

Bias

Warning: examples can feature triggering content.
They do not reflect the opinions of the speaker or
paper authors

Announcements

- Reading: Paper on bias
- Laptop policy: in effect for today to encourage discussion; you may bring your laptop after Thanksgiving
- Monday, Dec 2nd: Information extraction
- Wednesday, Dec 4th: Analysis of gang-involved social media posts and Final exam review
- Monday, Dec 9th: In-class final exam

Annotators needed

- Fact checking to reduce electricity consumption
- Tips to change electricity consumption mined from the internet: which are valid?
- \$15/hour for 4 hours of work
- Send email to hidey@cs.columbia.edu if interested

Today

- Attention a closer look
- Detecting bias in word and sentence embeddings
 - Semantics derived automatically from language corpora contain human biases
 - On measuring social biases in sentence embeddings
- Do de-biasing techniques actually work?
 - Lipstick on a Pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them

Attention

Aligning and Translating

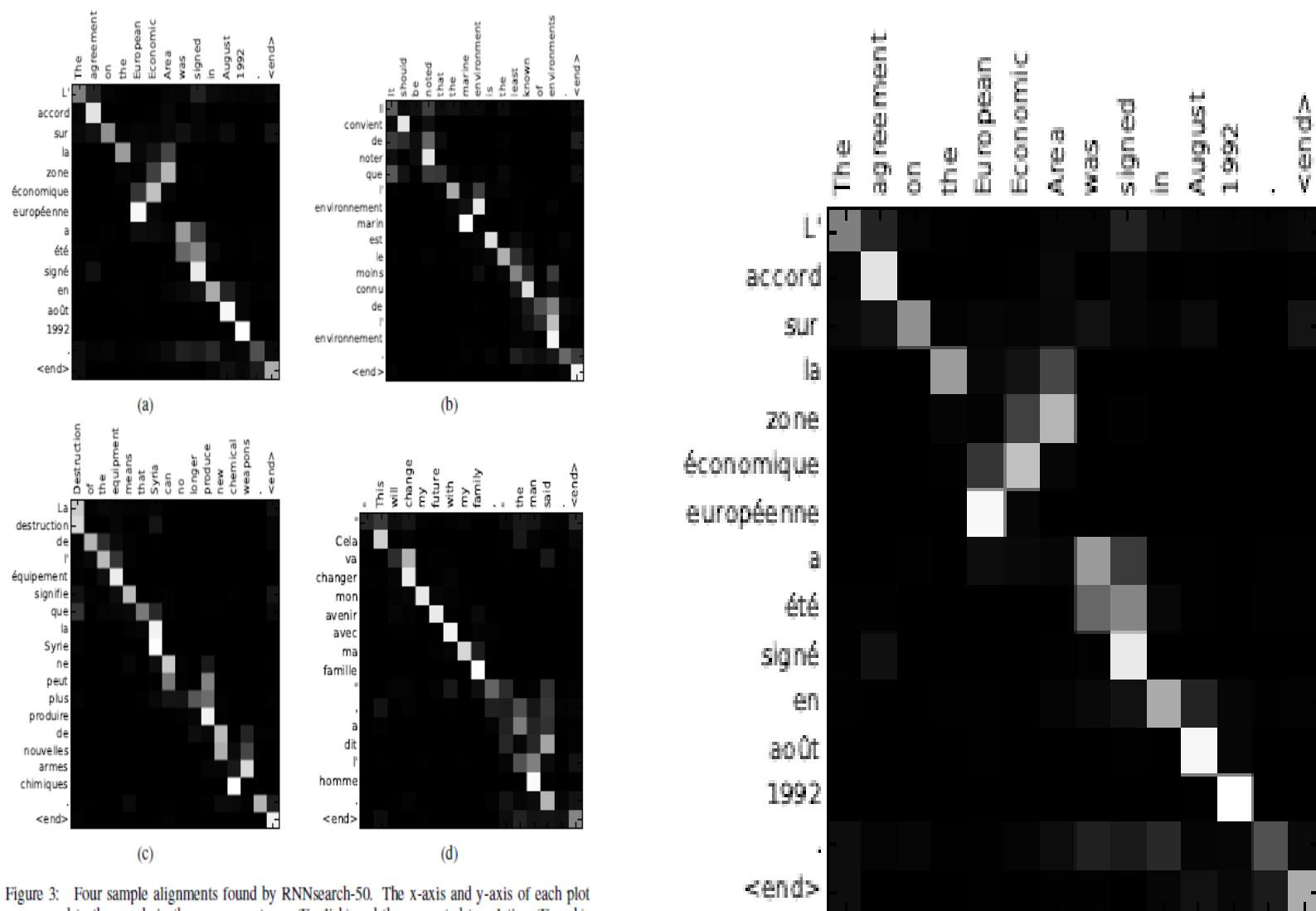
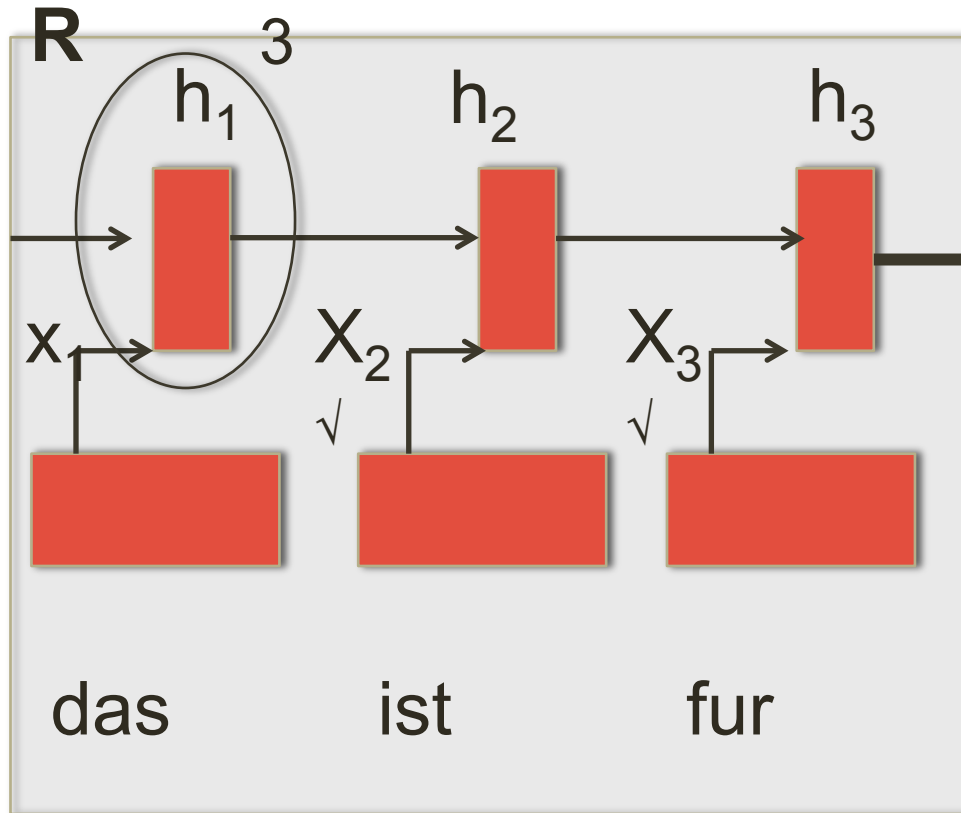


Figure 3: Four sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight α_{ij} of the annotation of the j -th source word for the i -th target word (see Eq. (6)), in grayscale (0: black, 1: white). (a) an arbitrary sentence. (b-d) three randomly selected samples among the sentences without any unknown words and of length between 10 and 20 words from the test set.

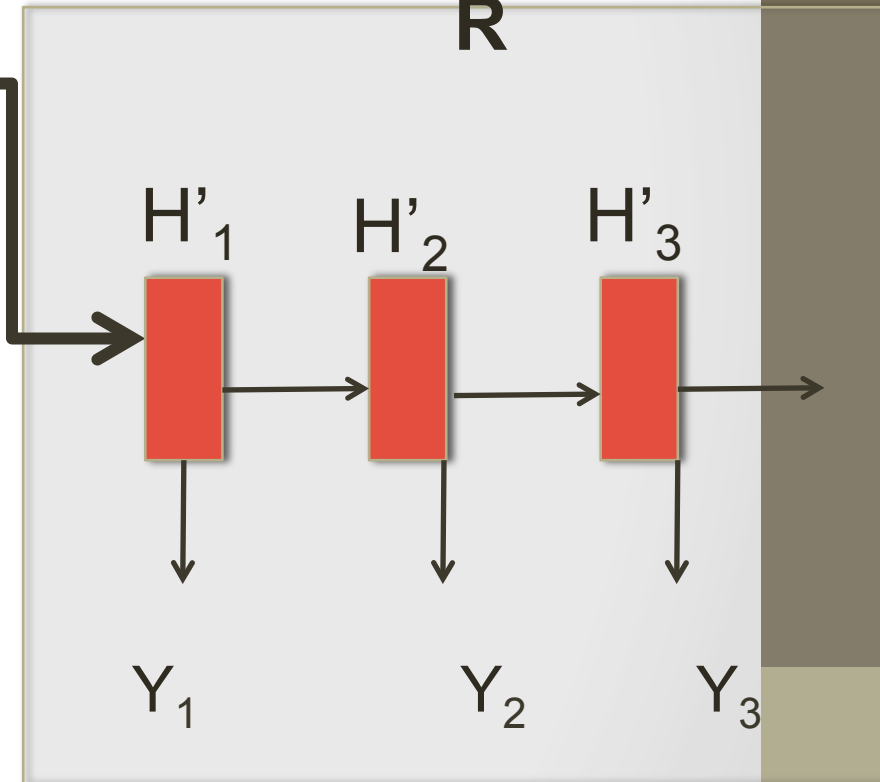
Attention Mechanism - Scoring

ENCODE



Score (h'_{t-1}, h_s)

DECODE
R

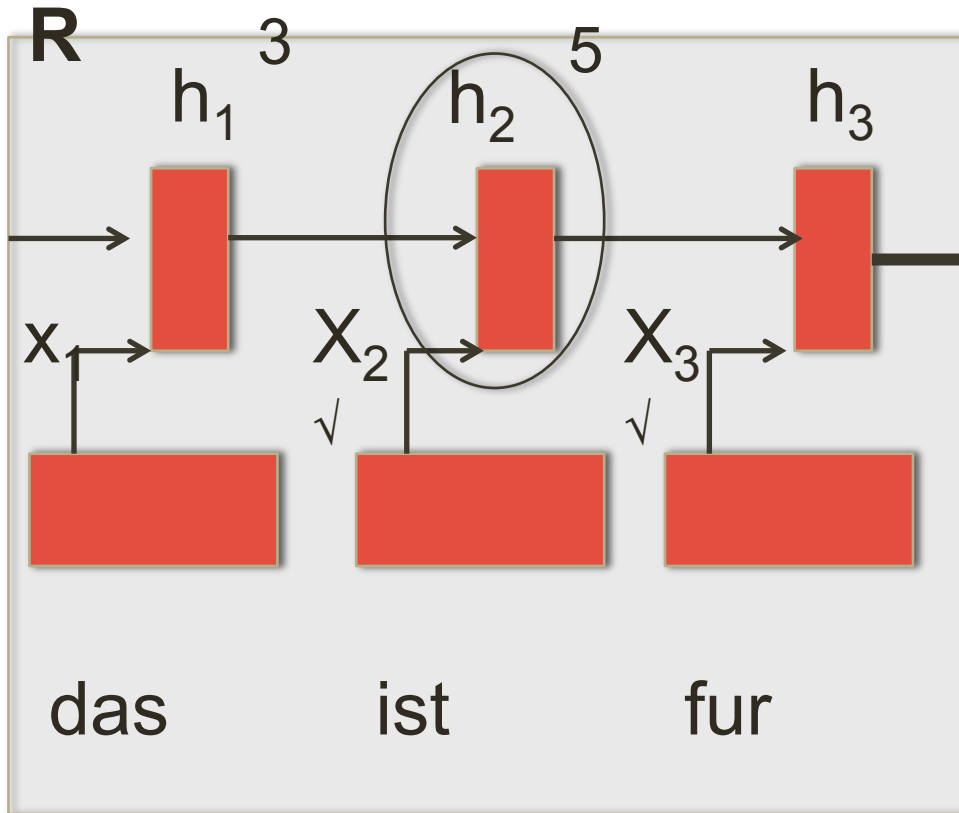


That

?

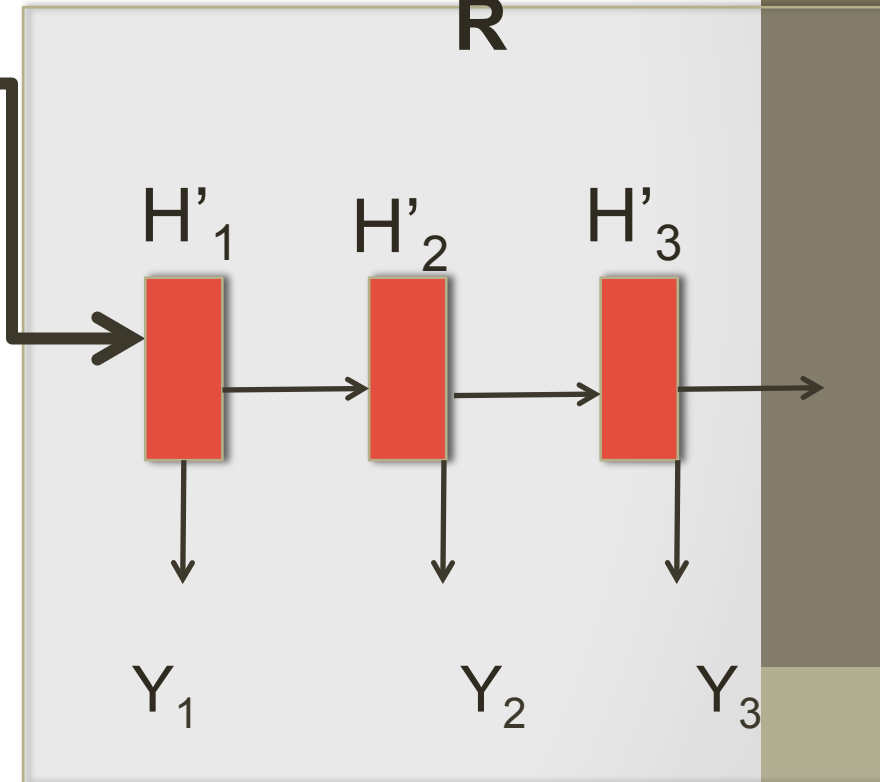
Attention Mechanism - Scoring

ENCODE



Score (h'_{t-1}, h_s)

DECODE

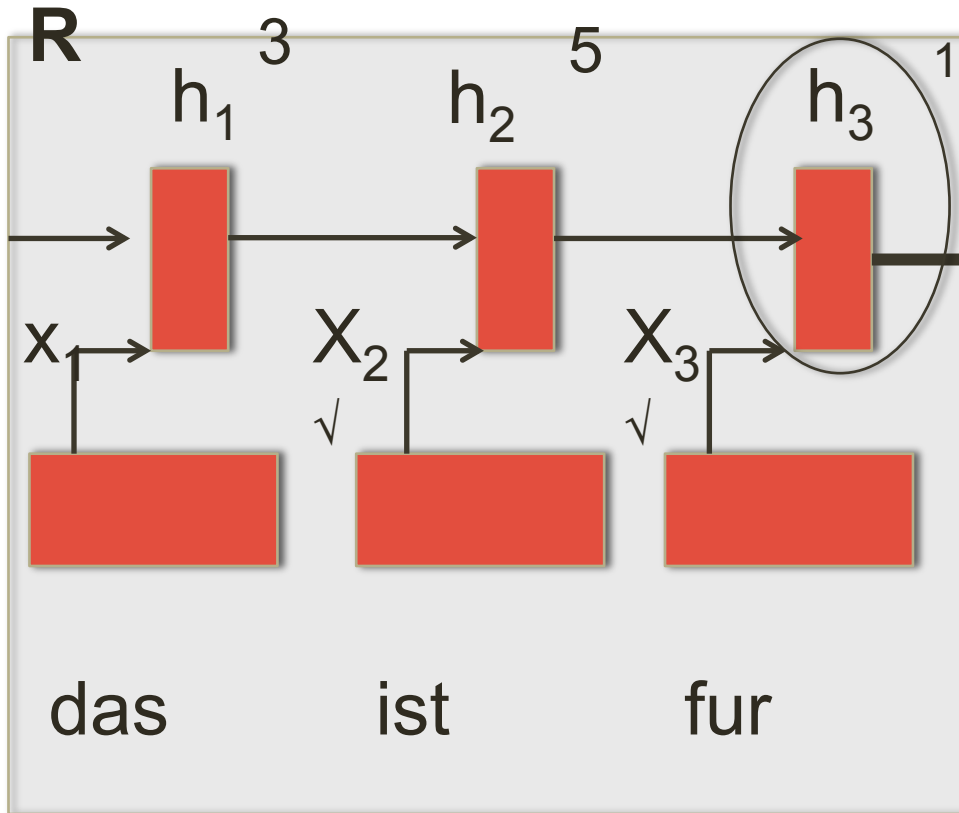


That

?

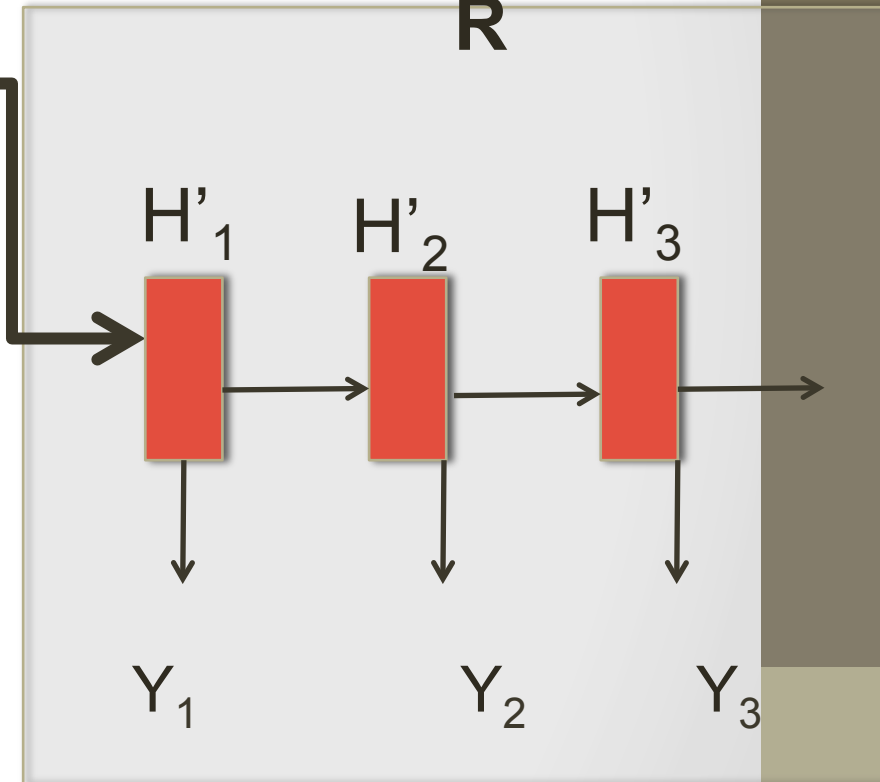
Attention Mechanism - Scoring

ENCODE



Score (h'_{t-1}, h_s)

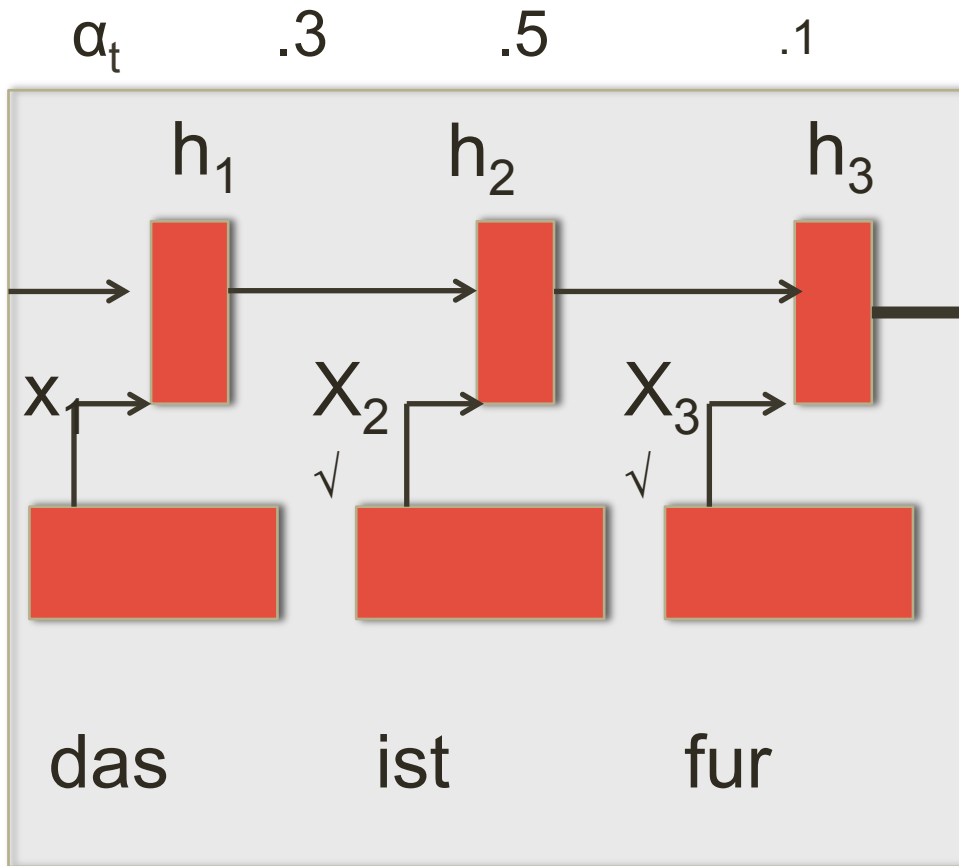
DECODE



That

?

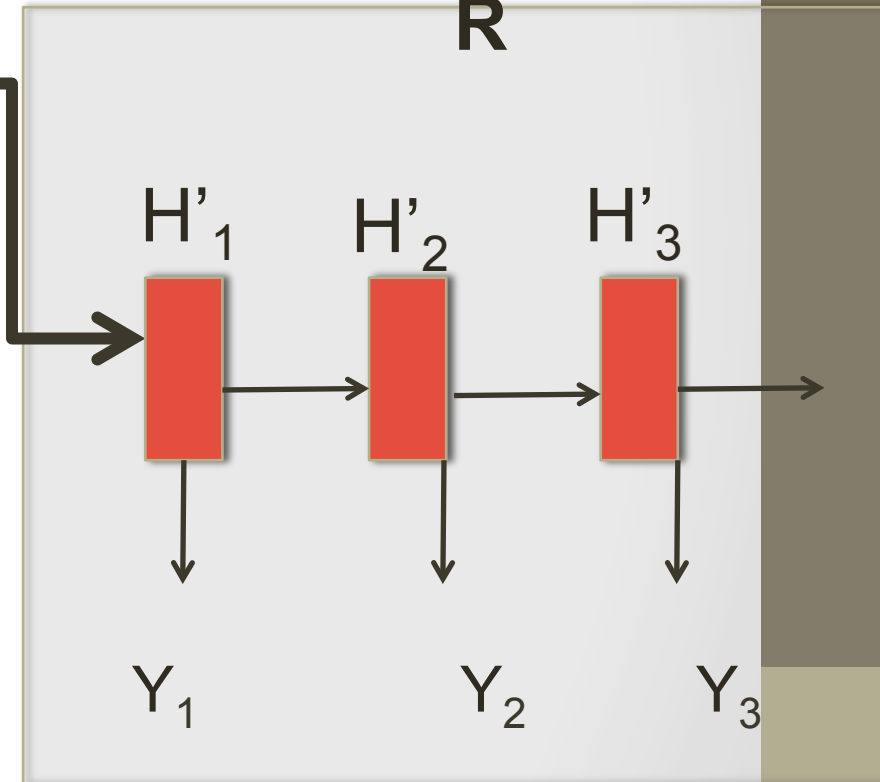
Attention Mechanism - Scoring



Convert into alignment weights

$$a_t(s) = \frac{e^{\text{score}(s)}}{\sum_{s'} e^{\text{score}(s')}}$$

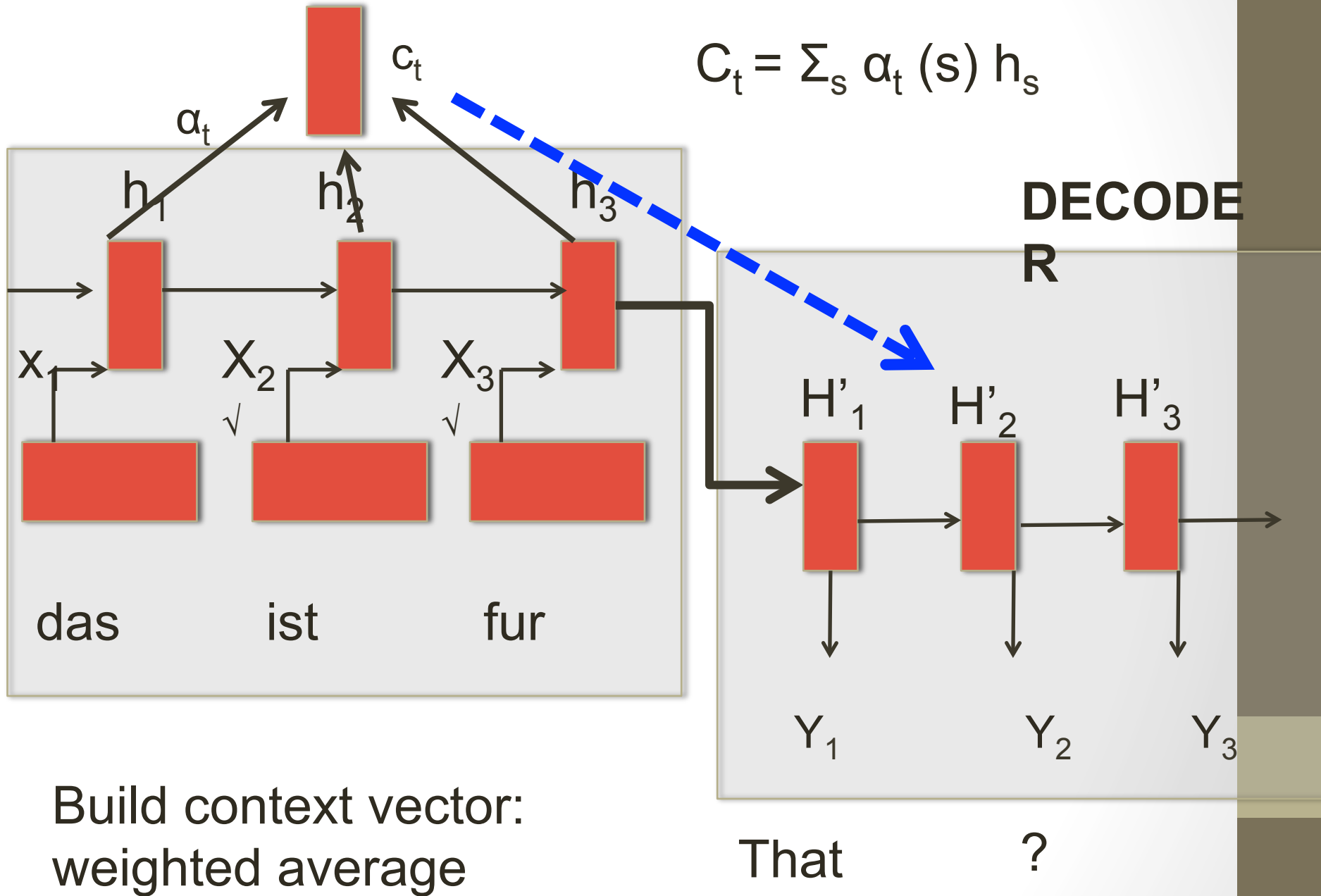
DECODE
R



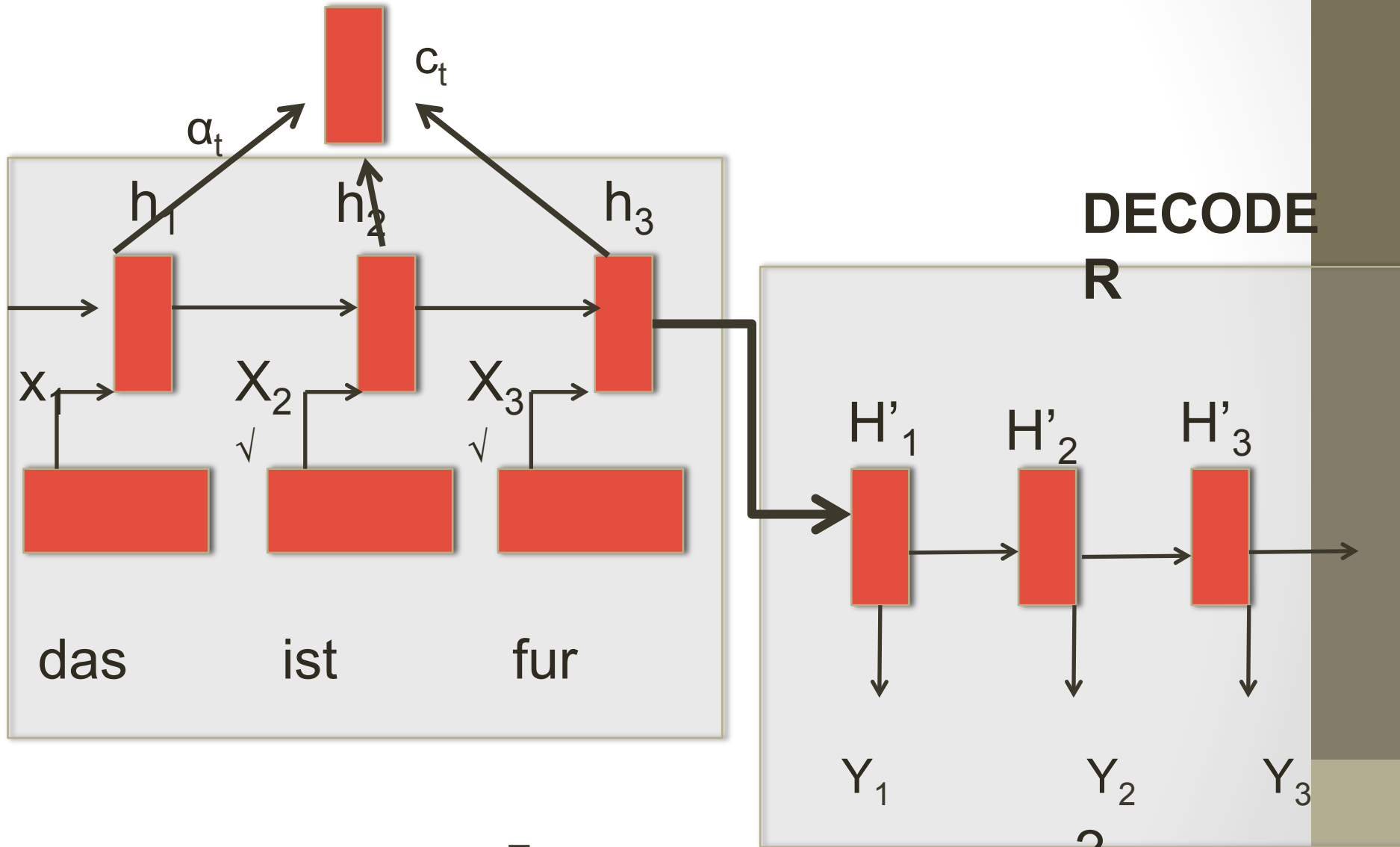
That

?

Attention Mechanism - Scoring



How do you score it?



$$\text{Score}(h_s, H'_t) = H'_t{}^T h_s$$

or

$$= H'_t{}^T W_\alpha h_s \text{ (Luong et al 2015)}$$

Bias in word embeddings

- Word embedding representations encode semantic analogies
- They also encode bias
 - Morally neutral (flowers vs insects)
 - Problematic (race, gender)
 - Reflecting status quo (e.g., in career)
- Analogies test: Word2Vec: “Man is to computer programmer as woman is to homemaker.”
- Measure bias using the Implicit Association Test
- Corpus linguistics has noted bias since 1996 but word embeddings amplify

Implicit Association Test

- Measures latency in reaction time to a presented pair of words
 - Flowers – pleasant
 - Flowers – unpleasant
 - Insects – pleasant
 - Insects – unpleasant
- Other word pairs
 - Instrument, weapons – pleasant, unpleasant
 - European American names, African American names – pleasant, unpleasant
 - Female names, male names – family, career
 - Female words (woman, girl), male words – arts, math

Latency and embedding comparisons

- Cohen's $d = (\text{mean log transformed latencies in milliseconds (MLTL) word 1 pair} - \text{MLTL word 2 pair}) / \text{standard deviation}$
 - .2 (small), .5 (medium), .8 (large)
- How would we compare two word embeddings?

Word Embedding Association Test

- Two sets of target words
 - Programmer, engineer, scientist
 - Nurse, teacher, librarian
- Two sets of attribute words
 - Man, male vs woman, female
- Null hypothesis: no difference in semantic similarity between target sets and attribute sets

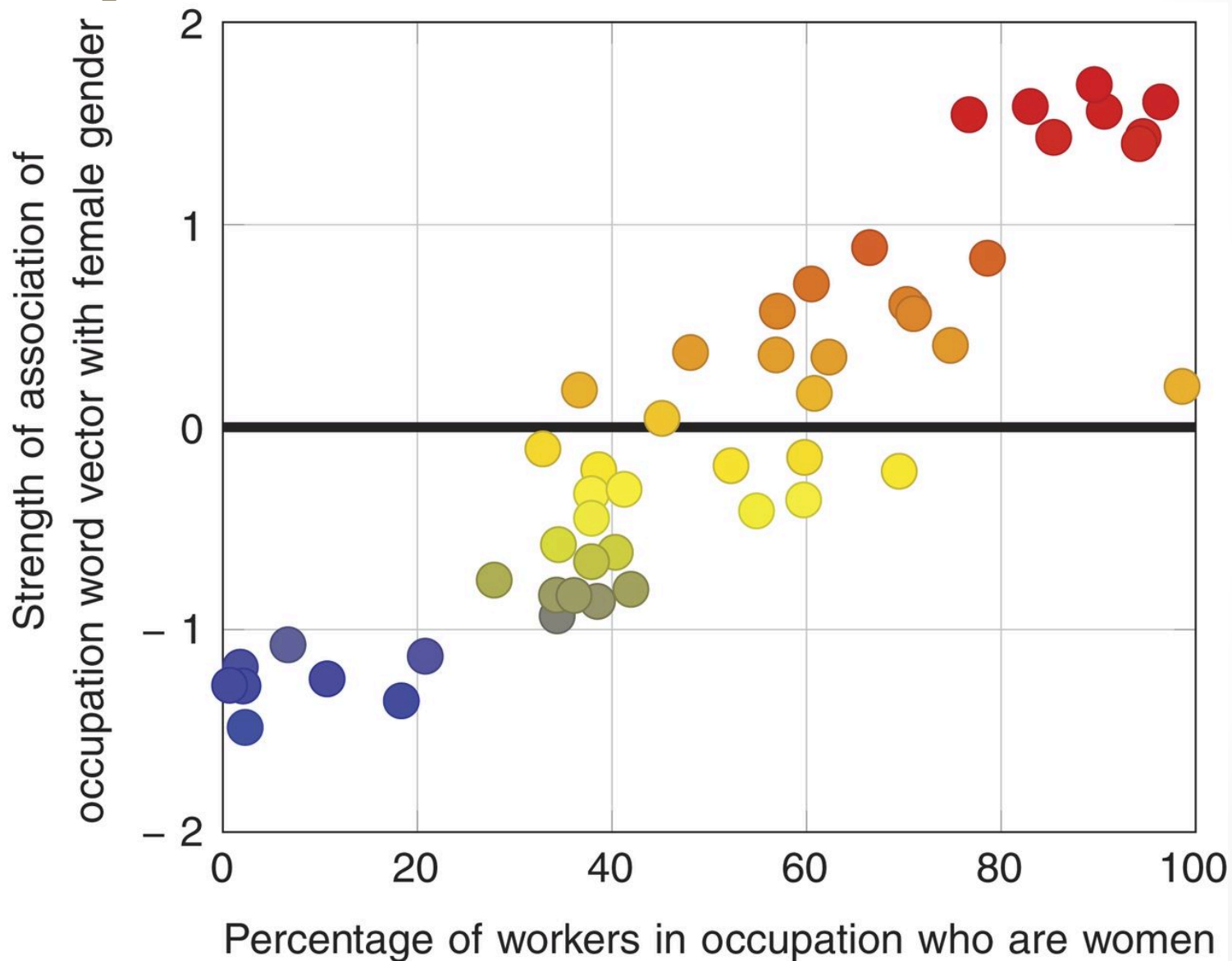
WEAT test statistic

- X, Y: target sets, A, B: attribute sets
- Differential association of target set/attribute set
- $S(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B)$
- $S(w,A,B) = \text{mean}_{a \in A} \cos(w,a) - \text{mean}_{b \in B} \cos(w,b)$
- Use the permutation test
- Normalized measure of how separated the distribution of X and A,B are vs Y and A,B
 - For all x in X and for all y in Y:
 - $\text{Mean } s(x,A,B) - \text{mean } s(y,A,B) / \text{SD}(w \in (X \cup Y) s(w,A,B))$

Results

- Flowers -> pleasant, insects -> unpleasant
- European American names -> pleasant
African American names -> unpleasant
- Female names -> family words
Male names -> career words
- Female words (woman, girl) -> arts words
Male words -> math words

Correlation between gender association of occupation word and labor force



Implications

- Results suggest that behavior can be driven by cultural history embedded in a term's use
- Histories can vary by language
- Sapir-Whorf hypothesis: “Human beings do not live in the objective world alone, nor alone in the world of social activity as ordinarily understood, but are very much at the mercy of the particular language which has become the medium of expression in their society. It is quite an illusion to imagine that one adjusts to reality essentially without the use of language and that language is merely an incidental means of solving specific problems of communication or reflection: The fact of the matter is that the ‘real world’ is to a large extent unconsciously built up on the language habits of the group. No two languages are ever sufficiently similar to be considered as representing the same social reality.”

Do you agree that bias/term use varies by language? If so, what languages seem different?

What implications would this have for NLP tasks?

- Translation: Chinese to English?
 - S/he performed brain surgery
- Pronoun disambiguation/generation
 - The nurse talked to the doctor. He said.
 - Does this change as society changes?
- Downstream NLP tasks: dialog?

How might bias affect pronoun use/coreference?



How might bias affect a downstream task like dialog?

Sentence level embeddings

- Does the same bias exist for sentence embeddings?
- Does it change if we use different encoding methods?
- SEAT: Sentence Encoder Association Test
 - Apply to simple sentence templates where the word has been inserted: “This is a <word>”.
 - Uses WEAT

Examples

- European American: “This is Katie.” “This is Adam.” “Paul is there.”
- African American names: “This is Jamal.” “That is Latisha.” “Lavon is there.”
- Pleasant: “There is love.” “That is happy.” “this is a friend.”
- Unpleasant: “This is evil.” “They are evil.” “That can kill.”



Why might we expect SEAT to be different from WEAT?

Two additional biases

- The “angry black woman” stereotype (Collins 2004, Madison 2009, Harris-Perry 2011, Hooks 2015, Gillespie 2016)
- A “double bind” on women in professional settings (Heilman et al 2004)

Testing

- The double bind
 - Targets: male/female names. “Kathy is an engineer with superior technical skills”
 - Attributes: likable and non-hostile terms: “the engineer is nice”
 - Target: “Kathy is an engineer”
 - Attributes: competent/achievement-oriented terms: “The engineer is high performing.”
- ABW
 - Same as example

Test	Context	CBoW	InferSent	GenSen	USE	ELMo	GPT	BERT
C1: Flowers/Insects	word	1.50**	1.56**	1.24**	1.38**	-0.03	0.20	0.22
C1: Flowers/Insects	sent	1.56**	1.65**	1.22**	1.38**	0.42**	0.81**	0.62**
C3: EA/AA Names	word	1.41**	1.33**	1.32**	0.52	-0.40	0.60*	-0.11
C3: EA/AA Names	sent	0.52**	1.07**	0.97**	0.32*	-0.38	0.19	0.05
C6: M/F Names, Career	word	1.81*	1.78*	1.84*	0.02	-0.45	0.22	0.21
C6: M/F Names, Career	sent	1.74**	1.69**	1.63**	0.83**	-0.38	0.35	0.08
ABW Stereotype	word	1.10*	1.18*	1.57**	-0.39	0.53	0.08	-0.32
ABW Stereotype	sent	0.62**	0.98**	1.05**	-0.19	0.52*	-0.07	-0.17
Double Bind: Competent	word	1.62*	1.09	1.49*	1.51*	-0.35	-0.28	-0.81
Double Bind: Competent	sent	0.79**	0.57*	0.83**	0.25	-0.15	0.10	0.39
Double Bind: Competent	sent (u)	0.84	1.42*	1.03	0.71	0.20	0.71	1.17*
Double Bind: Likable	word	1.29*	0.65	1.31*	0.16	-0.60	0.91	-0.55
Double Bind: Likable	sent	0.69*	0.37	0.25	0.32	-0.45	-0.20	-0.35
Double Bind: Likable	sent (u)	0.51	1.33*	0.05	0.48	-0.90	-0.87	0.99

Table 4: SEAT effect sizes for select tests, including word-level (word), bleached sentence-level (sent), and unbleached sentence-level (sent (u)) versions. *CN*: test from Caliskan et al. (2017, Table 1) row *N*; *: significant at 0.01, **: significant at 0.01 after multiple testing correction.

Tests based on given name have more of an effect
 Stronger evidence for Caliskan and ABW than the double
 bind: women are associated with incompetence
 regardless of context!

Test	Context	CBoW	InferSent	GenSen	USE	ELMo	GPT	BERT
C1: Flowers/Insects	word	1.50**	1.56**	1.24**	1.38**	-0.03	0.20	0.22
C1: Flowers/Insects	sent	1.56**	1.65**	1.22**	1.38**	0.42**	0.81**	0.62**
C3: EA/AA Names	word	1.41**	1.33**	1.32**	0.52	-0.40	0.60*	-0.11
C3: EA/AA Names	sent	0.52**	1.07**	0.97**	0.32*	-0.38	0.19	0.05
C6: M/F Names, Career	word	1.81*	1.78*	1.84*	0.02	-0.45	0.22	0.21
C6: M/F Names, Career	sent	1.74**	1.69**	1.63**	0.83**	-0.38	0.35	0.08
ABW Stereotype	word	1.10*	1.18*	1.57**	-0.39	0.53	0.08	-0.32
ABW Stereotype	sent	0.62**	0.98**	1.05**	-0.19	0.52*	-0.07	-0.17
Double Bind: Competent	word	1.62*	1.09	1.49*	1.51*	-0.35	-0.28	-0.81
Double Bind: Competent	sent	0.79**	0.57*	0.83**	0.25	-0.15	0.10	0.39
Double Bind: Competent	sent (u)	0.84	1.42*	1.03	0.71	0.20	0.71	1.17*
Double Bind: Likable	word	1.29*	0.65	1.31*	0.16	-0.60	0.91	-0.55
Double Bind: Likable	sent	0.69*	0.37	0.25	0.32	-0.45	-0.20	-0.35
Double Bind: Likable	sent (u)	0.51	1.33*	0.05	0.48	-0.90	-0.87	0.99

Table 4: SEAT effect sizes for select tests, including word-level (word), bleached sentence-level (sent), and unbleached sentence-level (sent (u)) versions. *CN*: test from Caliskan et al. (2017, Table 1) row *N*; *: significant at 0.01, **: significant at 0.01 after multiple testing correction.

Discrepancies: math/art -> male, female and science/art -> male/female. CBoW: same p-values; BERT, GenSim, GPT do not agree.

Other problems

- Caliskan's tests 3,4,5:
 - European American/African American -> pleasant/unpleasant
 - Test 3 has larger attribute sets than Test4
 - Test 4 has larger target concept sets than Test 5
 - Expect increasing p-values across 3,4,5
 - Target concepts and attributes of larger size -> higher power tests
- Yes for CBOW on word and sentence versions
- No for ELMo (decreasing p-values on word and sentence versions)

Cautions

- Are Bert and ELMo less likely to encode bias?
 - SEAT can confirm that bias exists, but negative results do not indicate no bias
- Discrepancies in results: results may not generalize beyond the specific words and sentences in the data
- Cosine similarity may not be a suitable model of representational similarity in recent models (e.g., BERT)
- ABW merits further study as an intersectional bias
 - Not well anticipated by an additive model of racism and sexism

Reactions?



Are there aspects of this work that you question?

Debiasing Methods

- Bolukbasi et al 2016:
 - Define gender bias w by its projection on the “gender direction”: w°_{he} , w°_{she} (the larger the projection the more biased)
 - Use post-processing for de-biasing
 - Change the word vectors for all words not inherently gendered (e.g., king, queen)
 - Zero the gender projection for each word on a pre-defined gender direction
 - Gender projection = top principal component for 10 gender pair difference vectors
 - Takes dozens of inherently gendered words and ensure that neutral words equally distant

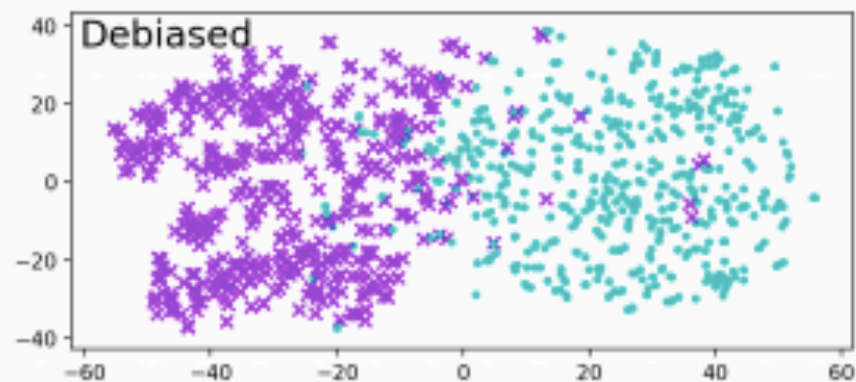
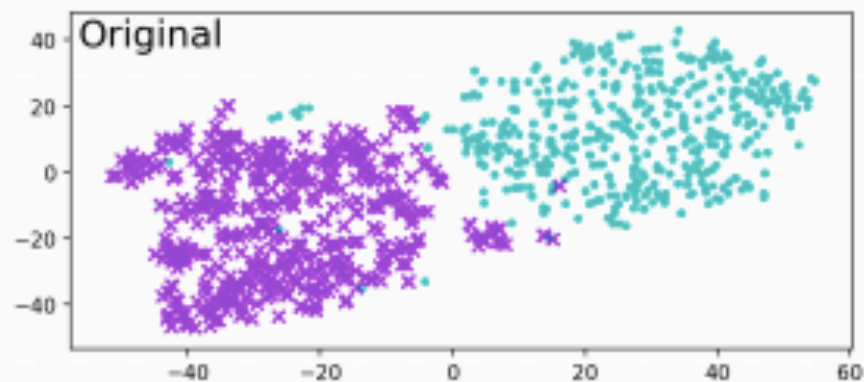
- Zhao et al 2018
 - Train debiased word embeddings from scratch
 - Change the loss function for Glove
 - To concentrate gender information in last coordinate
 - Two groups of male/female seed words
 - Encourage words in different groups to differ in last coordinate
 - Encourage neutral gender words to be orthogonal
 - When using the word embeddings, ignore the last coordinate

Do debiasing methods work?

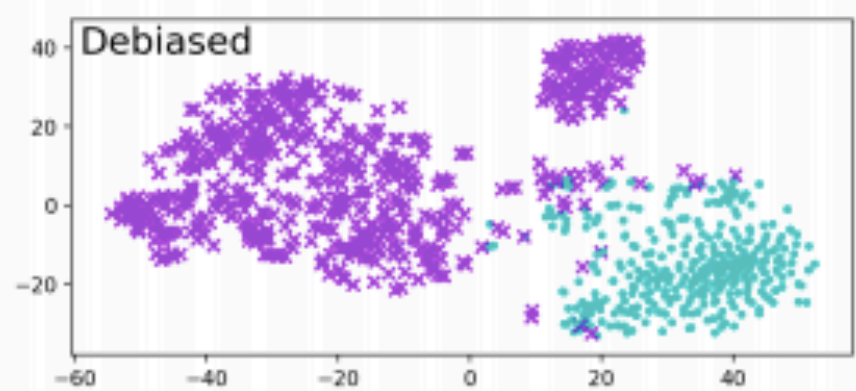
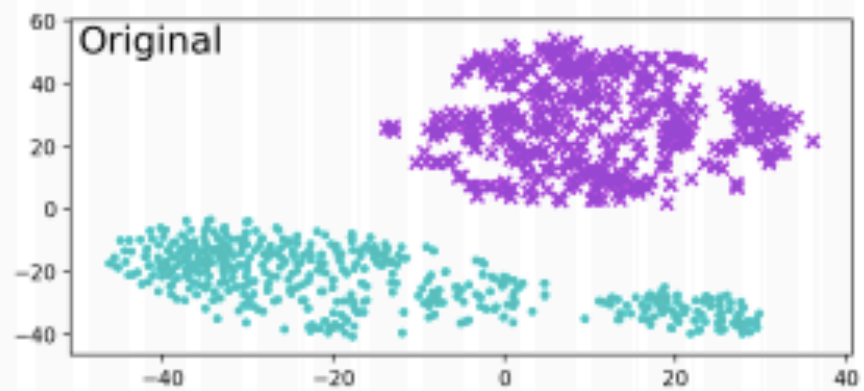
- Lipstick on a pig paper claims they do not
- Hides the bias
- Still reflected in similarities between gender neutral words
 - E.g., “math” “delicate”
- Most word pairs maintain their previous similarity

Experiments: do male and female-biased words cluster together?

- Take most biased words in the vocabulary according to the original bias
- 500 male biased, 500 female biased
- Cluster into 2 clusters using k-means



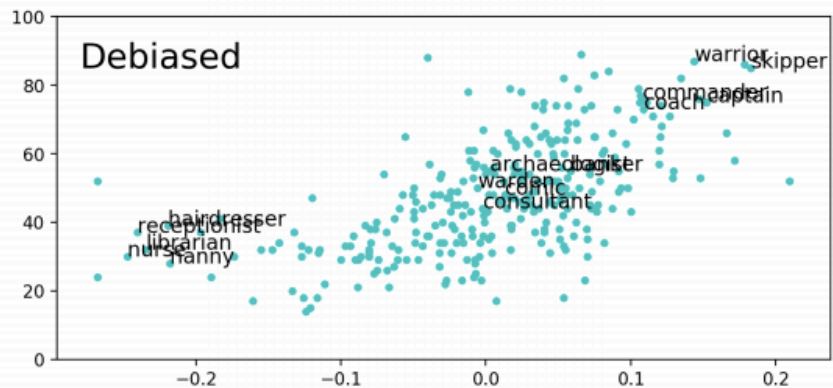
(a) Clustering for HARD-DEBIASED embedding, before (left hand-side) and after (right hand-side) debiasing.



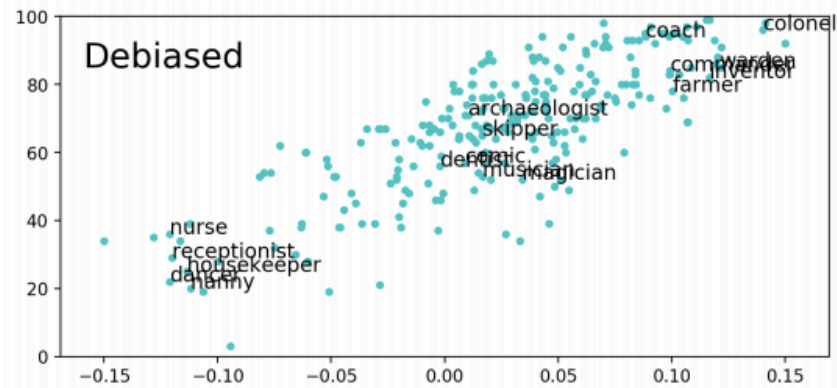
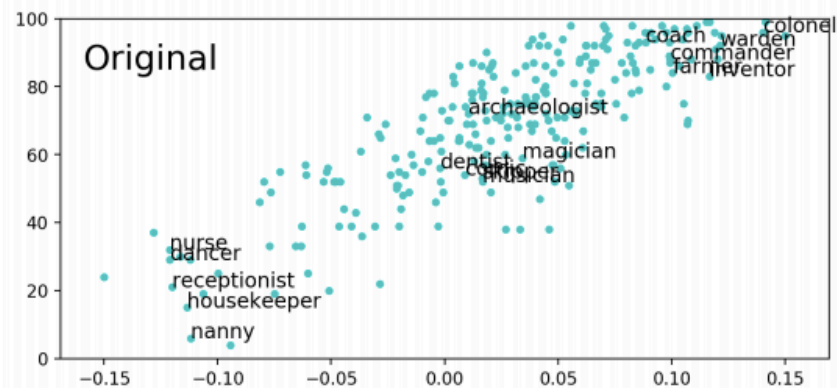
(b) Clustering for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.

Bias by neighbors

- Cannot directly observe the bias
- Bias still manifested by the word being close to socially-marked feminine words
- New mechanism for measuring bias: % male/female socially-biased words among the k nearest neighbors of the target word.



(a) The plots for HARD-DEBIASED embedding, before (top) and after (bottom) debiasing.



(b) The plots for GN-GLOVE embedding, before (top) and after (bottom) debiasing.

Can a classifier learn to predict gender of a word?

- Given some gendered words.
- Can it generalize to others based solely on representation
- Experiment: 5000 most biased words according to original experiments
 - Train an SVM on 1000 random sample, predict gender for remaining 4000


Results

- Hard-debiased
 - 88.8% accuracy vs 98.25% accuracy with non-debiased version
- GN-Glove
 - 96.53% accuracy vs. 98.65% accuracy with non-debiased version

Implications

- Bias is deeply ingrained in the embeddings space
- Real concern is not association with words such as “he”, “she”, “boy”, “girl”
- But of associating one implicitly gendered term with other implicitly gendered terms
 - Picking up gender specific regularities in the corpus
 - Conditioning on gender-biased words and generalizing to other gender-biased words

Reactions?



**Are there aspects that you question or that
you particularly like?**