Reference and Style

Two Challenge Problems for Natural Language Processing

Jacob Eisenstein
@jacobeisenstein

February 20, 2019

What's next for natural language processing?

The view from 2015:

- "NLP is kind of like a rabbit in the headlights of the Deep Learning machine, waiting to be flattened."
  –Neil Lawrence
- "Our field is the domain science of language technology... The domain problems will not go away."
  –Chris Manning

[1] Peters et al. 2018; Devlin et al. 2018.

What's next for natural language processing?

The view from 2015:

- "NLP is kind of like a rabbit in the headlights of the Deep Learning machine, waiting to be flattened."
  –Neil Lawrence

- "Our field is the domain science of language technology... The domain problems will not go away."
  –Chris Manning

Four years later, domain-neutral methods continue to advance.[1] What's left for domain science?

---

[1] Peters et al. 2018; Devlin et al. 2018.

Two hard problems for NLP

- **Reference resolution**: understanding which parts of a text refer to the same underlying semantic entity.
- **Style**: producing text that is stylistically coherent and controlled.

What do we know about reference?

Reference resolution draws on a wide range of linguistic cues:

- morphosyntactic constraints;
- discourse structure;
- pragmatics;
- semantics;
- world knowledge.

Reference, morphology, and syntax

Some basic constraints:

(1) a. Albert asked Becky to help him.

Reference, morphology, and syntax

Some basic constraints:

(1) a. Albert asked Becky to help him.

b. **Alice** asked Becky to help **her**.

Reference, morphology, and syntax

Some basic constraints:

(1) a. Albert asked Becky to help him.
    b. **Alice** asked Becky to help **her**.
    c. Alice asked Becky to help **herself**.

Reference and discourse

Entities acquire and lose “salience” throughout a discourse.

(2) The doctor found **an old map** in the captain’s chest.
Jim found **an even older map** hidden on the shelf.
It described an island.

- Recency is the strongest predictor of salience.
- However, syntactic function (e.g., subject, object)[2] and discourse relations[3] also play a role.

---

[2] Grosz, Weinstein, and Joshi 1995.

[3] Kehler and Rohde 2013.

Reference and pragmatics

Maxim of quantity: "be as informative as necessary, but not more so."[4]

(3) **Boris** opened the letter.
Mr Badenov was not pleased.

[4] Grice 1975.

Reference and pragmatics

Maxim of quantity: "be as informative as necessary, but not more so."[4]

(3) **Boris** opened the letter.
Mr Badenov was not pleased.

(4) **[Asha and [her friends]]** visited New York.
They ate several kinds of pizza.

[4] Grice 1975.

Reference, semantics, and world knowledge

Some "uphill battles" involve semantics and world knowledge.[5]

(5) a. Charlene loaned Doris a book on Spanish.
She is always helping people.

b. Charlene loaned Doris a book on Spanish.
She is visiting Mexico next month.

[5] Durrett and Klein 2013.

Reference, semantics, and world knowledge

Some "uphill battles" involve semantics and world knowledge.[5]

(5) a. Charlene loaned Doris a book on Spanish.
She is always helping people.

b. Charlene loaned Doris a book on Spanish.
She is visiting Mexico next month.

(6) Apple CEO Tim Cook jetted into China to resolve a raft of problems in the firm's biggest growth market.

[5] Durrett and Klein 2013.

Reference, semantics, and world knowledge

Some "uphill battles" involve semantics and world knowledge.[5]

(5) a. Charlene loaned Doris a book on Spanish.
She is always helping people.

b. Charlene loaned Doris a book on Spanish.
She is visiting Mexico next month.

(6) Apple CEO Tim Cook jetted into China to resolve a raft of problems in the firm's biggest growth market.

[5] Durrett and Klein 2013.

Reference, semantics, and world knowledge

Some "uphill battles" involve semantics and world knowledge.[5]

(5) a. Charlene loaned Doris a book on Spanish.
She is always helping people.

b. Charlene loaned Doris a book on Spanish.
She is visiting Mexico next month.

(6) Apple CEO Tim Cook jetted into China to resolve a raft of problems in the firm's biggest growth market.

[5] Durrett and Klein 2013.

What do we know about reference?

Reference resolution draws on a wide range of linguistic cues:

- morphosyntactic constraints;
- discourse structure;
- pragmatics;
- semantics;
- world knowledge.

What do we know about reference?

Reference resolution draws on a wide range of linguistic cues:

- morphosyntactic constraints;
- discourse structure;
- pragmatics;
- semantics;
- world knowledge.

Can self-attention implicitly learn them all?

A brief history of coreference resolution

1995-1997 First large-scale annotated data and metrics[6]

2001 First machine learning approach: mention-pair classification[7]

mid 2000s "Fancy" machine learning methods like structure prediction and nonparametric Bayes[8]

---

[146]Chinchor and Robinson 1997 [7]Soon, Ng, and Lim 2001; [8]Haghighi and Klein 2010; [9]Pradhan et al. 2011; [10]H. Lee et al. 2013; [11]Durrett and Klein 2013; [12]S. J. Wiseman et al. 2015; [13]K. Lee, He, Lewis, et al. 2017a; [14]K. Lee, He, and Zettlemoyer 2018

A brief history of coreference resolution

1995-1997 First large-scale annotated data and metrics[6]

2001 First machine learning approach: mention-pair classification[7]

mid 2000s "Fancy" machine learning methods like structure prediction and nonparametric Bayes[8]

2011 Larger-scale OntoNotes data;[9] Stanford's rule-based system beats all previous machine learning methods.[10]

mid 2010s Features and loss function hacking;[11] early neural approaches.[12]

2017-2018 Neural mention-pair model[13] and ELMo.[14]

---

[146] Chinchor and Robinson 1997 [7] Soon, Ng, and Lim 2001; [8] Haghighi and Klein 2010; [9] Pradhan et al. 2011; [10] H. Lee et al. 2013; [11] Durrett and Klein 2013; [12] S. J. Wiseman et al. 2015; [13] K. Lee, He, Lewis, et al. 2017a; [14] K. Lee, He, and Zettlemoyer 2018

How good is it really?

- Coreference evaluation is "opaque" at best[15]
- GAP dataset: two names, one pronoun[16]

(7) When **she** gets into an altercation with *Queenie*, Fiona makes her act as Queenie's slave...

  - Systems must identify which name, if any, is referenced by the pronoun.
  - Examples are balanced between masculine and feminine pronouns (in OntoNotes, only 25% of pronouns are feminine!)

[15] Stoyanov et al. 2009.
[16] Webster et al. 2018.

GAP results

	F_1^M	F_1^F	$\frac{F_1^F}{F_1^M}$	F_1
S. Wiseman, Rush, and Shieber 2016	67.8	59.1	0.87	63.6
K. Lee, He, Lewis, et al. 2017b	67.7	60.0	0.89	64.0
Syntactic parallelism rule	69.4	64.4	0.93	66.9
Transformer LM[17]	59.6	56.6	0.95	62.3

- **Syntactic parallelism** baseline: match subjects with subjects, objects with objects.
- This beats all pre-trained systems!

[17] The transformer sees gold spans, so results are not comparable.

Why hasn't black-box learning cracked coref?

- Requires many types of linguistic reasoning, not just semantic similarity.
- Significant differences across genres.[18]
- Most approaches build on the **mention-pair model**: train a classifier to decide if pairs of mentions corefer.
  - Clearly implausible from a cognitive perspective.
  - Expensive in both computation and labeled data.

[18] Moosavi and Strube 2017.

The Referential Reader

A Recurrent Entity Network for Anaphora Resolution

Fei Liu[19] Luke Zettlemoyer Jacob Eisenstein

[19] FAIR intern from the University of Melbourne

The Referential Reader

Design principles:

- store entities in a fixed-size memory network;[20]
- perform coreference resolution online, with linear time complexity;
- explicitly track entity salience while reading;
- multitask with a language modeling objective.

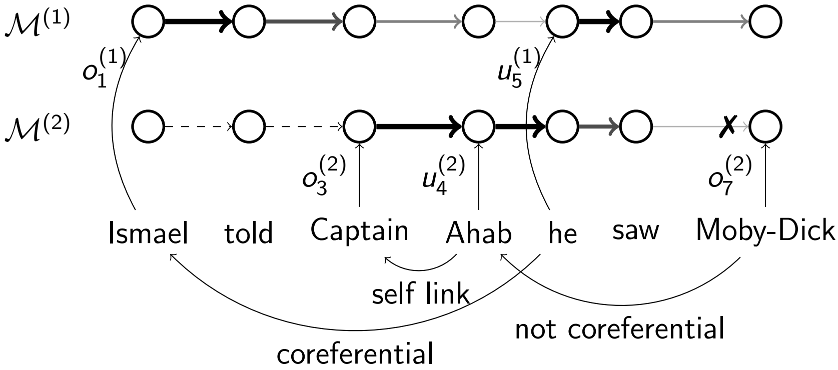
[20] Weston, Chopra, and Bordes 2015; Henaff et al. 2017.

The Referential Reader in action

At each token, **o**verwrite or **u**pdate an existing memory.

The Referential Reader in action

At each token, **o**verwrite or **u**pdate an existing memory.



These memory operations imply a coreference structure.

Memory structure

Each entry has a **k**ey, **v**alue,[21] and **s**alience.

$$\mathcal{M} = \{(\mathbf{k}^{(i)}, \mathbf{v}^{(i)}, s^{(i)})\}_{i=1}^{N}$$

[21] Miller et al. 2016.

Memory structure

Each entry has a **k**ey, **v**alue,[21] and **s**alience.

$$\mathcal{M} = \{(\mathbf{k}^{(i)}, \mathbf{v}^{(i)}, s^{(i)})\}_{i=1}^{N}$$

The memory is controlled by two operations:

- **Update** operations are compositional: after an update, the key and value are a function of the input and the previous key/value.
- **Overwrite** operations erase the current key and value.

Salience increases whenever a memory is accessed, and decreases otherwise.

[21] Miller et al. 2016.

Control flow

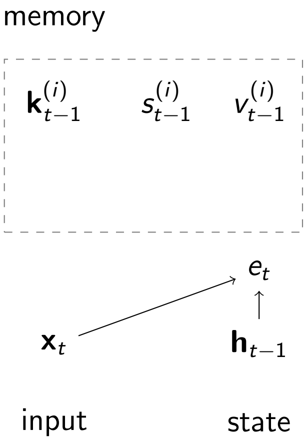
memory

$\mathbf{k}_{t-1}^{(i)}$ $s_{t-1}^{(i)}$ $v_{t-1}^{(i)}$

$\mathbf{x}_t$ $\mathbf{h}_{t-1}$

input state

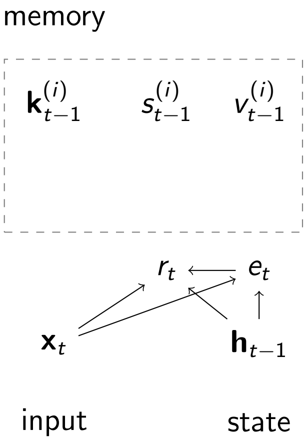
Control flow



- It is an entity mention?

$$e_t = \sigma(f_e(\mathbf{x}_t, \mathbf{h}_{t-1}))$$

Control flow



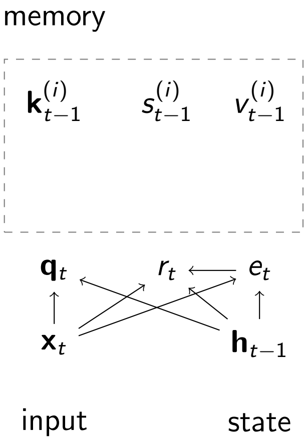
- It is an entity mention?

$$e_t = \sigma(f_e(\mathbf{x}_t, \mathbf{h}_{t-1}))$$

- Is it a reference?

$$r_t = e_t \times \sigma(f_r(\mathbf{x}_t, \mathbf{h}_{t-1}))$$

Control flow



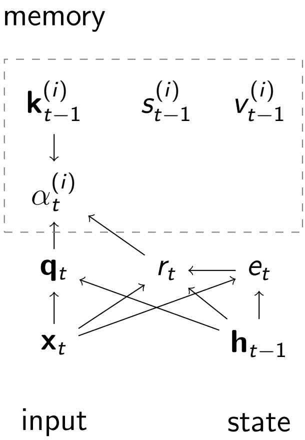
- It is an entity mention?

$$e_t = \sigma(f_e(\mathbf{x}_t, \mathbf{h}_{t-1}))$$

- Is it a reference?

$$r_t = e_t \times \sigma(f_r(\mathbf{x}_t, \mathbf{h}_{t-1}))$$

Control flow



- It is an entity mention?

$$e_t = \sigma(f_e(\mathbf{x}_t, \mathbf{h}_{t-1}))$$

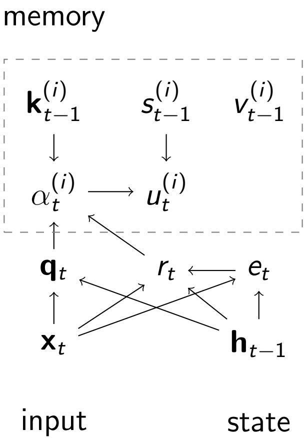
- Is it a reference?

$$r_t = e_t \times \sigma(f_r(\mathbf{x}_t, \mathbf{h}_{t-1}))$$

- Is it compatible with an existing memory?

$$\alpha_t^{(i)} = r_t \times \sigma(\mathbf{q}_t \cdot \mathbf{k}_{t-1}^{(i)})$$

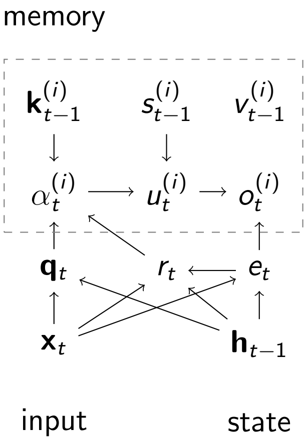
Control flow



- Update gate

$$u_t^{(i)} = \max(2s_{t-1}^{(i)}, \alpha_t^{(i)})$$

Control flow



- Update gate

$$u_t^{(i)} = \max(2s_{t-1}^{(i)}, \alpha_t^{(i)})$$

- Overwrite gate

$$o_t^{(i)} = (e_t - \sum_j u_t^{(j)}) \times \mathrm{GumbelSM}(\mathbf{s}_{t-1}, \tau)$$

Memory updates

- Candidate keys and values are computed from the input and hidden state,

$$\tilde{\mathbf{k}}_t = f_k(\mathbf{x}_t, \mathbf{h}_{t-1}), \quad \tilde{\mathbf{v}}_t = f_v(\mathbf{x}_t, \mathbf{h}_{t-1}).$$

Memory updates

- Candidate keys and values are computed from the input and hidden state,

$$\tilde{\mathbf{k}}_t = f_k(\mathbf{x}_t, \mathbf{h}_{t-1}), \quad \tilde{\mathbf{v}}_t = f_v(\mathbf{x}_t, \mathbf{h}_{t-1}).$$

- In an overwrite, the candidate **replaces** the current key and value; in an update, it is **composed** with the current key and value:

$$\begin{aligned}
\mathbf{k}_t =& o_t^{(i)}\tilde{\mathbf{k}}_t + u_t^{(i)} g_k(\mathbf{k}_{t-1}^{(i)}, \tilde{\mathbf{k}}_t) + (1 - o^{(i)} - u_t^{(i)})\mathbf{k}_{t-1}^{(i)} \\
\mathbf{v}_t =& o_t^{(i)}\tilde{\mathbf{v}}_t + u_t^{(i)} g_k(\mathbf{v}_{t-1}^{(i)}, \tilde{\mathbf{v}}_t) + (1 - o^{(i)} - u_t^{(i)})\mathbf{v}_{t-1}^{(i)}.
\end{aligned}$$

Memory updates

- Candidate keys and values are computed from the input and hidden state,

$$\tilde{\mathbf{k}}_t = f_k(\mathbf{x}_t, \mathbf{h}_{t-1}), \quad \tilde{\mathbf{v}}_t = f_v(\mathbf{x}_t, \mathbf{h}_{t-1}).$$

- In an overwrite, the candidate **replaces** the current key and value; in an update, it is **composed** with the current key and value:

$$\begin{aligned}
\mathbf{k}_t =& o_t^{(i)} \tilde{\mathbf{k}}_t + u_t^{(i)} g_k(\mathbf{k}_{t-1}^{(i)}, \tilde{\mathbf{k}}_t) + (1 - o^{(i)} - u_t^{(i)}) \mathbf{k}_{t-1}^{(i)} \\
\mathbf{v}_t =& o_t^{(i)} \tilde{\mathbf{v}}_t + u_t^{(i)} g_k(\mathbf{v}_{t-1}^{(i)}, \tilde{\mathbf{v}}_t) + (1 - o^{(i)} - u_t^{(i)}) \mathbf{v}_{t-1}^{(i)}.
\end{aligned}$$

- Salience decreases by exponential decay:

$$s_t^{(i)} = \lambda (1 - o_t^{(i)} - u_t^{(i)}) s_{t-1}^{(i)} + o_t^{(i)} + u_t^{(i)}.$$

Recurrent state

The memory is linked with a gated recurrent unit language model:[22]

$$\mathbf{h}_t = \mathrm{GRU}(\mathbf{x}_t, (1 - c_t)\mathbf{h}_{t-1} + c_t\mathbf{m}_t), \tag{1}$$

where $\mathbf{m}_t$ is a salience-weighted representation of the memory values,

$$\mathbf{m}_t = \sum_{i=1}^{N} s_t^{(i)} \mathbf{v}_t^{(i)}. \tag{2}$$

This architecture links coreference resolution to language modeling, facilitating multi-task training.

[22] Chung et al. 2015.

Coreference chains and supervision

The probability of coreference can be computed from the update and overwrite gates.

$$\psi_{t_1,t_2} = \sum_{i=1}^{N} \underbrace{\left(u_{t_1}^{(i)} + o_{t_1}^{(i)}\right)}_{t_1 \text{ in memory } i} \times \underbrace{u_{t_2}^{(i)}}_{t_2 \text{ updates } i} \times \underbrace{\prod_{\tau=t_1+1}^{t_2} \left(1 - o_{\tau}^{(i)}\right)}_{i \text{ is not overwritten}}.$$

Coreference chains and supervision

The probability of coreference can be computed from the update and overwrite gates.

$$\psi_{t_1,t_2} = \sum_{i=1}^{N} \underbrace{\left(u_{t_1}^{(i)} + o_{t_1}^{(i)}\right)}_{t_1 \text{ in memory } i} \times \underbrace{u_{t_2}^{(i)}}_{t_2 \text{ updates } i} \times \underbrace{\prod_{\tau=t_1+1}^{t_2} \left(1 - o_{\tau}^{(i)}\right)}_{i \text{ is not overwritten}}.$$

The model is trained on the cross-entropy,

$$\min \sum_{t_1, t_2 > t_1} -y_{t_1,t_2} \log \psi(t_1, t_2) - (1 - y_{t_1,t_2}) \log(1 - \psi(t_1, t_2)).$$

In addition, we train on a language modeling objective on large-scale unlabeled text.

Results

	F_1^M	F_1^F	$\frac{F_1^F}{F_1^M}$	F_1
Lee et al 2017, **pre**-trained	67.7	60.0	0.89	64.0
Lee et al 2017, **re**-trained	67.6	65.9	0.98	66.8
Parallelism	69.4	64.4	0.93	66.9

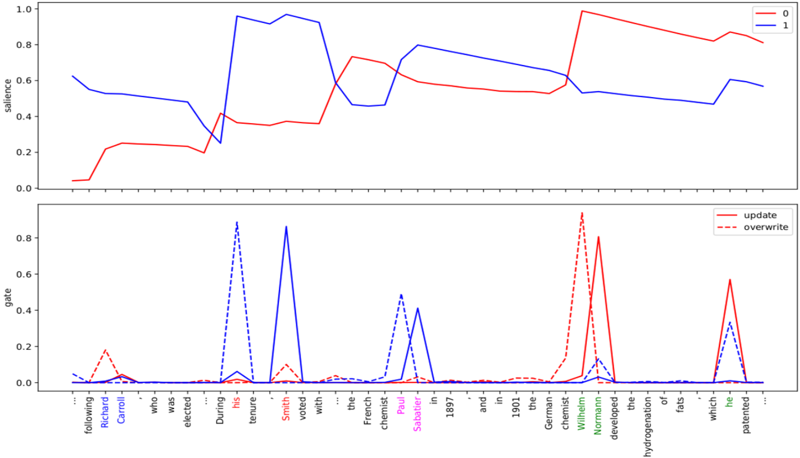
Results

	F_1^M	F_1^F	$\frac{F_1^F}{F_1^M}$	F_1
Lee et al 2017, **pre**-trained	67.7	60.0	0.89	64.0
Lee et al 2017, **re**-trained	67.6	65.9	0.98	66.8
Parallelism	69.4	64.4	0.93	66.9
Parallelism+URL	72.3	68.8	0.95	70.6

Results

	F_1^M	F_1^F	$\frac{F_1^F}{F_1^M}$	F_1
Lee et al 2017, **pre**-trained	67.7	60.0	0.89	64.0
Lee et al 2017, **re**-trained	67.6	65.9	0.98	66.8
Parallelism	69.4	64.4	0.93	66.9
Parallelism+URL	72.3	68.8	0.95	70.6
RefReader, LM	61.6	60.5	0.98	61.1
RefReader, coref	69.6	68.1	0.98	68.9
RefReader, LM & coref	**72.8**	**71.4**	**0.98**	**72.1**

- Referential Reader ($N = 2$) beats pre-trained systems and heuristics.
- Multitasking improves performance by $+3\, F_1$.



This example concatenates two GAP instances, but the reader "learns to forget" entities from the first instance so that it can correctly process the second instance.

What do we know about reference?

Reference resolution draws on a wide range of linguistic cues:

- morphosyntactic constraints;
- discourse structure;
- pragmatics;
- semantics;
- world knowledge.

What do we know about reference?

Reference resolution draws on a wide range of linguistic cues:

- ✓ morphosyntactic constraints;
- ✓ discourse structure;
- pragmatics;
- ✓ semantics;
- world knowledge.

Steps towards deeper linguistic reasoning:

- Explicit models of discourse relations and structure
- Multi-hop reasoning for pragmatic implicature?

Some speculative next steps

Human reading is not strictly left-to-right.

- Could memory network be linked to non-sequential generation?[23]
- What would it mean to do search in a model like this?

[23] Gu, Liu, and Cho 2019; Welleck et al. 2019.

[24] Fan, Lewis, and Dauphin 2019.

Some speculative next steps

Human reading is not strictly left-to-right.

- Could memory network be linked to non-sequential generation?[23]
- What would it mean to do search in a model like this?

"Back translation" for coreference?

- Generate hierarchically by first choosing entities, then generating mention strings[24]
- Use generated text as labeled examples to learn more about coreference.

[23] Gu, Liu, and Cho 2019; Welleck et al. 2019.

[24] Fan, Lewis, and Dauphin 2019.

Two hard problems for NLP

- **Reference resolution**: understanding which parts of a text refer to the same underlying semantic entity.
- **Style**: producing text that is stylistically coherent and controlled.

Two hard problems for NLP

- **Reference resolution**: understanding which parts of a text refer to the same underlying semantic entity.
- **Style**: producing text that is stylistically coherent and controlled.
  - What do the control knobs look like?
    How do we know if we have succeeded?
  - What are the perceptually and socially relevant dimensions of linguistic style?

Interactional Stancetaking in Online Forums[25]

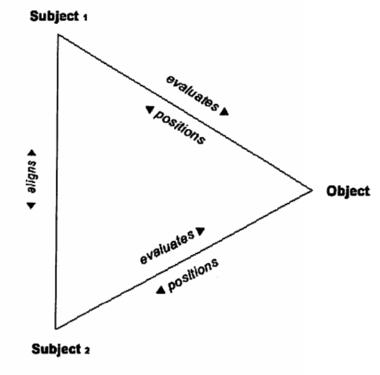
Scott Kiesling[26], Umashanthi Pavalanathan[27], Jim Fitzpatrick[26], Xiaochuang Han[27], Jacob Eisenstein[27]

[25] Scott F. Kiesling et al. (2018). "Interactional Stancetaking in Online Forums". In: Computational Linguistics 44 (4).

[26] University of Pittsburgh

[27] Georgia Tech

Interactional stancetaking[28]



- **Affect**: attitude towards stance object
- **Alignment**: attitude towards audience
- **Investment**: attitude towards talk

Stancetaking can be related to a wide range of "folk linguistic" constructs, like formality, agreeableness, and aggression.

---

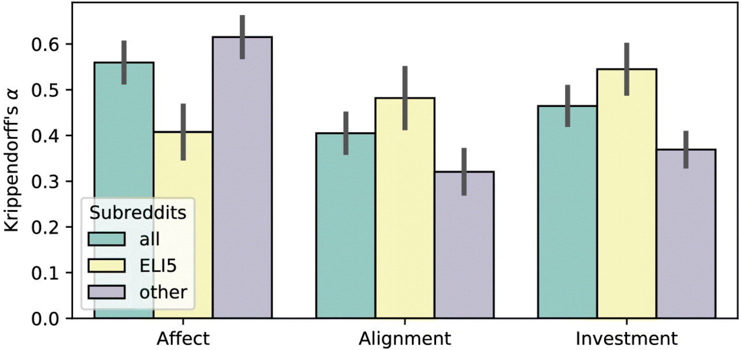
[28] Du Bois 2007.

A stancetaking corpus

Pilot corpus: annotations of the three stance dimensions on a 1-5 scale on 70 Reddit threads ($\sim$ 1400 turns)

ID-rep	Content	Stance focus	User	Karma	Affect	Investment	Alignment
008-02	People that are truly bad servers don't last very long at restaurants anyway. It is one of the jobs you actually need to be good at, you cant fake it like an engineer.	faking it like engineers	A_4	-2	3	3	3
009	You can't fake engineering	faking engineering	A_5	2	3	4	1
010	Lol... oh yes you can. You have to get the degree but that doesn't mean you are any use to anyone at your job.	faking engineering	A_4	2	3	3	1

Interrater agreement



- Annotation was largely performed by University of Pittsburgh undergraduates.
- Content matters: agreement varied widely by subreddit, and some were impossible to annotate reliably.

Linguistic properties

Affect		Investment		Alignment	
HIGH	LOW	HIGH	LOW	HIGH	LOW
thank	please	!	little	thank	evidence
!	worse	tell	limit	limit	wrong
sing	everyone	hope	ink	other	able
noise	nothing	better	maybe	!	not
stop	entire	never	may	absolutely	opinion

Linguistic properties

Affect		Investment		Alignment	
HIGH	LOW	HIGH	LOW	HIGH	LOW
thank	please	!	little	thank	evidence
!	worse	tell	limit	limit	wrong
sing	everyone	hope	ink	other	able
noise	nothing	better	maybe	!	not
stop	entire	never	may	absolutely	opinion

Dialogue properties:

- Alignment is "sticky": low-alignment turns tend to follow each other.
- Investment is "anti-sticky": high-investment turns to follow low-investment turns.

A stancetaking dialogue system?

(At least) two possible settings:

- Generate turns in a dialogue conditioned on the desired stancetaking attributes,[29] which could be controlled by the agent's personality.
- Generate turns conditioned on the stancetaking attributes of the **previous** turn, thereby modeling the contextual appropriateness of the response.

How much more annotated data do we need? How can we leverage unlabeled data?[30]

[29] Huang et al. 2018; Rashkin et al. 2018.

[30] Wolf et al. 2019.

Stance and stylistic variation in translation[31]

A: This sat work??

B: I kinda had plans already. Wassup with a weekday

A: I gotta beachouse 4th of July if y'all wanna come up

B: Ehhh I have off the 27th so maybe the 26th

[31] Song et al. 2018.

Stance and stylistic variation in translation[31]

A: This sat work??	Ce travail assis??	This job sitting??
B: I kinda had plans already. Wassup with a weekday	J'ai un peu deja des plans. Wassup avec un jour de semaine	I already have some plans. Wassup with a weekday
A: I gotta beachouse 4th of July if y'all wanna come up	Je dois beachouse le 4 juillet si vous voulez monter	I have beachouse on July 4th if you want to ride
B: Ehhh I have off the 27th so maybe the 26th	Ehhh j'ai le 27 alors peut-être le 26	Ehhh I have the 27th so maybe the 26th

[31] Song et al. 2018.

Giving MT a social life

Very limited public data for measuring translation of SMS-style text.[32]

- This summer, Facebook is leading a team at the Jelinek summer workshop to address this problem!

[32] Michel and Neubig 2018.
[33] Belinkov and Bisk 2017.
[34] Karpukhin et al. 2019.

Giving MT a social life

Very limited public data for measuring translation of SMS-style text.[32]

- This summer, Facebook is leading a team at the Jelinek summer workshop to address this problem!

How to handle character-level variation in MT?[33] (e.g., ehhh, wassup)

- Pinter, Guthrie, and Eisenstein 2017: "mimicking" word embeddings with sub-word RNNs
- This ACL: making character-level encoders more robust to natural noise, such as spelling errors.[34]

---

[32] Michel and Neubig 2018.
[33] Belinkov and Bisk 2017.
[34] Karpukhin et al. 2019.

Summary

It's an exciting time to work on NLP!

- New deep learning methods are making rapid progress on hard problems.
- Historically, each new wave of methodological innovation (e.g., probability, discriminative learning) has required rethinking the relationship between linguistics, computing, and data.
- The implications of deep learning for this relationship are still being worked out. NLP researchers should seek new syntheses that make language technology more flexible, robust, and data-efficient.

Acknowledgments

Co-authors Fei Liu, Marjan Ghazvininejad, Xiaochuang Han, Vladimir Karpukhin, Scott F. Kiesling, Omer Levy, Umashanthi Pavalanathan, Luke Zettlemoyer.

Discussion Y-Lan Boureau, Lucio Dery, Mandar Joshi, Michael Lewis, Xian Li, Yinhan Liu, Abdelrahman Mohamed, Sean Vasquez.

References I

- Belinkov, Yonatan and Yonatan Bisk (2017). "Synthetic and natural noise both break neural machine translation". In: arXiv preprint arXiv:1711.02173.
- Chinchor, Nancy and Patricia Robinson (1997). "MUC-7 named entity task definition". In: Proceedings of the 7th Conference on Message Understanding. Vol. 29.
- Chung, Junyoung et al. (2015). "Gated feedback recurrent neural networks". In: Proceedings of the International Conference on Machine Learning (ICML).
- Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: arXiv preprint arXiv:1810.04805.
- Du Bois, John W. (2007). "The stance triangle". In: Stancetaking in discourse. Ed. by Robert Engelbretson. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 139–182.

References II

- Durrett, Greg and Dan Klein (2013). "Easy victories and uphill battles in coreference resolution". In: Proceedings of Empirical Methods for Natural Language Processing (EMNLP
- Fan, Angela, Mike Lewis, and Yann Dauphin (2019). "Strategies for Structuring Story Generation". In: arXiv preprint arXiv:1902.01109.
- Grice, H Paul (1975). "Logic and Conversation". In: Syntax and Semantics Volume 3: Speech Acts. Ed. by P. Cole and J. L. Morgan. Academic Press, pp. 41–58.
- Grosz, Barbara J, Scott Weinstein, and Aravind K Joshi (1995). "Centering: A framework for modeling the local coherence of discourse". In: Computational linguistics 21.2, pp. 203–225.
- Gu, Jiatao, Qi Liu, and Kyunghyun Cho (2019). "Insertion-based Decoding with Automatically Inferred Generation Order". In: arXiv preprint arXiv:1902.01370.
- Haghighi, Aria and Dan Klein (2010). "Coreference resolution in a modular, entity-centered model". In: Proceedings of the North American Chapter of the Association for Computati pp. 385–393.

References III

Henaff, Mikael et al. (2017). “Tracking the World State with Recurrent Entity Networks”. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France.

Huang, Chenyang et al. (June 2018). “Automatic Dialogue Generation with Expressed Emotions”. In: Proceedings of the 2018 Conference of the North American Chapter of the As New Orleans, Louisiana: Association for Computational Linguistics, pp. 49–54.

Karpukhin, Vladimir et al. (2019). “Training on Synthetic Noise Improves Robustness to Natural Noise in Machine Translation”. In: arXiv preprint arXiv:1902.01509.

Kehler, Andrew and Hannah Rohde (2013). “A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation”. In: Theoretical Linguistics 39.1-2, pp. 1–37.

Kiesling, Scott F. et al. (2018). “Interactional Stancetaking in Online Forums”. In: Computational Linguistics 44 (4).

References IV

Lee, Heeyoung et al. (2013). "Deterministic coreference resolution based on entity-centric, precision-ranked rules". In: Computational Linguistics 39.4, pp. 885–916.

Lee, Kenton, Luheng He, Mike Lewis, et al. (2017a). "End-to-end Neural Coreference Resolution". In: Proceedings of Empirical Methods for Natural Language Processing (EMNLP

– (2017b). "End-to-end Neural Coreference Resolution". In: Proceedings of the 2017 Conference on Empirical Methods in Natural Langua Copenhagen, Denmark: Association for Computational Linguistics, pp. 188–197.

Lee, Kenton, Luheng He, and Luke Zettlemoyer (2018). "Higher-Order Coreference Resolution with Coarse-to-Fine Inference". In: Proceedings of the North American Chapter of the Association for Computati pp. 687–692.

Michel, Paul and Graham Neubig (2018). "MTNT: A Testbed for Machine Translation of Noisy Text". In: Proceedings of Empirical Methods for Natural Language Processing (EMNLP pp. 543–553.

References V

Miller, Alexander et al. (2016). "Key-value memory networks for directly reading documents". In: arXiv preprint arXiv:1606.03126.

Moosavi, Nafise Sadat and Michael Strube (2017). "Lexical features in coreference resolution: To be used with caution". In: arXiv preprint arXiv:1704.06779.

Peters, Matthew E et al. (2018). "Deep contextualized word representations". In: Proceedings of the North American Chapter of the Association for Computati

Pinter, Yuval, Robert Guthrie, and Jacob Eisenstein (2017). "Mimicking Word Embeddings using Subword RNNs". In: Proceedings of Empirical Methods for Natural Language Processing (EMNLP

Pradhan, Sameer et al. (2011). "CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes". In: Proceedings of the Conference on Natural Language Learning (CoNLL), pp. 1–27.

Rashkin, Hannah et al. (2018). "I know the feeling: Learning to converse with empathy". In: arXiv preprint arXiv:1811.00207.

References VI

- Song, Zhiyi et al. (2018). BOLT English SMS/chat LDC2018T19. Philadelphia.
- Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim (2001). "A machine learning approach to coreference resolution of noun phrases". In: Computational linguistics 27.4, pp. 521–544.
- Stoyanov, Veselin et al. (2009). "Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art". In: Proceedings of the Association for Computational Linguistics (ACL), pp. 656–664.
- Webster, Kellie et al. (2018). "Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns". In: Transactions of the ACL, to appear.
- Welleck, Sean et al. (2019). "Non-Monotonic Sequential Text Generation". In: arXiv preprint arXiv:1902.02192.
- Weston, Jason, Sumit Chopra, and Antoine Bordes (2015). "Memory networks". In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA.

References VII

Wiseman, Sam Joshua et al. (2015). “Learning anaphoricity and antecedent ranking features for coreference resolution”. In: Proceedings of the Association for Computational Linguistics (ACL).

Wiseman, Sam, Alexander M. Rush, and Stuart M. Shieber (2016). “Learning Global Features for Coreference Resolution”. In: Proceedings of the 2016 Conference of the North American Chapter of the As San Diego, California: Association for Computational Linguistics, pp. 994–1004.

Wolf, Thomas et al. (2019). “TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents”. In: arXiv preprint arXiv:1901.08149.