Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution

Sam Wiseman, Alexander M. Rush,
Stuart M. Shieber, and Jason Weston

HARVARD
School of Engineering
and Applied Sciences

Facebook AI Research

A Preliminary Example (CoNLL Dev Set, wsj/2404)

Cadillac posted a 3.2% increase despite new competition from Lexus, the fledgling luxury-car division of Toyota Motor Corp. Lexus sales weren't available; the cars are imported and Toyota reports their sales only at month-end.

With Coreferent Mentions Annotated

Cadillac posted a 3.2% increase despite new competition from [Lexus, the fledgling luxury-car division of [Toyota Motor Corp]] . [Lexus] sales weren't available; the cars are imported and [Toyota] reports [their] sales only at month-end.

Summary of (Informal) Terminology

Cadillac posted a 3.2% increase despite new competition from [Lexus, the fledgling luxury-car division of [Toyota Motor Corp]] . [Lexus] sales weren't available; the cars are imported and [Toyota] reports [their] sales only at month-end.

- **mention**: a span of text that can refer or be referred to
- **anaphoric**: a mention is anaphoric if it is coreferent with a previous mention
- **antecedent**: a mention to which an anaphoric mention refers

Mention Ranking [Denis and Baldridge 2008; Bengtson and Roth 2008]

- Model each mention x as having a single "true" antecedent
- Score potential antecedents y of x with scoring function $s(x, y)$
- If only clusters annotated, "true" antecedent a latent variable [Yu and Joachims 2009; Chang et al. 2013; Durrett and Klein 2013]

$s(x, y_1) = 0.4$

$s(x, y_2) = 0.9$

... [Lexus] sales weren't available ... [Toyota] reports [their]

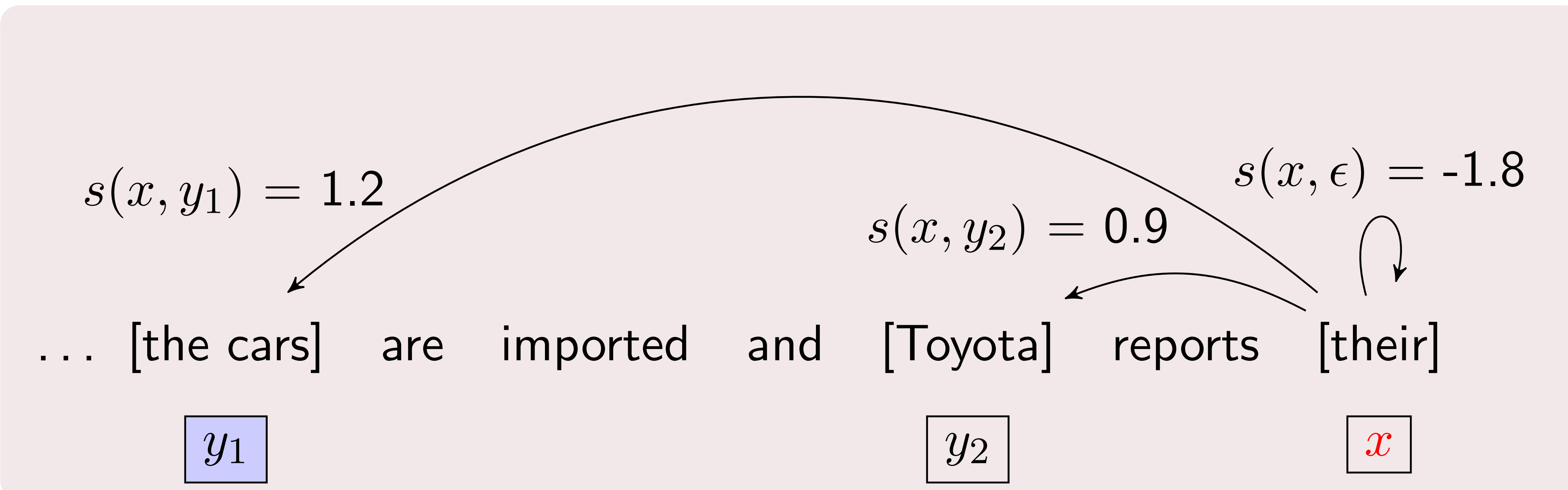
$\boxed{y_1}$ $\boxed{y_2}$ $\boxed{x}$

But Wait: Non-Anaphoric Mentions

[Cadillac] posted a [3.2% increase] despite [new competition from [Lexus, the fledgling luxury-car division of [Toyota Motor Corp]]] . [[Lexus] sales] weren't available; [the cars] are imported and [Toyota] reports [[their] sales] only at [month-end] .

Mention Ranking II

- Also score possibility that x non-anaphoric, denoted by $y = \epsilon$
- Again predict $y^* = \arg\max_{y \in \mathcal{Y}(x)} s(x, y)$



Mention Ranking III

- Common to use scoring function $s_{\text{lin}}(x, y) \triangleq \boldsymbol{w}^{\mathsf{T}} \widetilde{\boldsymbol{\phi}}(x, y)$
- Can duplicate features for a more flexible model:

$$s_{\text{lin+}}(x, y) \triangleq \begin{cases} \boldsymbol{u}^{\mathsf{T}} \begin{bmatrix} \widetilde{\phi}_{\text{a}}(x) \\ \widetilde{\phi}_{\text{p}}(x, y) \end{bmatrix} & \text{if } y \neq \epsilon \\ \boldsymbol{v}^{\mathsf{T}} \widetilde{\phi}_{\text{a}}(x) & \text{if } y = \epsilon \end{cases}$$

- $\widetilde{\phi}_{\text{a}}$ features on mention context (capture anaphoricity info)
- $\widetilde{\phi}_{\text{p}}$ features on mention, antecedent pair (capture pairwise affinity)
- Above equivalent to model of Durrett and Klein [2013]

Mention Ranking III

- Common to use scoring function $s_{\mathrm{lin}}(x, y) \triangleq \boldsymbol{w}^{\mathsf{T}} \widetilde{\boldsymbol{\phi}}(x, y)$
- Can duplicate features for a more flexible model:

$$s_{\mathrm{lin+}}(x, y) \triangleq \begin{cases} \boldsymbol{u}^{\mathsf{T}} \begin{bmatrix} \widetilde{\boldsymbol{\phi}}_{\mathrm{a}}(x) \\ \widetilde{\boldsymbol{\phi}}_{\mathrm{p}}(x, y) \end{bmatrix} & \text{if } y \neq \epsilon \\ \boldsymbol{v}^{\mathsf{T}} \widetilde{\boldsymbol{\phi}}_{\mathrm{a}}(x) & \text{if } y = \epsilon \end{cases}$$

- $\widetilde{\boldsymbol{\phi}}_{\mathrm{a}}$ features on mention context (capture anaphoricity info)
- $\widetilde{\boldsymbol{\phi}}_{\mathrm{p}}$ features on mention, antecedent pair (capture pairwise affinity)
- Above equivalent to model of Durrett and Klein [2013]

Mention Ranking III

- Common to use scoring function $s_{\text{lin}}(x, y) \triangleq \boldsymbol{w}^{\mathsf{T}} \widetilde{\boldsymbol{\phi}}(x, y)$
- Can duplicate features for a more flexible model:

$$s_{\text{lin+}}(x, y) \triangleq \begin{cases} \boldsymbol{u}^{\mathsf{T}} \begin{bmatrix} \widetilde{\boldsymbol{\phi}}_{\text{a}}(x) \\ \widetilde{\boldsymbol{\phi}}_{\text{p}}(x,y) \end{bmatrix} & \text{if } y \neq \epsilon \\ \boldsymbol{v}^{\mathsf{T}} \widetilde{\boldsymbol{\phi}}_{\text{a}}(x) & \text{if } y = \epsilon \end{cases}$$

- $\widetilde{\boldsymbol{\phi}}_{\text{a}}$ features on mention context (capture anaphoricity info)
- $\widetilde{\boldsymbol{\phi}}_{\text{p}}$ features on mention, antecedent pair (capture pairwise affinity)
- Above equivalent to model of Durrett and Klein [2013]

Non-Anaphoric Mentions Still Problematic

[Cadillac] posted a [3.2% increase] despite [new competition from [Lexus, the fledgling luxury-car division of [Toyota Motor Corp]]] . [[Lexus] sales] weren't available; [the cars] are imported and [Toyota] reports [[their] sales] only at [month-end] .

Ratio of non-anaphoric to anaphoric mentions in CoNLL train set over 3.5:1!

Problems with Simple Features

[Cadillac] posted a [3.2% increase] despite [new competition from [Lexus, the fledgling luxury-car division of [Toyota Motor Corp]]] . [[Lexus] sales] weren't available; [the cars] are imported and [Toyota] reports [[their] sales] only at [month-end] .

Misleading Head Matches

[Lexus **sales**] and [their **sales**] not coreferent!

Problems with Simple Features

[Cadillac] posted a [3.2% increase] despite [new competition from [Lexus, the fledgling luxury-car division of [Toyota Motor Corp]]] . [[Lexus] sales] weren't available; [the cars] are imported and [Toyota] reports [[their] sales] only at [month-end] .

Misleading Number Matches

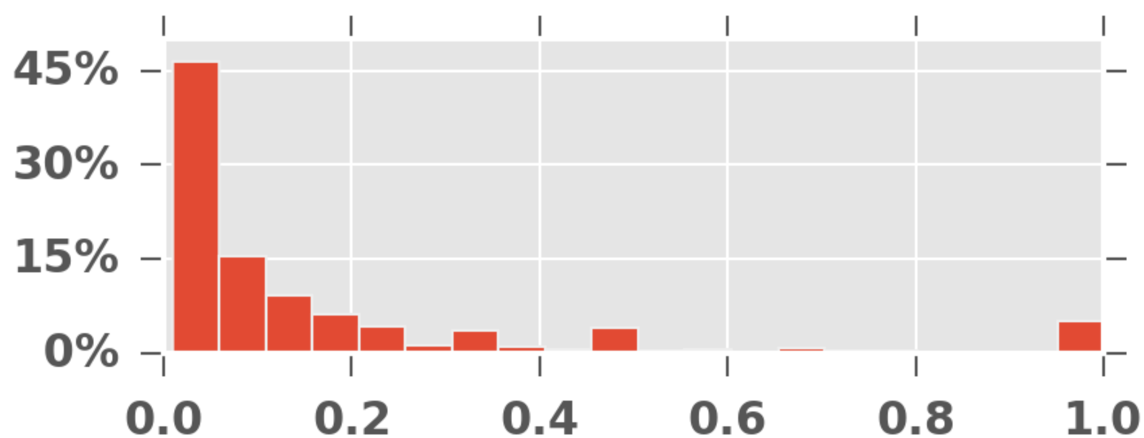
[the **cars**] and [**their**] not coreferent!

Simple Antecedent/Pairwise Features Not Discriminative

E.g., is [Lexus sales] the antecedent of [their sales]?

- Common pairwise features: String/Head Match, Sentences Between, Mention-Antecedent Numbers/Heads/Genders, etc.

$$\phi_{\mathrm{p}}(\text{[their sales],[Lexus sales]}) = \left\{ \begin{array}{c} \text{string-match=false} \\ \text{head-match=true} \\ \text{sentences-between=0} \\ \text{ment-ant-numbers=plur.,plur.} \\ \vdots \end{array} \right\}$$



Dealing with the Feature Problem

Finding discriminative features is a major challenge for coreference systems [Fernandes et al. 2012; Durrett and Klein 2013]

- Typical to define (or search for) feature conjunction-schemes to improve predictive performance [Fernandes et al. 2012; Durrett and Klein 2013; Björkelund and Kuhn 2014]. For instance:
  - string-match$(x, y) \wedge$ type$(x) \wedge$ type(y) [Durrett and Klein 2013], where

$$\text{type}(x) = \begin{cases} \text{Nom.} & \text{if } x \text{ is nominal} \\ \text{Prop.} & \text{if } x \text{ is proper} \\ \text{citation-form}(x) & \text{if } x \text{ is pronominal} \end{cases}$$

  - substring-match(head$(x), y) \wedge$ substring-match$(x$, head$(y)) \wedge$ coarse-type$(y) \wedge$ coarse-type(x) [Björkelund and Kuhn 2014]
- Not just a problem for Mention Ranking systems!

Dealing with the Feature Problem

Finding discriminative features is a major challenge for coreference systems [Fernandes et al. 2012; Durrett and Klein 2013]

- Typical to define (or search for) feature conjunction-schemes to improve predictive performance [Fernandes et al. 2012; Durrett and Klein 2013; Björkelund and Kuhn 2014]. For instance:
    - string-match(x, y) $\wedge$ type(x) $\wedge$ type(y) [Durrett and Klein 2013], where

$$\text{type}(x) = \begin{cases} \text{Nom.} & \text{if } x \text{ is nominal} \\ \text{Prop.} & \text{if } x \text{ is proper} \\ \text{citation-form}(x) & \text{if } x \text{ is pronominal} \end{cases}$$

    - substring-match(head$(x), y)$ $\wedge$ substring-match$(x,$ head$(y))$ $\wedge$ coarse-type(y) $\wedge$ coarse-type(x) [Björkelund and Kuhn 2014]
- Not just a problem for Mention Ranking systems!

Our Approach

Motivation: Current conjunction schemes perhaps not optimal, and in any case hard to scale as more features added.

Accordingly, we:

- Develop a model that learns good representations automatically
- Use only raw, unconjoined features
- Introduce pre-training scheme to improve quality of learned representations

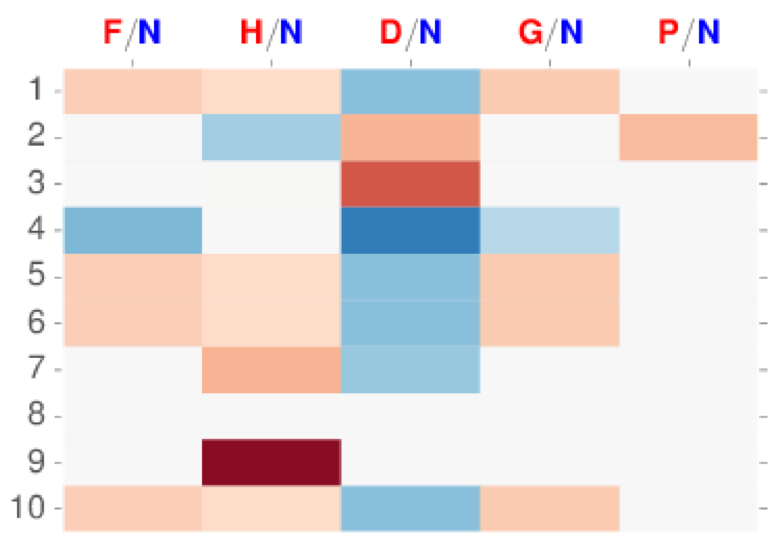
Extending the Piecewise Model I

Goal: learn higher order feature representations

We first define the following nonlinear feature representations:

$$\boldsymbol{h}_{\mathrm{a}}(x) \triangleq \tanh(\boldsymbol{W}_{\mathrm{a}}\,\boldsymbol{\phi}_{\mathrm{a}}(x) + \boldsymbol{b}_{\mathrm{a}})$$

$$\boldsymbol{h}_{\mathrm{p}}(x, y) \triangleq \tanh(\boldsymbol{W}_{\mathrm{p}}\,\boldsymbol{\phi}_{\mathrm{p}}(x, y) + \boldsymbol{b}_{\mathrm{p}})$$



- **Here, $\phi_{\mathrm{a}}, \phi_{\mathrm{p}}$ are raw, unconjoined features!**

Extending the Piecewise Model II

Use the scoring function

$$s(x,y) \triangleq \begin{cases} \boldsymbol{u}^\mathsf{T} \boldsymbol{g}\left(\begin{bmatrix} \boldsymbol{h}_\mathrm{a}(x) \\ \boldsymbol{h}_\mathrm{p}(x,y) \end{bmatrix}\right) + u_0 & \text{if } y \neq \epsilon \\ \boldsymbol{v}^\mathsf{T} \boldsymbol{h}_\mathrm{a}(x) + v_0 & \text{if } y = \epsilon \end{cases}$$

$(\boldsymbol{g}_1)$ If $\boldsymbol{g}$ is identity, obtain version of $s_{\mathrm{lin+}}$ with nonlinear features.

$(\boldsymbol{g}_2)$ If $\boldsymbol{g}$ is an additional hidden layer, further encourage nonlinear interactions between $\boldsymbol{h}_\mathrm{a}, \boldsymbol{h}_\mathrm{p}$

Training Objective

To train, we use the following margin-based loss:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \max_{\hat{y} \in \mathcal{Y}(x_n)} \Delta(x_n, \hat{y})(1 + s(x_n, \hat{y}) - s(x_n, y_n^{\ell})) + \lambda ||\boldsymbol{\theta}||_1$$

- y_n^{ℓ} a latent antecedent: equal to highest scoring antecedent in same cluster (or ϵ) [Yu and Joachims 2009; Fernandes et al. 2012; Chang et al. 2013; Durrett and Klein 2013]
- Slack-rescale with a mistake-specific cost function $\Delta(x_n, \hat{y})$
- Note that even if s were linear, would still be non-convex!

Pre-training Subtasks I

Two very natural subtasks for pre-training $\boldsymbol{h}_{\mathrm{p}}$ and $\boldsymbol{h}_{\mathrm{a}}$

Antecedent Ranking

Predict antecedents of known anaphoric mentions with scoring function

$$s_{\mathrm{p}}(x, y) \triangleq \boldsymbol{u}_{\mathrm{p}}{}^{\mathsf{T}} \boldsymbol{h}_{\mathrm{p}}(x, y) + \upsilon_0$$

$s(x, y_1) = 0.4$

$s(x, y_2) = 0.9$

. . . [Lexus] sales weren't available . . . [Toyota] reports [their]

$\boxed{y_1}$ $\boxed{y_2}$ $\boxed{x}$

- Very similar to "gold mention" version of coreference task (but slightly easier)

Pre-training Subtasks II

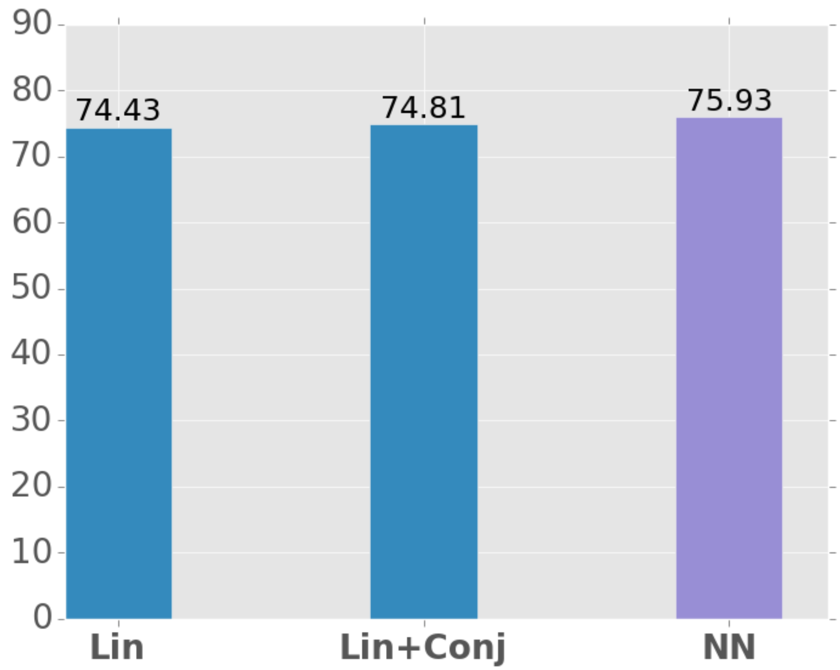
Anaphoricity Detection

Predict anaphoricity of mentions with scoring function

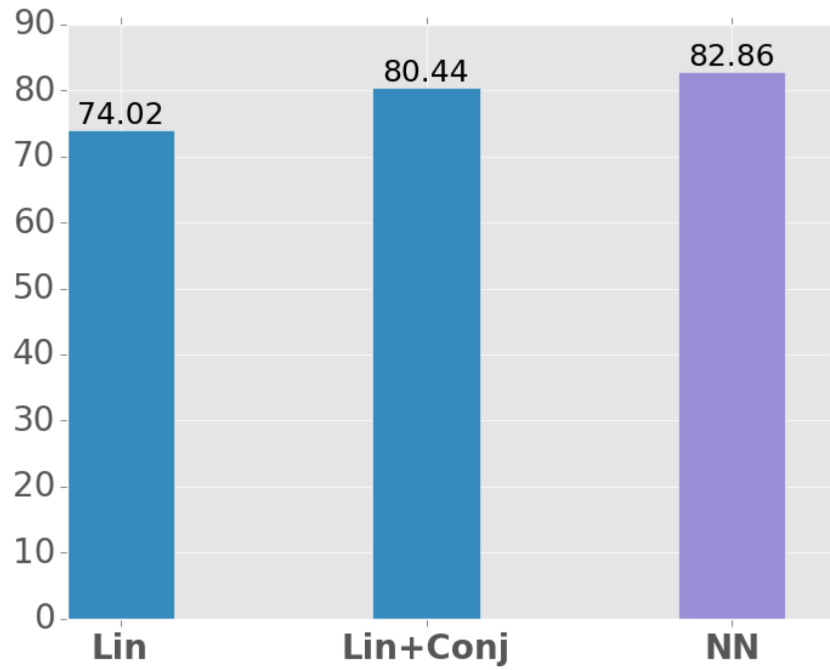
$$s_{\mathrm{a}}(x) \triangleq \boldsymbol{v}_{\mathrm{a}}{}^{\mathsf{T}} \boldsymbol{h}_{\mathrm{a}}(x) + \nu_0$$

- Anaphoricity/Singleton detection has a long history in coreference resolution.
  - Generally an initial step in a pipeline [Ng and Cardie 2002; Rahman and Ng 2009; Recasens et al. 2013; Lee et al. 2013; Ma et al. 2014]
- We use similar, margin-based objectives for both pre-training tasks

Subtask Performance (CoNLL Development Set)



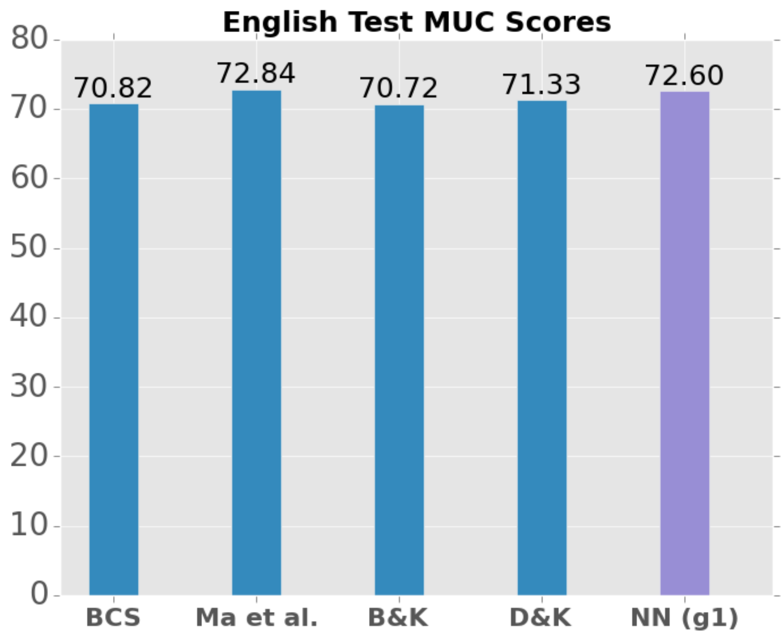
Anaphoricity Detection F_1 Score

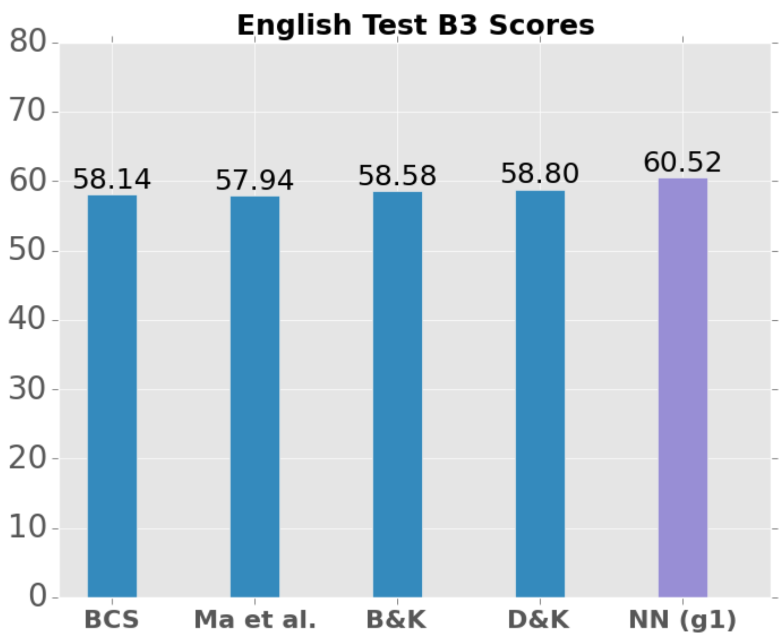


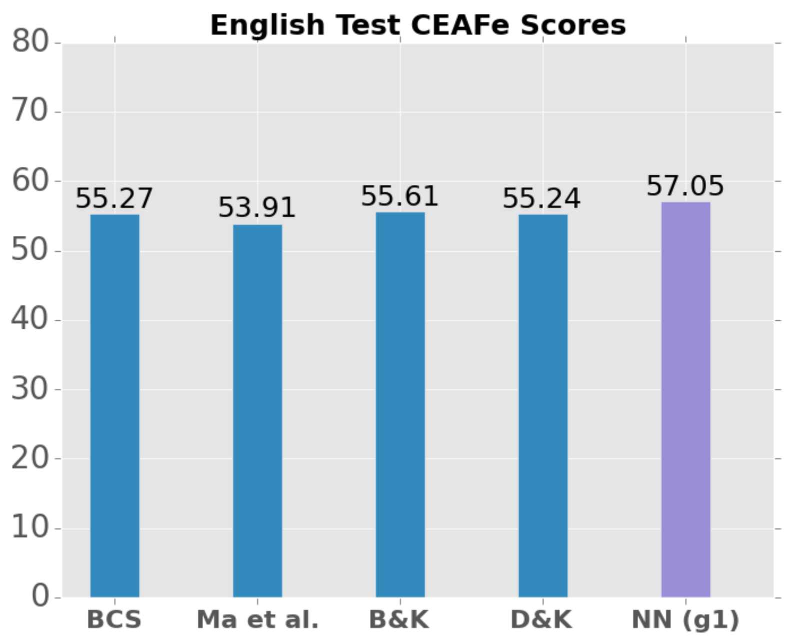
Antecedent Ranking Accuracy

- We compare with linear baseline with and without D&K (2013) conjunctions (over same features)

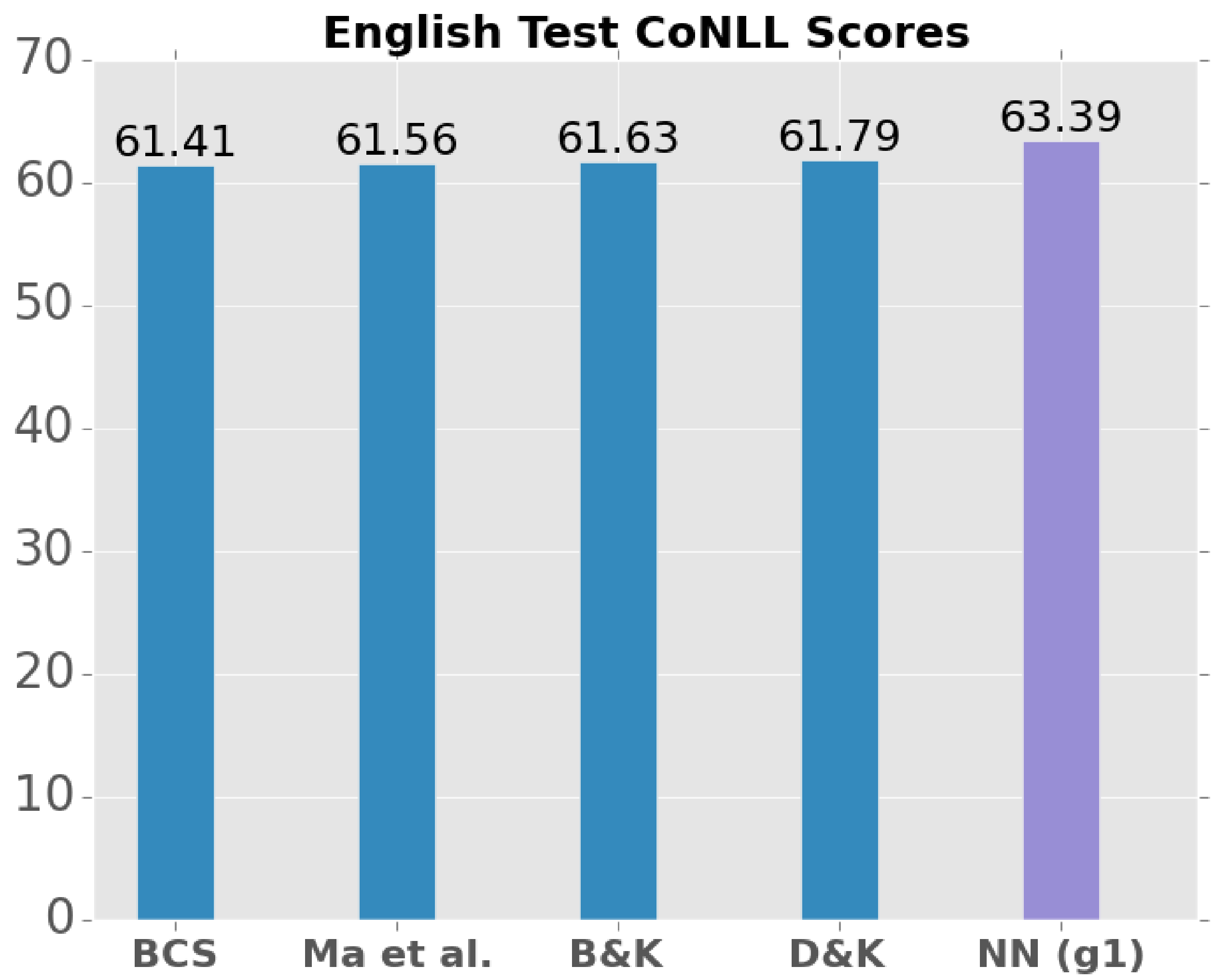
Main Results (MUC, B^3, $CEAF_e$)







Main Results (CoNLL Score)



Results on CoNLL 2012 English test set. We compare with (in order) Durrett and Klein [2013], Ma et al. [2014], Björkelund and Kuhn [2014], and Durrett and Klein [2014]. F_1 gains are significant ($p < 0.05$) compared with both B&K and D&K for all metrics.

Model Ablations

	Model	MUC	B^3	$CEAF_e$	CoNLL
(A)	1 Layer MLP	71.80	60.93	57.51	63.41
	2 Layer MLP	71.77	60.84	57.05	63.22
(B)	$\boldsymbol{g}_1$	71.92	61.06	57.59	63.52
	$\boldsymbol{g}_1$ + pre-train	72.74	61.77	58.63	64.38
(C)	$\boldsymbol{g}_2$	72.31	61.79	58.06	64.05
	$\boldsymbol{g}_2$ + pre-train	72.68	61.70	58.32	64.23

F_1 performance on CoNLL 2012 development set

- Table (A) examines whether separating $\boldsymbol{h}_{\mathrm{p}}, \boldsymbol{h}_{\mathrm{a}}$ (in first layer) actually helpful
- Tables (B) and (C) examine whether pre-training is helpful

Scaling to More Features

Model	Features	MUC	B^3	$CEAF_e$	CoNLL
Lin.		70.44	59.10	55.57	61.71
NN ($\boldsymbol{g}_2$)	BASIC	71.59	60.56	57.45	63.20
NN ($\boldsymbol{g}_1$)		71.86	60.9	57.90	63.55
Lin.		70.92	60.05	56.39	62.45
NN ($\boldsymbol{g}_2$)	BASIC+	72.68	61.70	58.32	64.23
NN ($\boldsymbol{g}_1$)		72.74	61.77	58.63	64.38

F_1 performance comparison between state-of-the-art linear mention-ranking model Durrett and Klein [2013] and our full models on CoNLL 2012 development set for different feature sets.

Discussion: What are we getting wrong?

Mention Ranking models make error analysis very simple:

- Highest percentage error ($\frac{736}{1000}$) on anaphoric mentions with no previous occurring head-match
  - e.g., [the team] and [the New York Giants]
- Highest number of errors ($\frac{1823}{9900}$) were mis-predicted links of pronominal mentions
  - Almost all were errors on pronouns that can be used pleonastically ("it", "you"), and almost all predicted antecedents were another instance of same pronoun.
  - An argument for more structure?
  - Note 30% of anaphoric pronominal mentions in CoNLL dev data are in pronoun-only clusters!

Summary

(1) Possible to achieve state-of-the-art performance with

- Very simple, local model and powerful scoring function
  - Note most recent state-of-the-art models non-local!
- Only raw, unconjoined features
- Over 1.5 pt increase over previous state-of-the-art in CoNLL score

(2) Separating anaphoricity and antecedent ranking (learned) representations beneficial

- Natural to pre-train on corresponding subtasks

Thanks!

Additional Slides

Discussion: preliminaries

Note that Mention Ranking models make error analysis very simple!

Three Kinds of Errors Possible

(Adopting terminology of Durrett and Klein [2013]):

(FL) **False Link** errors: predicting a mention to be anaphoric when it is non-anaphoric

(FN) **False New** errors: predicting a mention to be non-anaphoric when it is anaphoric

(WL) **Wrong Link** errors: predicting an incorrect antecedent for an anaphoric mention

Discussion: What are we getting wrong?

	Singleton		1st in clust.		Anaphoric	
	FL	#	FL	#	FN + WL	#
Ment. w/ prev. head match	817	8.2K	147	0.8K	700 + 318	4.7K
Ment. w/o prev. head match	86	19.8K	41	2.4K	677 + 59	1.0K
Pronominal mentions	948	2.6K	257	0.5K	434 + 875	7.3K

Largest % error on anaphoric mentions with no previous head match

- The classic "hard" coreference case, presumably requiring knowledge, understanding

But make most errors (by far) on pronouns!

All Features

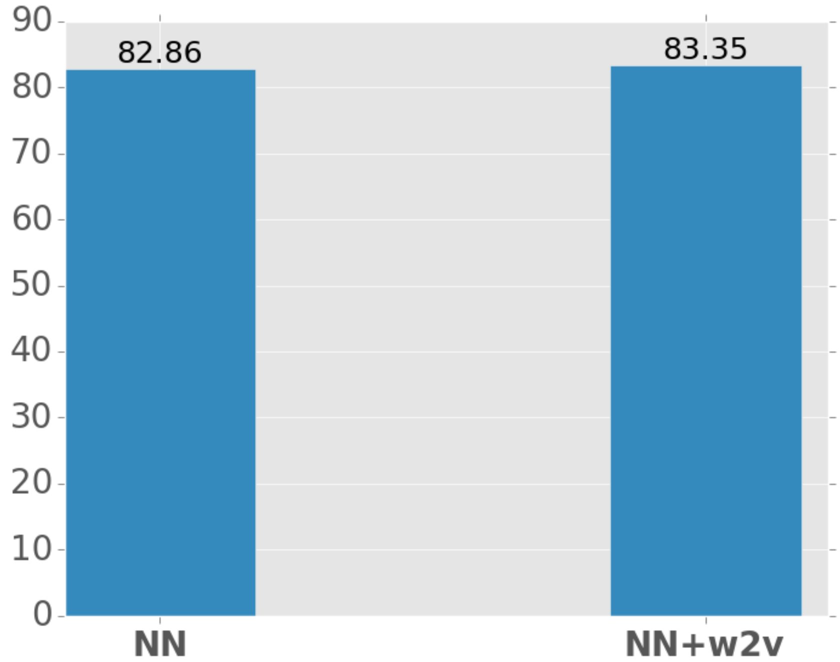
Mention Features (ϕ_a)
Mention Head
Mention First Word
Mention Last Word
Word Preceding Mention
Word Following Mention
Words in Mention
Mention Synt. Ancestry
Mention Type
Mention Governor
Mention Sentence Index
Mention Entity Type
Mention Number
Mention Animacy
Mention Gender
Mention Person

Pairwise Features (ϕ_p)
ϕ_a(Mention); ϕ_a(Antecedent)
Mentions between Ment., Ante.
Sentences between Ment., Ante.
i-within-i
Same Speaker
Document Type
Ante., Ment. String Match
Ante. contains Ment.
Ment. contains Ante.
Ante. contains Ment. Head
Mention contains Ante. Head
Ante., Ment. Head Match
Ante., Ment. Synt. Ancestries; Numbers; Genders; Persons; Entity Types; Heads; Types

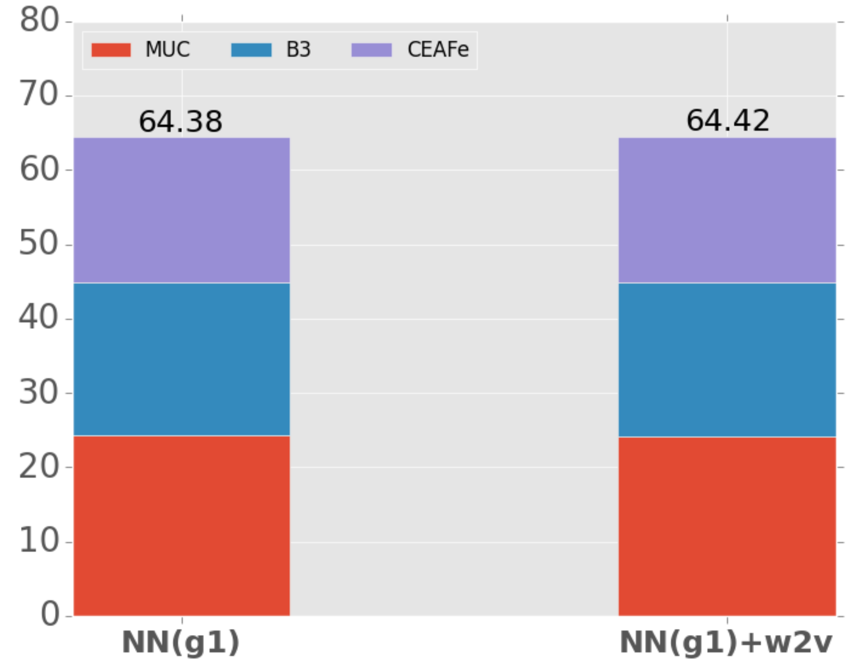
g_2 Error Analysis

NN ($\boldsymbol{g}_2$)	Singleton		1^{st} in clust.		Anaphoric	
	FL	#	FL	#	FN + WL	#
HM	770	8.2K	130	0.8K	803 + 306	4.7K
No HM	73	19.8K	39	2.4K	699 + 52	1.0K
Pron.	896	2.6K	249	0.5K	456 + 869	7.3K

Preliminary Embeddings Experiments



Antecedent Ranking Accuracy



CoNLL Scores on Dev

Embeddings Error Analysis

$\boldsymbol{g}_1$+w2v	Singleton		1st in clust.		Anaphoric	
	FL	#	FL	#	FN + WL	#
HM	801	8.2K	141	0.8K	742 + 333	4.7K
No HM	98	19.8K	51	2.4K	648 + 66	1.0K
Pron.	933	2.6K	251	0.5K	475 + 852	7.3K

Experimental Setup

- Used standard CoNLL 2012 English dataset experimental split
- Results scored with CoNLL 2012 scoring script v8.01
- Used Berkeley Coreference System [Durrett and Klein 2013] for mention extraction
- All optimization with Composite Mirror-Descent flavor of AdaGrad
- All hyperparameters (learning rates and regularization coefficients) tuned with grid-search on development set

Main Results (Full Table)

	MUC			B^3			$CEAF_e$			
	P	R	F_1	P	R	F_1	P	R	F_1	CoNLL
BCS	74.89	67.17	70.82	64.26	53.09	58.14	58.12	52.67	55.27	61.41
Ma et al.	81.03	66.16	**72.84**	66.90	51.10	57.94	68.75	44.34	53.91	61.56
B&K	74.30	67.46	70.72	62.71	54.96	58.58	59.40	52.27	55.61	61.63
D&K	72.73	69.98	71.33	61.18	56.60	58.80	56.20	54.31	55.24	61.79
NN($\boldsymbol{g}_2$)	76.96	68.10	72.26	66.90	54.12	59.84	59.02	53.34	56.03	62.71
NN($\boldsymbol{g}_1$)	76.23	69.31	72.60	66.07	55.83	**60.52**	59.41	54.88	**57.05**	**63.39**

Results on CoNLL 2012 English test set. We compare with (in order) Durrett and Klein [2013], Ma et al. [2014], Björkelund and Kuhn [2014], and Durrett and Klein [2014]. F_1 gains are significant ($p < 0.05$ under the bootstrap resample test Koehn [2004]) compared with both B&K and D&K for all metrics.

Eric Bengtson and Dan Roth. Understanding the Value of Features for Coreference Resolution. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 294–303. Association for Computational Linguistics, 2008.

Anders Björkelund and Jonas Kuhn. Learning structured perceptrons for coreference Resolution with Latent Antecedents and Non-local Features. ACL, Baltimore, MD, USA, June, 2014.

Kai-Wei Chang, Rajhans Samdani, and Dan Roth. A Constrained Latent Variable Model for Coreference Resolution. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 601–612, 2013.

Pascal Denis and Jason Baldridge. Specialized Models and Ranking for Coreference Resolution. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 660–669. Association for Computational Linguistics, 2008.

Greg Durrett and Dan Klein. Easy Victories and Uphill Battles in Coreference Resolution. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1971–1982, 2013.

Greg Durrett and Dan Klein. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. Transactions of the Association for Computational Linguistics, 2:477–490, 2014.

Eraldo Rezende Fernandes, Cícero Nogueira Dos Santos, and Ruy Luiz Milidiú. Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution. In Joint Conference on EMNLP and CoNLL-Shared Task, pages 41–48. Association for Computational Linguistics, 2012.

Philipp Koehn. Statistical Significance Tests for Machine Translation Evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 388–395. Citeseer, 2004.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic Coreference Resolution based on Entity-centric, Precision-ranked Rules. Computational Linguistics, 39(4):885–916, 2013.

Chao Ma, Janardhan Rao Doppa, J Walker Orr, Prashanth Mannem, Xiaoli Fern, Tom Dietterich, and Prasad Tadepalli. Prune-and-score: Learning for Greedy Coreference Resolution. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014.

Vincent Ng and Claire Cardie. Identifying Anaphoric and Non-anaphoric Noun Phrases to Improve Coreference Resolution. In Proceedings of the 19th international conference on Computational linguistics-Volume 1, pages 1–7. Association for Computational Linguistics, 2002.

Altaf Rahman and Vincent Ng. Supervised Models for Coreference Resolution. In Proceedings of the 2009 Conference on Empirical

Methods in Natural Language Processing: Volume 2-Volume 2, pages 968–977. Association for Computational Linguistics, 2009.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. The Life and Death of Discourse Entities: Identifying Singleton Mentions. In HLT-NAACL, pages 627–633, 2013.

Chun-Nam John Yu and Thorsten Joachims. Learning Structural SVMs with Latent Variables. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 1169–1176. ACM, 2009.