Learning Global Features for Coreference Resolution

Sam Wiseman, Alexander M. Rush, Stuart M. Shieber

HARVARD
School of Engineering
and Applied Sciences

Nominal-Nominal Coreference (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.

- **mention:** a syntactic unit 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

Nominal-Nominal Coreference (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.

- **mention:** a syntactic unit 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

Pronominal Coreference (CoNLL Dev Set, msnbc/0000)

Dan Abrams: *It's because of what [both of you] are doing to have things change. um and [I] think that is what's - Go ahead Linda.*

Linda Walker: *Well and uh thanks goes to [you] and to the media to help [us] .*

Erin Runnion: *Absolutely.*

Linda Walker: *Obviously [we] couldn't scream loud enough to bring the attention. So [our] hat is off to all of you as well.*

Mention Ranking [Denis and Baldridge 2008; Rahman and Ng 2009]

- Consider each mention x in turn
- Use *local* scoring function $f(x, y)$ to score compatibility of x and each *previous* mention y
- Also score possibility that x non-anaphoric: $f(x, \epsilon)$
- Predict $y^* = \arg\max_{y \in \mathcal{Y}(x)} f(x, y)$

y_1 y_2

. . . division of [Toyota Motor Corp]. [Lexus] sales weren't available;

$f(x, y_1) = 1.2$ $f(x, y_2) =$ -0.1

the cars are imported and [Toyota] . . . $f(x, \epsilon) = 0.4$

x

Parameterizing f

In previous work [Wiseman et al. 2015] we defined

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

- Feature embeddings $\boldsymbol{h}_\mathrm{a}$ and $\boldsymbol{h}_\mathrm{p}$ of sparse mention-level features $\boldsymbol{\phi}_\mathrm{a}(x)$ and sparse pairwise feature $\boldsymbol{\phi}_\mathrm{p}(x, y)$, respectively.

Let's Consider the Pronominal Example Again

Dan Abrams: *It's because of what [both of you] are doing to have things change. um and [I] think that is what's - Go ahead [Linda] .*

Linda Walker: *Well and uh thanks goes to [you] and to the media to help [us] .*

Erin Runnion: *Absolutely.*

Linda Walker: *Obviously [we] couldn't scream loud enough to bring the attention. So [our] hat is off to all of* **[you]** *as well.*

Let's Consider the Pronominal Example Again

Dan Abrams: *It's because of what [both of you] are doing to have things change. um and [I] think that is what's - Go ahead [Linda] .*

Linda Walker: *Well and uh thanks goes to* **[you]** *and to the media to help [us] .*

Erin Runnion: *Absolutely.*

Linda Walker: *Obviously [we] couldn't scream loud enough to bring the attention. So [our] hat is off to all of* **[you]** *as well.*

Let's Consider the Pronominal Example Again

Dan Abrams: *It's because of what [both of you] are doing to have things change. um and [I] think that is what's - Go ahead [Linda] .*

Linda Walker: *Well and uh thanks goes to* [**you**] *and to the media to help [us] .*

Erin Runnion: *Absolutely.*
Linda Walker: *Obviously [we] couldn't scream loud enough to bring the attention. So [our] hat is off to all of* [**you**] *as well.*

Mention Ranking with Global Features

Mention ranking has some important benefits:

- Simple, left-to-right inference; works well in practice

Idea: augment local ranking score with global term examining state of clusters/entities so far

$$\text{score}(x_n, y) = f(x_n, y) + g(x_n, y, \boldsymbol{z}_{1:n-1}).$$

- $\boldsymbol{z}$ is a clustering: $z_n = k$ iff mention x_n predicted to be in cluster k.

Mention Ranking with Global Features

Mention ranking has some important benefits:

- Simple, left-to-right inference; works well in practice

Idea: augment local ranking score with global term examining state of clusters/entities so far

$$\mathrm{score}(x_n, y) = f(x_n, y) + g(x_n, y, \boldsymbol{z}_{1:n-1}).$$

- $\boldsymbol{z}$ is a clustering: $z_n = k$ iff mention x_n predicted to be in cluster k.

Global Information Actually Useful?

Until recently, incorporating global information was unnecessary for obtaining SOTA performance [Durrett and Klein 2014; Clark and Manning 2015; Wiseman et al. 2015; Peng et al. 2015].

- Perhaps due to inherent difficulty of crafting discrete features for clusters, which can be of any size.
- Past global feature strategies either too coarse or too sparse:
  - Quantifier features [Luo 2005; Rahman and Ng 2011]: `most-true-gender-match`$(x, X^{(i)})$ `= true`
  - Concatenating mention-level features [Björkelund and Kuhn 2014]: {*the president*, *he*, *Obama*} $\Rightarrow$ `Common-Pron-Prop = true`

Our Strategy

Idea: Embed entity/cluster $X^{(i)}$ by running an RNN over its sequence of mentions

- Get a representation of $X^{(i)}$ after each mention is added
- Only need *mention-level* features!

RNN Reminder: Let $(\boldsymbol{m}_j)_{j=1}^{J}$ be a sequence of J input vectors $\boldsymbol{m}_j \in \mathbb{R}^D$, and let $\boldsymbol{h}_0 = \boldsymbol{0}$. Applying an RNN to any such sequence yields

$$\boldsymbol{h}_j \leftarrow \mathbf{RNN}(\boldsymbol{m}_j, \boldsymbol{h}_{j-1}; \boldsymbol{\theta}).$$

Our Strategy

Idea: Embed entity/cluster $X^{(i)}$ by running an RNN over its sequence of mentions

- Get a representation of $X^{(i)}$ after each mention is added
- Only need *mention-level* features!

RNN Reminder: Let $(\boldsymbol{m}_j)_{j=1}^J$ be a sequence of J input vectors $\boldsymbol{m}_j \in \mathbb{R}^D$, and let $\boldsymbol{h}_0 = \boldsymbol{0}$. Applying an RNN to any such sequence yields

$$\boldsymbol{h}_j \leftarrow \mathbf{RNN}(\boldsymbol{m}_j, \boldsymbol{h}_{j-1}; \boldsymbol{\theta}).$$

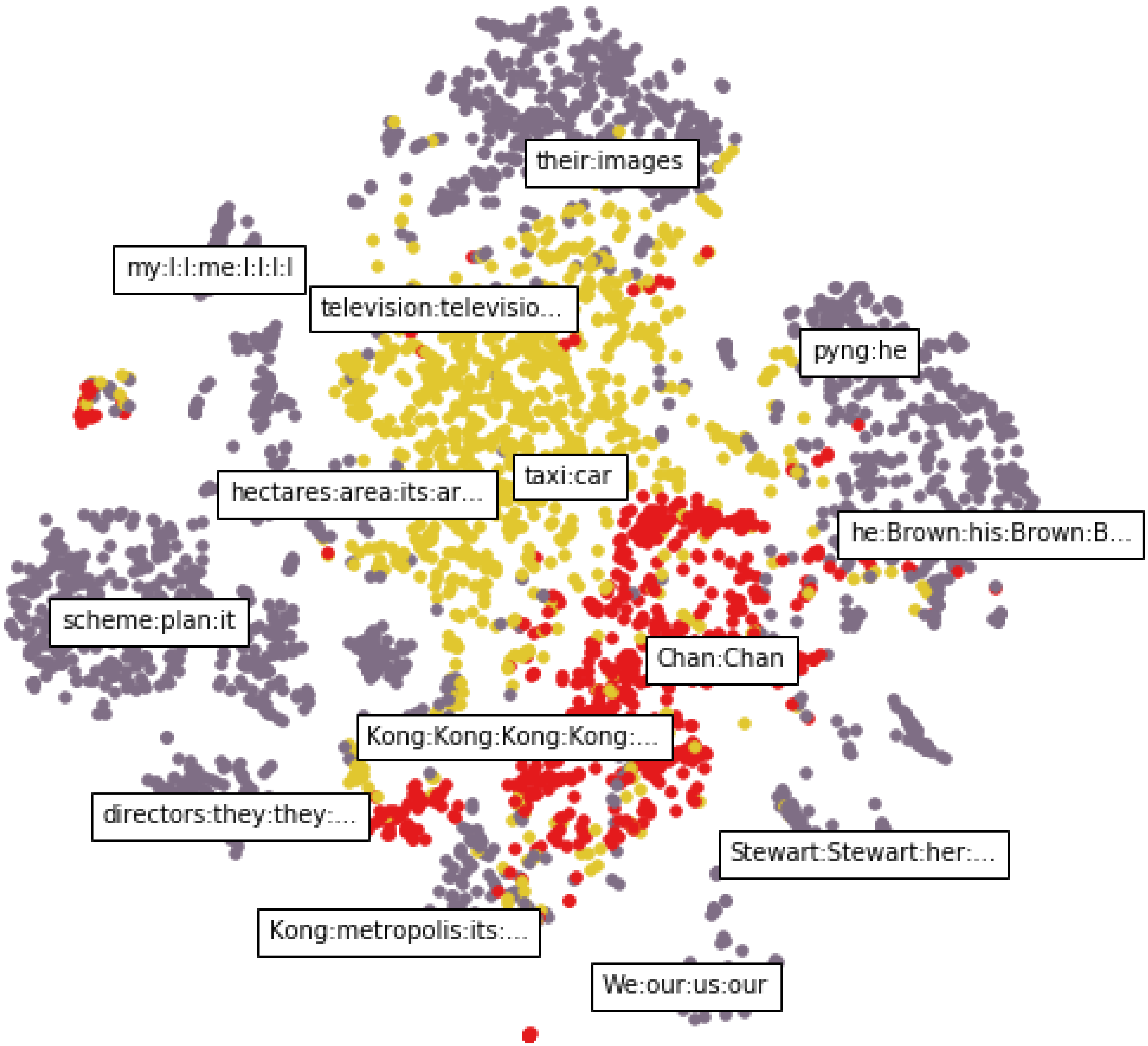
Cluster RNNs

Dan Abrams: *um and [I]*$_1$ *think that is what's - Go ahead [Linda]*$_2$*.*

Linda Walker: *Well and thanks goes to [you]*$_1$ *and to [the media]*$_3$ *to help [us]*$_4$*...So [our]*$_4$ *hat is off to all of [you]*$_5$*...*

$$\begin{array}{ccc} & X^{(1)} & \\ \boldsymbol{h}_1^{(1)} & \longrightarrow & \boldsymbol{h}_2^{(1)} \\ \uparrow & & \uparrow \\ \text{[I]} & & \text{[you]} \end{array} \qquad \begin{array}{c} X^{(2)} \\ \boldsymbol{h}_1^{(2)} \\ \uparrow \\ \text{[Linda]} \end{array} \qquad \begin{array}{c} X^{(3)} \\ \boldsymbol{h}_1^{(3)} \\ \uparrow \\ \text{[the media]} \end{array} \qquad \begin{array}{ccc} & X^{(4)} & \\ \boldsymbol{h}_1^{(4)} & \longrightarrow & \boldsymbol{h}_2^{(4)} \\ \uparrow & & \uparrow \\ \text{[us]} & & \text{[our]} \end{array}$$

Visualizing Cluster Embeddings



Defining g

We define:

$$g(x_n, y, \boldsymbol{z}_{1:n-1}) \triangleq \begin{cases} \boldsymbol{h}_{\mathrm{c}}(x_n)^{\mathsf{T}} \boldsymbol{h}_{<n}^{(z_y)} & \text{if } y \neq \epsilon \\ \boldsymbol{q}^{\mathsf{T}} \tanh\left(\boldsymbol{W}_s \begin{bmatrix} \boldsymbol{\phi}_{\mathrm{a}}(x) \\ \sum_{m=1}^{M} \boldsymbol{h}_{<n}^{(m)} \end{bmatrix} + \boldsymbol{b}_s\right) & \text{if } y = \epsilon, \end{cases}$$

where

- $\boldsymbol{h}_{<n}^{(z_y)}$ is state of RNN corresponding to y's cluster after consuming its last mention *before* x_n
- Intuitively, we express compatibility of x_n with a potential cluster with a dot-product

g in Action

Dan Abrams: *um and [I]$_1$ think that is what's - Go ahead [Linda]$_2$.*

Linda Walker: *Well and thanks goes to [you]$_1$ and to [the media]$_3$ to help [us]$_4$...So [our]$_4$ hat is off to all of [you]$_5$...*

$X^{(1)}$

$$\begin{tikzcd} \boldsymbol{h}_1^{(1)} \arrow[r] & \boldsymbol{h}_2^{(1)} \\ {[\text{I}]} \arrow[u] & {[\text{you}]} \arrow[u] \end{tikzcd}$$

$X^{(2)}$

$$\begin{tikzcd} \boldsymbol{h}_1^{(2)} \\ {[\text{Linda}]} \arrow[u] \end{tikzcd}$$

$X^{(3)}$

$$\begin{tikzcd} \boldsymbol{h}_1^{(3)} \\ {[\text{the media}]} \arrow[u] \end{tikzcd}$$

$X^{(4)}$

$$\begin{tikzcd} \boldsymbol{h}_1^{(4)} \arrow[r] & \boldsymbol{h}_2^{(4)} \\ {[\text{us}]} \arrow[u] & {[\text{our}]} \arrow[u] \end{tikzcd}$$

[I], $\boldsymbol{h}_2^{(1)}$ [Linda], $\boldsymbol{h}_1^{(2)}$ [you], $\boldsymbol{h}_2^{(1)}$ [the media], $\boldsymbol{h}_1^{(3)}$ [us], $\boldsymbol{h}_2^{(4)}$ [our], $\boldsymbol{h}_2^{(4)}$ x_n = [you] ϵ, $\mathrm{NA}(x_n)$

Full Model

Define:

$$\text{score}(x_n, y) = f(x_n, y) + g(x_n, y, \boldsymbol{z}_{1:n-1})$$

Train with loss:

$$\sum_{n=1}^{N} \max_{\hat{y} \in \mathcal{Y}(x_n)} \Delta(x_n, \hat{y}) \left(1 + \text{score}(x_n, \hat{y}) - \text{score}(x_n, y_n^{\ell})\right)$$

- We use an LSTM to embed cluster states
- We train with *oracle* history: when predicting x_n, have access to $\boldsymbol{z}_{1:n-1}^{(o)}$
- For main results we simply use greedy inference at test-time

Main Results (F_1), English CoNLL 2012 Test Set

	MUC	B^3	CEAF$_e$	CoNLL
Björkelund & Kuhn (2014)	70.72	58.58	55.61	61.63
Martschat & Strube (2015)	72.17	59.58	55.67	62.47
Clark & Manning (2015)	72.59	60.44	56.02	63.02
Peng et al. (2015)	72.22	60.50	56.37	63.03
Wiseman et al. (2015)	72.60	60.52	57.05	63.39
This work	**73.42**	**61.50**	**57.70**	**64.21**

But see Clark and Manning (2016)!

Comparison with Baselines on CoNLL Development Set

	CoNLL
Mention Ranking	64.90
Mean Pooling, Oracle History	65.07
RNN, Greedy	65.47
RNN, Oracle History	65.90

Errors on Non-Anaphoric Mentions (Development Set)

	Non-Anaphoric (FL)		
	Nom. HM	Nom. No HM	Pron.
Mention Ranking	1061	130	1075
Mean Pooling, Oracle History	983	140	1011
RNN, Greedy	914	125	893
RNN, Oracle History	913	130	842
# Mentions	9.0K	22.2K	3.1K

An Example, CoNLL Development Set (wsj/2418)

"I had no idea I was getting in so deep," says Mr. Kaye, who founded Justin in 1982. Mr. Kaye had sold Capetronic Inc., a Taiwan electronics Maker, and retired, only to find he was bored. With Justin, he began selling toys and electronics made mostly in Hong Kong, beginning with Mickey Mouse radios. The company has grown -- to about 40 employees, from four initially, Mr. Kaye says. Justin has been profitable since 1986, adds the official, who shares **[his]** office... (nw/wsj/2418)

Conclusion

- With good representations, global information helps
- Most pronounced improvement on pronouns
- RNNs provide a simple, efficient way of learning cluster representations

Thanks!

All Features

Mention Features (ϕ_a)
Mention Head, First, Last Words
Word Preceding, Following Mention
Words in Mention
Mention Synt. Ancestry
Mention Type
Mention Governor
Mention Sentence Index
Mention Entity Type
Mention Number, Gender, Person
Mention Animacy
Document Genre
Speaker
Mention contains Speaker
Normalized Document Position of Mention

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. String Match with non-current Speaker

Anaphoric Mention Error Analysis

	Anaphoric (FN + WL)		
Model	Nom. HM	Nom. No HM	Pron.
MR	665+326	666+56	533+796
Avg, OH	781+300	641+60	578+744
RNN, GH	767+303	648+57	664+727
RNN, OH	750+289	648+52	611+686
# Mentions	4.7K	1.0K	7.3K

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

Kevin Clark and Christopher D. Manning. Entity-centric coreference resolution with model stacking. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), pages 1405–1415, 2015.

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

Xiaoqiang Luo. On Coreference Resolution Performance Metrics. In Proceedings of the conference on Human Language Technology and

Empirical Methods in Natural Language Processing, pages 25–32. Association for Computational Linguistics, 2005.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. A joint framework for coreference resolution and mention head detection. In Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL), pages 12–21, 2015.

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.

Altaf Rahman and Vincent Ng. Narrowing the modeling gap: A cluster-ranking approach to coreference resolution. J. Artif. Intell. Res. (JAIR), 40:469–521, 2011.

Sam Wiseman, Alexander M. Rush, Stuart M. Shieber, and Jason Weston. Learning anaphoricity and antecedent ranking features for coreference resolution. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), pages 1416–1426, 2015.