6	**57.8**
Duplicated Haghighi and Klein	50.8	40.7	**45.2**	43.0	40.9	**41.9**
+ Relaxed Head Generation	48.3	45.7	**47.0**	40.9	50.0	**45.0**
+ Agreement Constraints	50.4	47.5	**48.9**	41.7	51.2	**46.0**
+ Pronoun-only Salience	52.2	53.0	**52.6**	44.3	57.3	**50.0**
Fully Supervised Model	53.0	70.3	**60.4**	53.1	70.5	**60.6**

- Trained using C4.5, entire ACE training set, 34 features
- Outperforms the unsupervised models by 3-8%

MUC, CEAF, CEAF-Variant F-Scores

Experiments on System Mentions	Broadcast News			Newswire		
	MUC	CEAF	CEAFV	MUC	CEAF	CEAFV
Heuristic Baseline	36.4	48.4	46.3	43.2	54.2	50.3
Our EM-based Model	51.6	55.7	52.9	57.8	59.6	52.8
Duplicated Haghighi and Klein	45.2	45.2	39.0	41.9	48.8	41.7
+ Relaxed Head Generation	47.0	47.5	42.3	45.0	52.6	46.3
+ Agreement Constraints	48.9	51.4	47.0	46.0	54.5	48.4
+ Pronoun-only Salience	52.6	54.7	51.1	50.0	57.4	51.2
Fully Supervised Model	60.4	61.8	59.9	60.6	64.5	60.6

MUC, CEAF, CEAF-Variant F-Scores

Experiments on System Mentions	Broadcast News			Newswire		
	MUC	CEAF	CEAFV	MUC	CEAF	CEAFV
Heuristic Baseline	36.4	48.4	46.3	43.2	54.2	50.3
Our EM-based Model	51.6	55.7	52.9	57.8	59.6	52.8
Duplicated Haghighi and Klein	45.2	45.2	39.0	41.9	48.8	41.7
+ Relaxed Head Generation	47.0	47.5	42.3	45.0	52.6	46.3
+ Agreement Constraints	48.9	51.4	47.0	46.0	54.5	48.4
+ Pronoun-only Salience	52.6	54.7	51.1	50.0	57.4	51.2
Fully Supervised Model	60.4	61.8	59.9	60.6	64.5	60.6

MUC, CEAF, CEAF-Variant F-Scores

Experiments on System Mentions	Broadcast News			Newswire		
	MUC	CEAF	CEAFV	MUC	CEAF	CEAFV
Heuristic Baseline	36.4	48.4	46.3	43.2	54.2	50.3
Our EM-based Model	51.6	55.7	52.9	57.8	59.6	52.8
Duplicated Haghighi and Klein	45.2	45.2	39.0	41.9	48.8	41.7
+ Relaxed Head Generation	47.0	47.5	42.3	45.0	52.6	46.3
+ Agreement Constraints	48.9	51.4	47.0	46.0	54.5	48.4
+ Pronoun-only Salience	52.6	54.7	51.1	50.0	57.4	51.2
Fully Supervised Model	60.4	61.8	59.9	60.6	64.5	60.6

MUC, CEAF, CEAF-Variant F-Scores

Experiments on System Mentions	Broadcast News			Newswire		
	MUC	CEAF	CEAFV	MUC	CEAF	CEAFV
Heuristic Baseline	36.4	48.4	46.3	43.2	54.2	50.3
Our EM-based Model	51.6	55.7	52.9	57.8	59.6	52.8
Duplicated Haghighi and Klein	45.2	45.2	39.0	41.9	48.8	41.7
+ Relaxed Head Generation	47.0	47.5	42.3	45.0	52.6	46.3
+ Agreement Constraints	48.9	51.4	47.0	46.0	54.5	48.4
+ Pronoun-only Salience	52.6	54.7	51.1	50.0	57.4	51.2
Fully Supervised Model	60.4	61.8	59.9	60.6	64.5	60.6

MUC, CEAF, CEAF-Variant F-Scores

Experiments on System Mentions	Broadcast News			Newswire		
	MUC	CEAF	CEAFV	MUC	CEAF	CEAFV
Heuristic Baseline	36.4	48.4	46.3	43.2	54.2	50.3
Our EM-based Model	51.6	55.7	52.9	57.8	59.6	52.8
Duplicated Haghighi and Klein	45.2	45.2	39.0	41.9	48.8	41.7
+ Relaxed Head Generation	47.0	47.5	42.3	45.0	52.6	46.3
+ Agreement Constraints	48.9	51.4	47.0	46.0	54.5	48.4
+ Pronoun-only Salience	52.6	54.7	51.1	50.0	57.4	51.2
Fully Supervised Model	60.4	61.8	59.9	60.6	64.5	60.6

- Similar performance trends across the 3 scoring programs

Experiments using Perfect Mentions

- Similar performance trends observed
  - except that the unsupervised models perform comparably to the fully-supervised resolver

- Conclusions drawn from system mentions are not always generalizable to perfect mentions and vice versa

Summary

- Presented an EM-based model for unsupervised coreference resolution that
  - outperforms Haghighi and Klein’s coreference model
  - compares favorably to a modified version of their model