Coreference Resolution

Emory Center of Language and Information Research
NLP Reading Group Presentation

Henry Chen

Overview

1. Significance of coreference resolution
2. Connection with NER
3. Resolution Structure
   A. Mention Detection
   B. Coreference Resolution
4. Current Approaches
   C. Feature-based
   D. Unsupervised learning
5. Progression of Coreference Resolution Models (2007-2009)
6. Stanford Sieve System (2011-2013)
7. Topics yet to be Explored

Importance

Purposes:

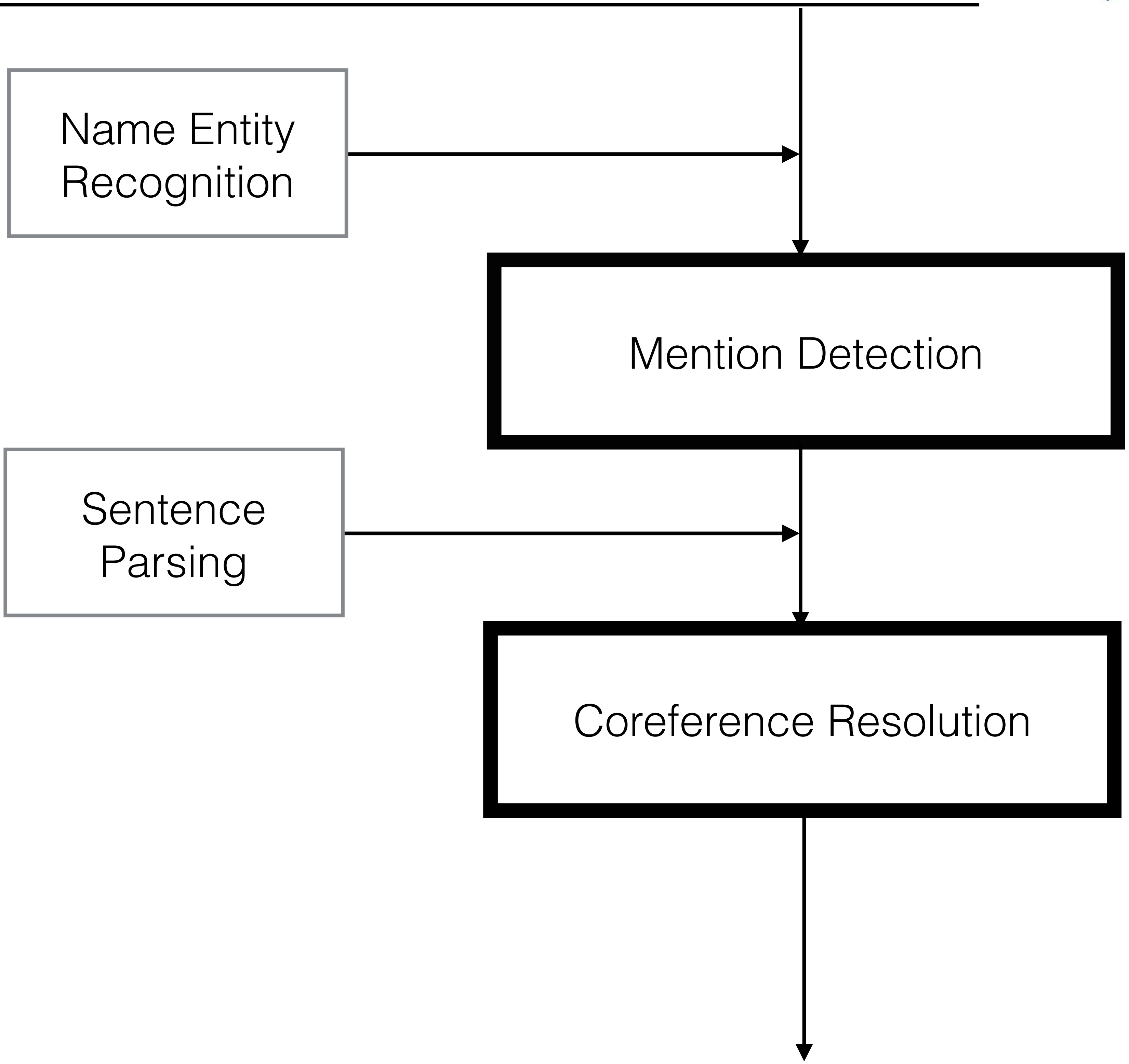
- Finding mentions in text
- Grouping all relevant mentions together
- Referring them back to the real-world entities

- Connecting Entities across sentences
- A preliminary step for knowledge graph generation

Connection with NER

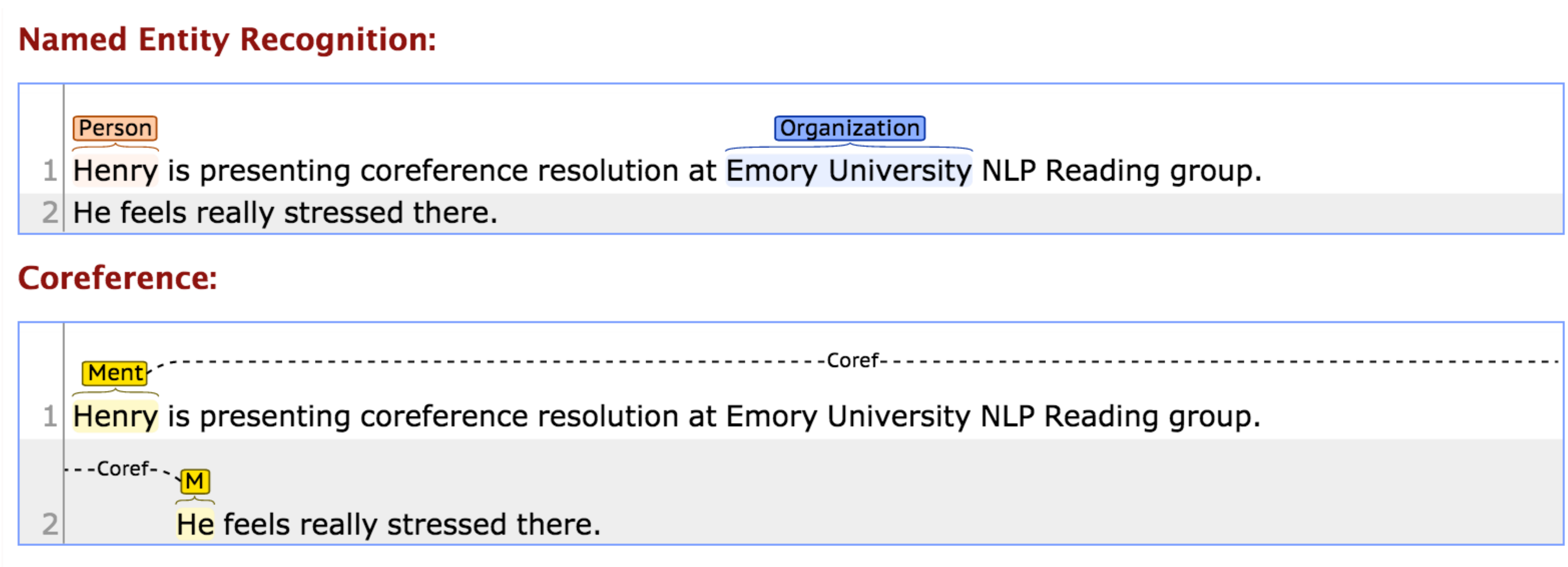
- NER is a sub-topic of coreference resolution
- Extension to Mention Detection

Structure


Name Entity
Recognition
Mention Detection
Sentence
Parsing
Coreference Resolution

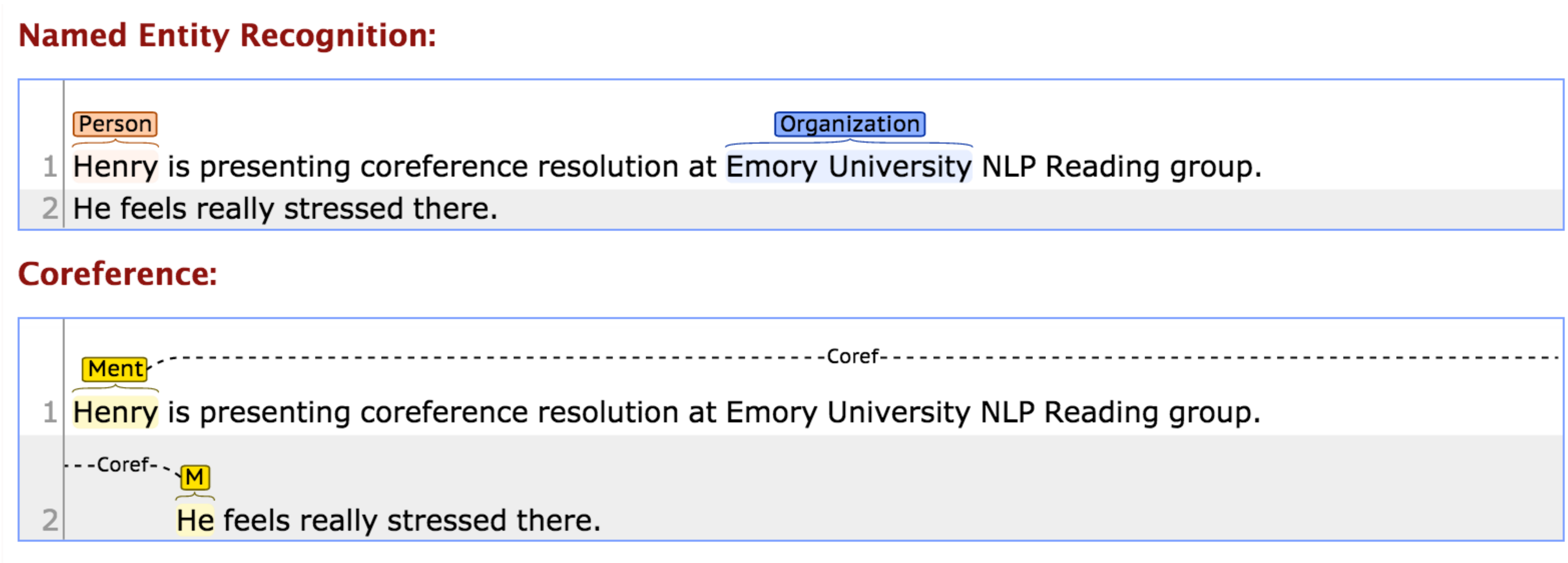
Mention Detection

- 3 types of mentions:
  - Name mention
  - Nominal mention   i.e. the guy wearing a blue shirt
  - Pronoun mention



Coreference Resolution

- Grouping of mentions
  - Directed / Undirected
- Merge/Connect entities together



Approaches

1. Incorporating more feature into the models, such as mention-pair model and cluster ranking models
   - Knowledge sources for the features
     A. Semantic knowledge sources. i.e. FrameNet, WordNet, etc.
     B. Annotated corpuses. i.e. ACE, MUC, OntoNotes, etc.
     C. Unannotated corpuses. (represented in horn rules or relations)

2. Shifting from the supervised learning setting to unsupervised learning

Evaluation methods

MUC

- *Counts the minimum number of links between mentions*
- *Cannot represent singleton entities (Only mentioned once)*
- *Give no credit for separating singleton entities from a chains*

B^3

- *Overcome the short-coming of MUC score*
- *Compute precision and recall for all mentions*

$$Precision(m_i) = \frac{|R_{m_i} \cap K_{m_i}|}{|R_{m_i}|}$$

$$Recall(m_i) = \frac{|R_{m_i} \cap K_{m_i}|}{|K_{m_i}|}$$

R: Respones, K: Key

Progression of Coreference Resolution

Statistical approach (binary classification) (Denis and Baldridge, 2007)

- Asymmetric interpretation (Nguyen and Kim, 2008)
- Symmetric interpretation

-> Pairwise formulation
-> feature independence assumption

Cluster-based features

- Convert classification of pairwise models into cluster of mention
- Cluster of mention -> an individual entity (Culotta et al., 2007)

Clark and Gonzalez-Brenes. *Coreference Resolution: Current Trends and Future Directions.* 2008

Progression of Coreference Resolution

First-Order Probabilistic Models for Coreference Resolution

- Soften the independence assumption of the pairwise model by allowing feature based on first-order logic
- Creating a model over clusters instead of over pairs

$$p(y_{ij}|x_{ij}) = \frac{1}{Z_{x_{ij}}} \exp \sum_k \lambda_k f_k(x_{ij}, y_{ij}) \longrightarrow p(y_j|x^j) = \frac{1}{Z_{x^j}} \exp \sum_k \lambda_k f_k(x^j, y_j)$$

Training Technique

- Sampling positive and negative examples uniformly using Margin Infused Relaxed Algorithms (MIRA)
- Rank positive example higher than the negative ones

  -> True mentions in bad clusters are not unjustly penalized

Aron Culotta, Michael Wick, Robert Hall, and Andrew McCallum. 2007. *First-order probabilistic models for coreference resolution*. In Proceedings of the Conference on Human Language Technology.

Progression of Coreference Resolution

EMORY UNIVERSITY

First-Order Probabilistic Models for Coreference Resolution

Result

1. Allows to model constraints like: “do not prefer entities whose only mentions are pronouns”
2. Introduction of a error-driven, rank-based training technique

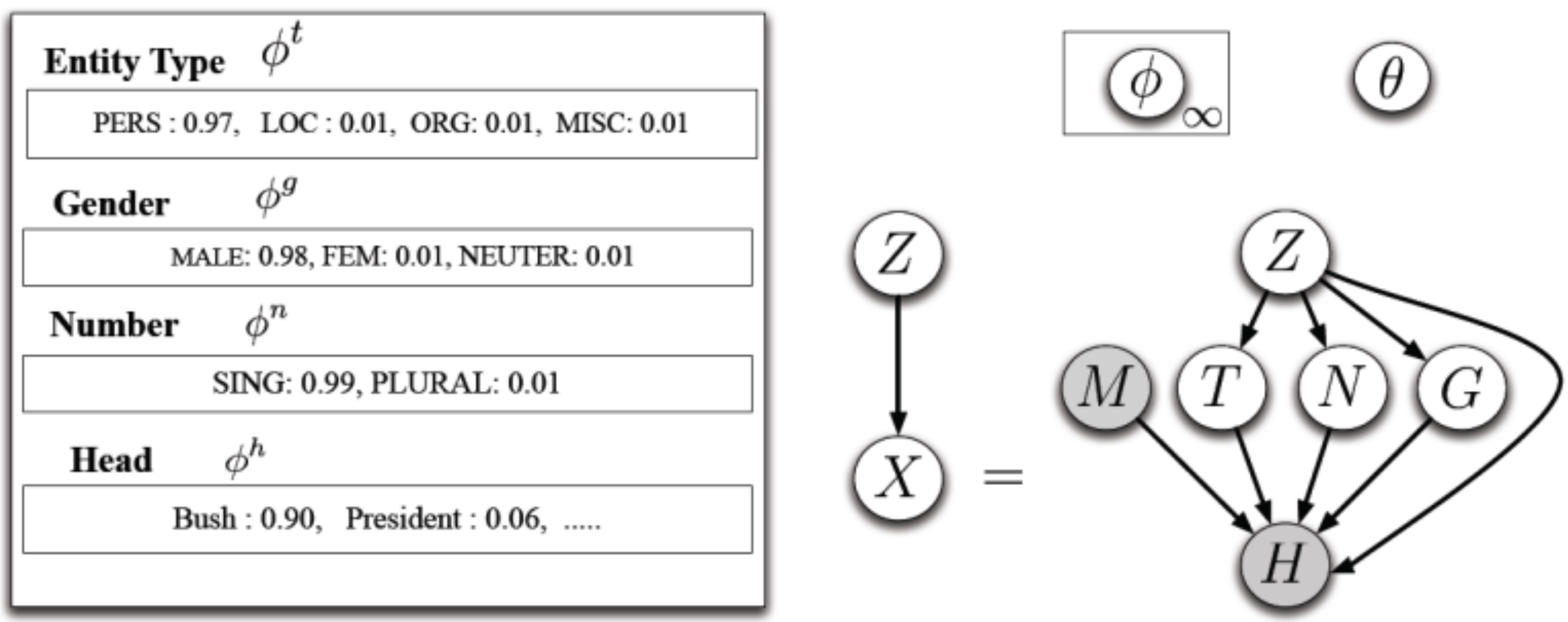
	F1	**Prec**	**Rec**
First-Order MIRA	**79.3**	86.7	73.2
Pairwise MIRA	72.5	92.0	59.8
First-Order Uniform	**69.2**	79.0	61.5
Pairwise	62.4	62.5	62.3

Aron Culotta, Michael Wick, Robert Hall, and Andrew McCallum. 2007. *First-order probabilistic models for coreference resolution*. In Proceedings of the Conference on Human Language Technology.

Progression of Coreference Resolution

Unsupervised Coreference Resolution in a Nonparametric Bayesian Model

- 1st unsupervised method that perform as good as supervised approach
- Proposed a hierarchical Dirichlet Process to find the referents of mentions within a document. Then extends to cross documents



Z: random variable that takes the value of the the index of an entity.
X: collection of variables associated with a mention in the model (namely entity type T, numberN, gender G, head H, mention M).

Aria Haghighi and Dan Klein. 2007. *Unsupervised coreference resolution in a nonparametric bayesian model*.
In Proceedings of the Association for Computational Linguistics

Progression of Coreference Resolution

Unsupervised Coreference Resolution in a Nonparametric Bayesian Model

Dataset	Num Docs.	Prec.	Recall	F_1
MUC-6	60	80.8	52.8	63.9
+DRYRUN-TRAIN	251	79.1	59.7	68.0
+ENGLISH-NWIRE	381	80.4	62.4	70.3

(a)

Dataset	Prec.	Recall	F_1
ENGLISH-NWIRE	66.7	62.3	64.2
ENGLISH-BNEWS	63.2	61.3	62.3
CHINESE-NWIRE	71.6	63.3	67.2
CHINESE-BNEWS	71.2	61.8	66.2

(b)

Table 2: Formal Results: Our system evaluated using the MUC model theoretic measure Vilain et al. (1995). The table in (a) is our performance on the thirty document MUC-6 formal test set with increasing amounts of training data. In all cases for the table, we are evaluating on the same thirty document test set which is included in our training set, since our system in unsupervised. The table in (b) is our performance on the ACE 2004 training sets.

Aria Haghighi and Dan Klein. 2007. *Unsupervised coreference resolution in a nonparametric bayesian model*.
In Proceedings of the Association for Computational Linguistics

Progression of Coreference Resolution

EMORY UNIVERSITY

Unsupervised Models for Coreference Resolution

- Conceptualize the problem as inducing coreference partitions on unlabeled document instead of mention pair classification
- Modified Expectation-Maximization (EM) algorithm, so the number of cluster does not have to be predetermined

Given a document D, a clustering C, let θ be the model parameters.

E-step: Compute the posterior probabilities of the clusterings, $P(C|D,\theta)$ based on the current θ.

M-step: Using the value of $P(C|D,\theta)$ computed in the E-step, find the θ' that maximizes the expected log-likelihood: $\sum_C P(C|D,\theta) log P(D,C|\theta')$.

Vincent Ng. 2008. *Unsupervised models for coreference resolution*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Unsupervised Models for Coreference Resolution

		Broadcast News (BNEWS)						Newswire (NWIRE)					
		True Mentions			System Mentions			True Mentions			System Mentions		
	Experiments	R	P	F	R	P	F	R	P	F	R	P	F
1	Heuristic Baseline	27.8	72.0	40.1	30.9	44.3	36.4	31.2	70.3	43.3	36.3	53.4	43.2
2	Degenerate EM Baseline	63.6	53.1	57.9	70.8	36.3	48.0	64.5	42.6	51.3	69.0	25.1	36.8
3	Our EM-based Model	56.1	71.4	**62.8**	42.4	66.0	51.6	47.0	68.3	**55.7**	55.2	60.6	**57.8**
4	Haghighi and Klein Baseline	49.4	60.2	54.3	50.8	40.7	45.2	44.7	55.5	49.5	43.0	40.9	41.9
5	+ Relaxed Head Generation	53.0	65.4	58.6	48.3	45.7	47.0	45.1	62.5	52.4	40.9	50.0	45.0
6	+ Agreement Constraints	53.6	68.7	60.2	50.4	47.5	48.9	44.6	63.7	52.5	41.7	51.2	46.0
7	+ Pronoun-only Salience	56.8	68.3	62.0	52.2	53.0	**52.6**	46.8	66.2	54.8	44.3	57.3	50.0
8	Fully Supervised Model	53.7	70.8	61.1	53.0	70.3	60.4	52.0	69.6	59.6	53.1	70.5	60.6

Table 4: Results obtained using the MUC scoring program for the Broadcast News and Newswire data sets

		Broadcast News (BNEWS)						Newswire (NWIRE)					
		True Mentions			System Mentions			True Mentions			System Mentions		
	Experiments	R	P	F	R	P	F	R	P	F	R	P	F
1	Heuristic Baseline	42.1	75.8	54.1	44.2	48.7	46.3	43.9	73.4	54.9	47.5	53.4	50.3
2	Degenerate EM Baseline	51.2	43.1	46.8	53.7	26.8	35.8	51.0	30.5	38.2	45.1	18.6	26.3
3	Our EM-based Model	53.3	60.5	**56.7**	47.5	59.6	**52.9**	49.2	60.7	54.4	53.5	52.1	**52.8**
4	Haghighi and Klein Baseline	43.7	48.8	46.1	46.0	33.9	39.0	45.5	51.7	48.4	44.6	39.2	41.7
5	+ Relaxed Head Generation	45.8	52.4	48.9	45.4	39.6	42.3	46.0	57.0	50.9	44.5	48.3	46.3
6	+ Agreement Constraints	51.8	60.5	55.8	50.6	43.8	47.0	47.8	60.1	53.2	46.5	50.4	48.4
7	+ Pronoun-only Salience	53.9	59.9	**56.7**	52.3	49.9	51.1	49.6	62.8	**55.4**	47.4	55.7	51.2
8	Fully Supervised Model	55.0	63.3	58.8	56.2	64.2	59.9	54.7	64.7	59.3	56.5	65.4	60.6

Table 5: Results obtained using the CEAF scoring program for the Broadcast News and Newswire data sets

Vincent Ng. 2008. *Unsupervised models for coreference resolution*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Progression of Coreference Resolution

Joint Unsupervised Coreference Resolution with Markov Logic

- Unsupervised model using Markov Logic Network (MLN), first-order knowledge base with a weight attached to each formula
- Behave exactly as first-order logic when weight is infinite
- Soften the constraint imposed by a sets of first-order logic formula

Table 5: Our coreference results in precision, recall, and F1 for pairwise resolution.

Pairwise	Prec.	Rec.	F1
MUC-6	63.0	57.0	59.9
EN-BNEWS	51.2	36.4	42.5
EN-NWIRE	62.6	38.9	48.0
BNEWS	44.6	32.3	37.5
NWIRE	59.7	42.1	49.4
NPAPER	64.3	43.6	52.0

Table 6: Average gold number of clusters per document vs. the mean absolute error of our system.

# Clusters	MUC-6	EN-BN	EN-NW
Gold	15.4	22.3	37.2
Mean Error	4.7	3.0	4.8
# Clusters	BNEWS	NWIRE	NPAPER
Gold	20.4	39.2	55.2
Mean Error	2.5	5.6	6.6

Hoifung Poon and Pedro Domingos. 2008. *Joint unsupervised coreference resolution with markov logic*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Progression of Coreference Resolution

Joint Determination of Anaphoricity and Coreference Resolution using Integer Programming

- Using Integer Linear Programming (ILP) to optimize and join together multiple classifiers in coreference
- Model anaphoricity and coreference as a joint task
- Become intractable unless using pairwise independence assumption

System	BNEWS			NPAPER			NWIRE		
	R	P	F	R	P	F	R	P	F
COREF-PAIRWISE	54.4	77.4	63.9	58.1	80.7	67.6	53.8	78.2	63.8
COREF-ILP	62.2	75.5	68.2	67.1	77.3	71.8	60.1	74.8	66.8
JOINT-ILP	62.1	78.0	69.2	68.0	77.6	72.5	60.8	75.8	67.5

Table 2: Recall (R), precision (P), and f-score (F) on the three ACE datasets for the basic coreference system (COREF-PAIRWISE), the coreference only ILP system (COREF-ILP), and the joint anaphoricity-coreference ILP system (JOINT-ILP). All f-score differences are significant ($p < .05$).

Pascal Denis and Jason Baldridge. 2007. *Joint determination of anaphoricity and coreference resolution using integer programming*. In Proceedings of the North American Association for Computational Linguistics and the Conference on Human Language Technology.

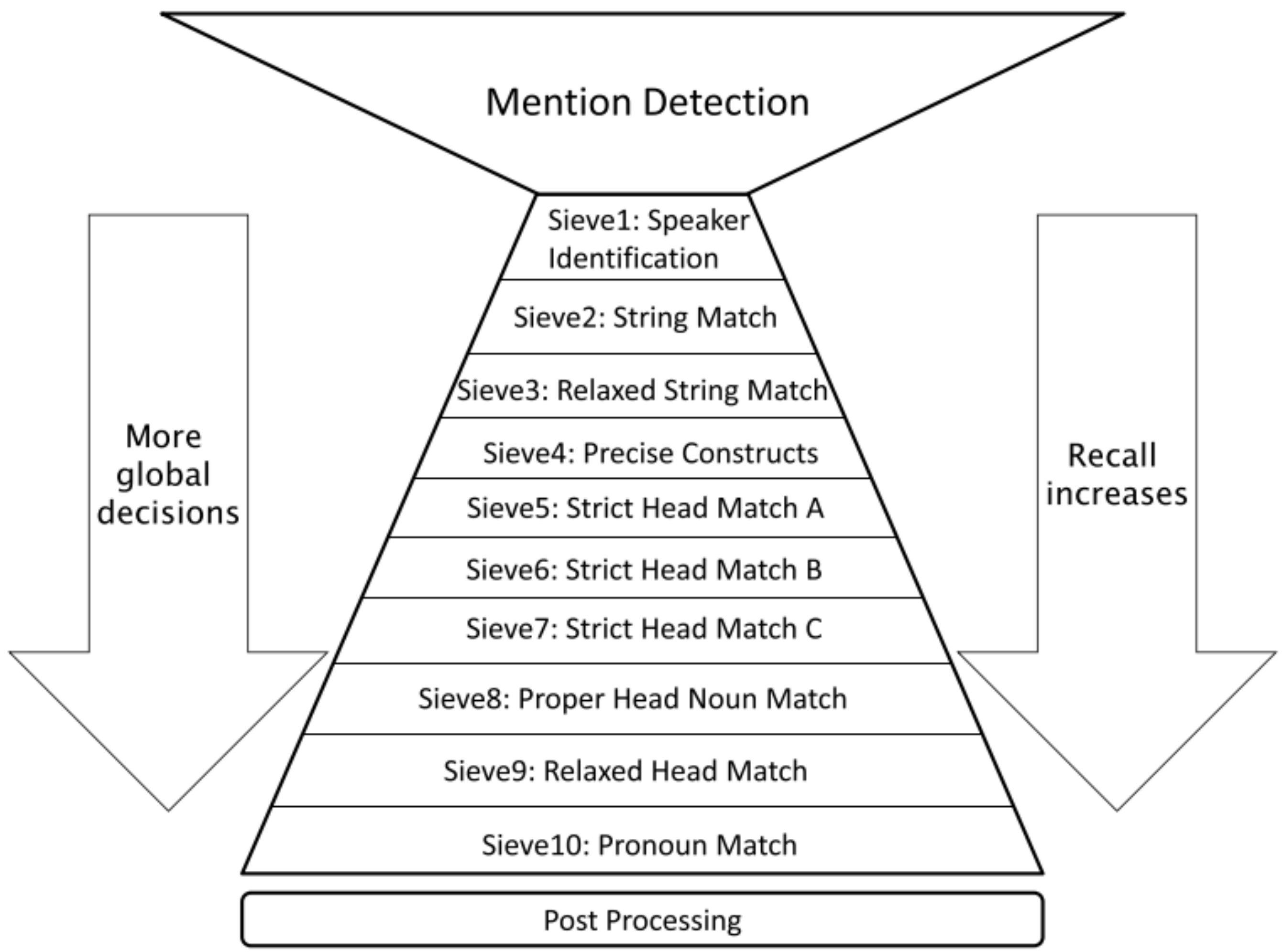
Progression of Coreference Resolution

EMORY UNIVERSITY

	F1	Precision	Recall
First-Order Probabilistic Models	79.3 (B3) ~69.0 (MUC)	86.7 (B3) ~ 76.0 (MUC)	73.2 (B3) ~ 63.0 (MUC)
Unsupervised Nonparametric Bayesian Model	70.3	80.4	62.4
Unsupervised Expectation-Maximization Models	62.8	71.4	56.1
Joint Unsupervised Model with Markov Logic	59.9	63.0	57.0
Joint Determination model using Integer Programming	69.2	78.9	62.1

Stanford Sieve System

Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules



Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Henry Chen

Stanford Sieve System

EMORY UNIVERSITY

Overview

- A entity-centric coreference model
    - Each decision can be globally informed by previously clustered mentions and their shared attributes
- Using solely simple deterministic rules
- A highly modular system to provide flexibility
    - Lack of language-specific lexical features (language independent)
- Maintain the balance of high precision and recall

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Mention Detection

- Value high recall over high precision
- Mark all NPs, pronouns, and name entity mentions
- Apply exclusion rules to remove unwanted mentions

  i.e.
  - Numeric entities
  - Pre-defined dictionary (i.e. American, hmm, etc.)
  - Partitive or quantifier expression (i.e. none of…)

*** The rules might change depending on the corpus annotation

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Coreference model

Mention Selection in a Given Sieve

- Take advantage of the partially clustered mentions
- Look at only the the first mention of each cluster to minimize error

$$\{m_1^1, m_2^2, m_3^2, m_4^3, m_5^1, m_6^2\}$$

- Further prune the search space with a model of *discourse salience*
  - *Disable coreference for ambiguous pronouns/articles (* some, other, a, an, etc.)

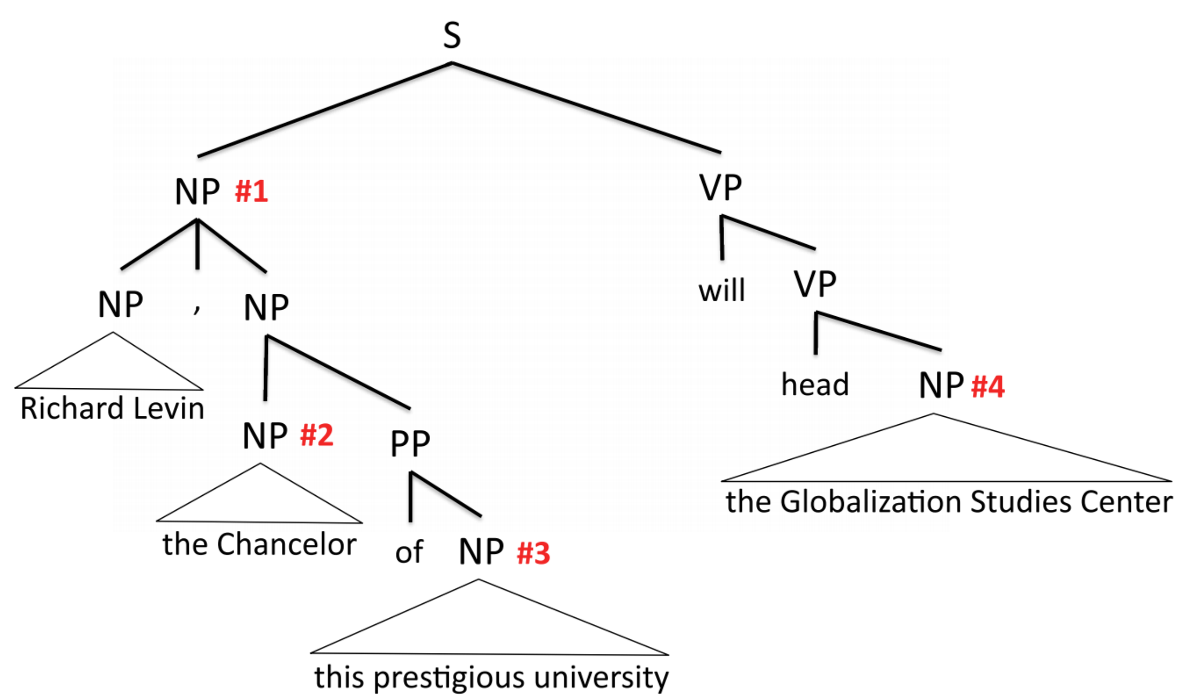
Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Coreference model

Antecedent Selection for a Given Sieve

- Recursively find all possible antecedents
- Antecedents are sorted based on their textual proximity to the anaphoric mention
- The selection stops when a match is found



Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Henry Chen

Stanford Sieve System

Coreference model

Feature Sharing in the Entity-Centric Model

- Merge all mention attributes as a joint, shared attributes for the cluster to be more inclusive for later resolution
- To avoid negligence of correct links due to missing or mismatch attributes

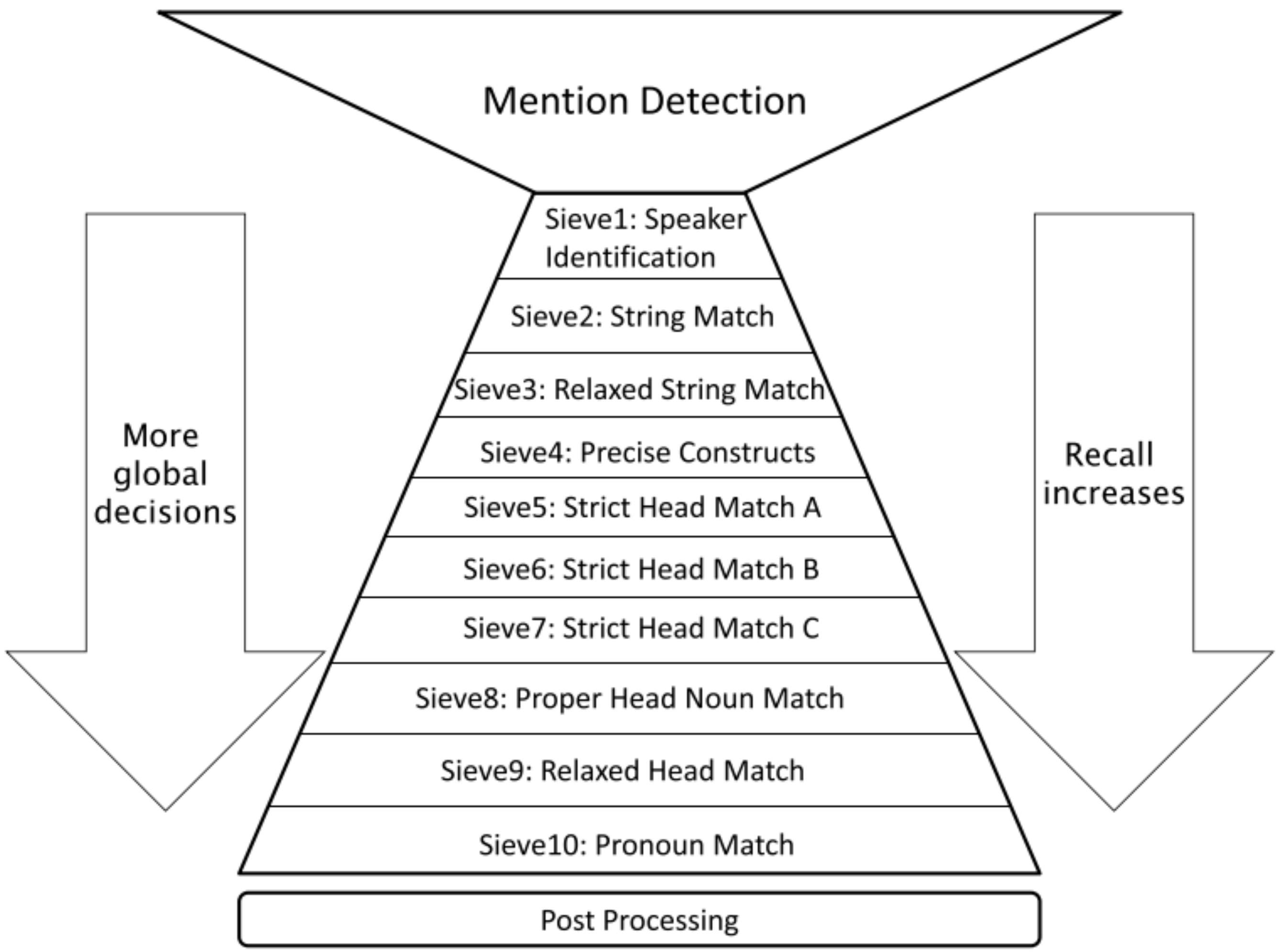
For example:

a group of student: { singular }
five students : { plural }
a group of student <- five students: { singular, plural }

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Coreference sieves



Sieve 1. Speaker Identification

Sieve 2. Exact String Match

Sieve 3. Relaxed String Match

Sieve 4. Precise Constructs

Sieve 5-7. Strict Head Match A-C

Sieve 8. Proper Head Match

Sieve 9. Relaxed Head Match

Sieve 10. Pronoun Resolution

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Sieve 1. Speaker Identification

- *Searches for subjects of reporting verbs in same or neighboring sentence to a quotation*

Sieve 2. Exact String Match

- *Two mentions are linked if all text (including modifiers) are the same*

Sieve 3. Relaxed String Match

- *Two mentions are linked if string obtained after dropping head words are identical*

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Sieve 4. Precise Constructs

- *Matches mentions based on syntactic structures*
    1. *Appositive*
    2. *Predicate nominative*
    3. *Relative pronoun*
    4. *Acronym*
    5. *Demonym*

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Sieve 5-9. Head Word Match Variants

- *Mention head matches any word in the antecedent entity*
- *No location or numeric mismatches*
- *Cannot be child NP in another NP's constituent*

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Sieve 10. Pronomial Coreference Resolution

- *Number*
- *Gender*
- *Person*
- *Animacy*
- *NER labels*
- *Pronoun proximity*

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Post processing

- Implementations of several transformations
- Ensure the output of the model matches the annotation specification of corresponding corpus

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Performance

- CoNLL '11

Table 4
Performance of the top systems in the CoNLL-2011 shared task. All these systems use automatically detected mentions. We report results for both the closed and the open tracks, which allowed the use of resources not provided by the task organizers. MD indicates mention detection, and gold boundaries indicate that mention boundary information is given.

System	MD			MUC			B^3			CEAF-ϕ_4			BLANC			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Closed Track																
This paper	75.1	66.8	70.7	61.8	57.5	59.6	68.4	68.2	68.3	43.4	47.8	45.5	70.6	76.2	73.0	57.8
Sapena	92.4	28.2	43.2	56.3	63.2	59.6	62.8	72.1	67.1	44.8	38.4	41.3	69.5	73.1	71.1	56.0
Chang	68.1	62.0	64.9	57.2	57.2	57.2	67.1	70.5	68.8	41.9	41.9	41.9	71.2	77.1	73.7	56.0
Nugues	69.9	68.1	69.0	60.2	57.1	58.6	66.7	64.2	65.5	38.1	41.1	39.5	72.0	70.3	71.1	54.5
Santos	67.8	63.3	65.5	59.2	54.3	56.7	68.8	62.8	65.7	35.9	40.2	37.9	73.4	66.9	69.5	53.4
Song	57.8	80.4	67.3	53.7	67.8	60.0	60.7	66.1	63.2	43.4	30.7	36.0	69.5	59.7	61.5	53.1
Stoyanov	70.8	65.0	67.8	63.6	54.0	58.4	72.6	53.3	61.4	32.0	40.8	35.9	73.2	58.9	60.9	51.9
Sobha	67.8	62.1	64.8	51.1	49.9	50.5	62.6	65.4	64.0	40.7	41.8	41.2	61.4	68.4	63.9	51.9
Kobdani	62.1	60.0	61.0	55.6	51.5	53.5	69.7	62.4	65.9	32.3	35.4	33.8	61.9	63.5	62.6	51.0
Zhou	61.1	63.6	62.3	45.7	52.8	49.0	57.1	72.9	64.1	43.2	36.8	39.7	61.1	73.9	64.7	50.9
Charton	65.9	62.8	64.3	55.1	50.1	52.5	66.3	58.4	62.1	34.3	39.1	36.5	69.9	62.2	64.8	50.4
Yang	71.9	57.5	63.9	59.9	46.4	52.3	71.6	55.1	62.3	30.3	42.4	35.3	71.1	61.8	64.6	50.0
Hao	64.5	64.1	64.3	57.9	51.4	54.5	67.8	55.4	61.0	30.1	35.8	32.7	72.6	62.4	65.4	49.4
Xinxin	65.5	58.7	61.9	48.5	44.9	46.6	61.6	62.3	61.9	35.2	38.6	36.8	63.0	65.8	64.3	48.5
Zhang	55.4	68.3	61.1	42.0	55.6	47.9	52.6	73.1	61.1	42.0	30.3	35.2	62.8	69.2	65.2	48.1
Kummerfeld	69.8	57.0	62.7	46.4	39.6	42.7	63.6	57.3	60.3	35.1	42.3	38.3	58.7	61.6	59.9	47.1
Zhekova	67.5	37.6	48.3	28.9	20.7	24.1	67.1	56.7	61.5	31.6	41.2	35.8	52.8	57.1	53.8	40.4
Irwin	17.1	61.1	26.7	12.5	50.6	20.0	35.1	89.9	50.5	45.8	17.4	25.2	51.5	56.8	51.1	31.9
Open Track																
This paper	74.3	67.9	70.9	62.8	59.3	61.0	68.9	69.0	68.9	43.3	46.8	45.0	71.9	76.6	74.0	58.3
Cai	67.2	67.6	67.4	56.7	58.9	57.8	64.6	71.0	67.7	42.7	40.7	41.7	69.8	74.0	71.6	55.7
Uryupina	70.6	66.3	68.4	59.7	55.7	57.6	66.3	64.1	65.2	38.3	42.2	40.2	69.2	68.5	68.9	54.3
Klenner	64.4	60.3	62.3	49.0	50.7	49.9	61.7	68.6	65.0	41.3	39.7	40.5	66.1	73.9	69.1	51.8
Irwin	24.6	62.3	35.3	18.6	51.0	27.2	39.0	85.6	53.6	43.3	19.4	26.8	51.6	52.9	51.8	35.8
Closed Track - gold boundaries																
This paper	79.5	71.3	75.2	65.9	62.1	63.9	69.5	70.6	70.0	46.3	50.5	48.3	72.0	78.6	74.8	60.7
Nugues	74.2	70.7	72.4	64.3	60.1	62.1	68.3	65.2	66.7	39.9	44.2	41.9	72.5	71.0	71.8	56.9
Chang	63.4	73.2	67.9	55.0	65.5	59.8	62.2	76.7	68.7	46.8	37.2	41.4	71.0	79.3	74.3	56.6
Santos	65.8	69.9	67.8	57.8	61.4	59.5	64.5	70.3	67.3	41.4	38.2	39.7	72.7	72.0	72.3	55.5
Kobdani	67.1	65.1	66.1	62.6	56.8	59.6	73.2	62.2	67.3	32.9	37.3	34.9	64.1	64.1	64.1	53.9
Stoyanov	76.9	64.7	70.3	69.8	55.0	61.5	77.1	52.5	62.5	31.0	44.8	36.6	76.6	60.3	63.0	53.6
Zhang	59.6	71.2	64.9	46.1	58.8	51.6	53.9	73.4	62.2	43.5	32.1	37.0	64.1	70.5	66.5	50.3
Song	58.4	77.6	66.7	46.7	68.4	55.5	54.4	70.2	61.3	43.8	25.9	32.5	66.3	58.8	60.2	49.8
Zhekova	69.2	57.3	62.7	33.5	37.2	35.2	55.5	68.2	61.2	38.3	34.7	36.4	53.5	63.3	54.8	44.3

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Performance

- ACE & MUC

Table 5
Comparison of our system with the other reported results on the ACE and MUC corpora. All these systems use gold mention boundaries.

System	MUC			B^3		
	R	P	F1	R	P	F1
ACE2004-Culotta-Test						
This paper	70.2	82.7	75.9	74.5	88.7	81.0
Haghighi and Klein (2009)	77.7	74.8	79.6	78.5	79.6	79.0
Culotta et al. (2007)	–	–	–	73.2	86.7	79.3
Bengston and Roth (2008)	69.9	82.7	75.8	74.5	88.3	80.8
ACE2004-nwire						
This paper	75.1	84.6	79.6	74.1	87.3	80.2
Haghighi and Klein (2009)	75.9	77.0	76.5	74.5	79.4	76.9
Poon and Domingos (2008)	70.5	71.3	70.9	–	–	–
Finkel and Manning (2008)	58.5	78.7	67.1	65.2	86.8	74.5
MUC6-Test						
This paper	69.1	90.6	78.4	63.1	90.6	74.4
Haghighi and Klein (2009)	77.3	87.2	81.9	67.3	84.7	75.0
Poon and Domingos (2008)	75.8	83.0	79.2	–	–	–
Finkel and Manning (2008)	55.1	89.7	68.3	49.7	90.9	64.3

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

EMORY UNIVERSITY

Performance

- Sieves comparison

Table 8
Cumulative performance as sieves are added to the system.

	MUC			B^3			CEAF-ϕ_4			BLANC			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Sieve 1	8.7	72.7	15.5	32.4	96.4	48.5	50.6	15.4	23.7	57.2	80.3	60.2	29.2
+ Sieve 2	29.5	71.8	41.9	46.4	90.4	61.4	51.8	23.8	32.6	63.0	82.2	67.8	45.3
+ Sieve 3	29.7	71.2	41.9	46.7	90.1	61.5	51.6	24.0	32.7	63.0	82.0	67.8	45.4
+ Sieve 4	30.2	71.0	42.3	47.1	89.9	61.8	51.5	24.1	32.9	63.2	81.7	68.0	45.7
+ Sieve 5	34.4	66.1	45.2	51.5	86.6	64.6	50.8	27.6	35.8	64.1	80.8	68.8	48.5
+ Sieve 6	34.9	65.8	45.6	51.9	86.1	64.8	50.4	27.8	35.9	64.2	80.6	68.9	48.8
+ Sieve 7	35.8	64.0	45.9	53.3	85.0	65.5	49.8	28.9	36.6	64.4	80.3	69.1	49.3
+ Sieve 8	36.2	63.5	46.1	53.7	84.5	65.7	49.4	29.1	36.6	64.6	79.9	69.2	49.5
+ Sieve 9	36.7	63.2	46.5	54.2	84.0	65.9	49.2	29.4	36.8	64.7	79.5	69.2	49.7
+ Sieve 10	59.6	60.9	60.3	68.6	73.3	70.9	47.5	46.2	46.9	73.5	79.3	76.0	59.3

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

Performance

- Feature groups

Table 9
Contribution of each feature group. This is an ablative experiment, that is, each feature group is analyzed by removing it from the complete system listed in the first row.

	MUC			B^3			CEAF-ϕ_4			BLANC			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Complete system	59.6	60.9	60.3	68.6	73.3	70.9	47.5	46.2	46.9	73.5	79.3	76.0	59.3
– Number	57.0	56.4	56.7	66.2	68.6	67.4	45.6	46.2	45.9	67.6	72.6	69.7	56.7
– Gender	59.3	60.2	59.7	68.2	72.3	70.2	47.2	46.3	46.7	72.6	77.8	74.9	58.9
– Animacy	58.2	58.6	58.4	67.8	71.6	69.6	47.1	46.8	47.0	71.6	77.3	74.0	58.3
– NE	58.5	60.4	59.5	67.5	73.3	70.3	47.6	45.7	46.6	72.3	78.8	75.1	58.8

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Stanford Sieve System

EMORY UNIVERSITY

Error Analysis

Table 11
Distribution of errors.

Error type	Percentage
Semantics, discourse	41.7
Pronominal resolution errors	28.7
Non-referential mentions	14.8
Event mentions	6.1
Miscellaneous	8.7

Table 12
Examples of errors in each class. The mention to be resolved is in **boldface**, its correct antecedent is in *italics*, and we underlined <u>the incorrect antecedent</u> from our system result.

Error type	Example
Semantics, discourse	• Lincoln's parent company, American Continental Corp., entered bankruptcy - law proceedings this April 13, and regulators seized *the thrift* the next day. . . . Mr. Keating has filed his own suit, alleging that **his property** was taken illegally. • *New pictures* reveal the sheer power of that terrorist bomb . . . In **these photos obtained by NBC News**, the damage much larger than first imagined . . . • Of all the one-time expenses incurred by a corporation or professional firm, few are larger or longer term than *the purchase of real estate or the signing of a commercial lease* . . . To take full advantage of the financial opportunities in **this commitment**, . . .
Pronominal resolution errors	Under the laws of <u>the land</u>, *the ANC* remains an illegal organization , and **its** headquarters are still in Lusaka, Zambia.
Non-referential mentions	When <u>you</u> become a federal judge, all of a sudden **you** are relegated to a paltry sum.
Event mentions	"*Support* the troops, not <u>the regime</u>" **That** 's a noble idea until you're supporting the weight of an armoured vehicle on your chest.
Miscellaneous (inconsistent annotations, parser or NER errors, enumerations)	• Inconsistent annotation - Inclusion of *'s*: . . . that's without adding in [*Business Week 's*] charge . . . Small wonder that [*Britain*] 's Labor Party wants credit controls. • Parser or NER error: Um alright uh *Mister Zalisko* do **you** know anything from your personal experience of having been on the cruise as to what happened? – *Mister Zalisko* is not recognized as a PERSON • Enumerations: This year, the economies of the five large special economic zones, namely, Shenzhen, <u>Zhuhai, Shantou, Xiamen</u> and Hainan, have maintained strong growth momentum. . . . A three dimensional traffic frame in **Zhuhai** has preliminarily taken shape and the investment environment improves daily.

Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

Topics yet to be Explored

- Common noun coreference resolution
- Coreference resolution with NER integration
- Wildcard pronoun coreference resolution
- Object hierarchy handling (i.e. gems, diamonds, rubies)

Wildcard Pronoun Resolution

EMORY UNIVERSITY

I. Relatives & Superlatives
- Tome gives more than Sam.
- What is the smallest?

II. Quantity Modifiers
- He gave 3 to Jessica.
- Henry has none.

III. Abstract Quantities
- I have only a little.
- He ate a few.

IV. Unit-like Pronouns
- Johnny has the same amount as Mike
- She only received a portion

V. Inclusive & Status Pronoun
- We have the rest.
- They brought home the remaining.

Q & A