Ethics

CS4248 Natural Language Processing

Week 13

Min-Yen KAN

13

Recap of Week 12

Contextual Word Embeddings

Machine Translation

Question Answering II

Week 13 Agenda

NLP Ethics

Mitigating Word Embedding Bias

Revision (Separate Deck)

NLP Ethics

How I learned to stop worrying and love natural language processing

Why does a discussion about ethics need to be a part of NLP?

The decisions we make about our methods — training data, algorithm, evaluation — are often tied up with its use and impact in the world.

Slide Credits: David Bamman (UC Berkeley)

The common misconception is that language has to do with words and what they mean.

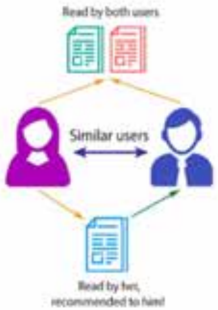
It doesn't.

It has to do with people and what they mean.

Clark & Schober, 1982

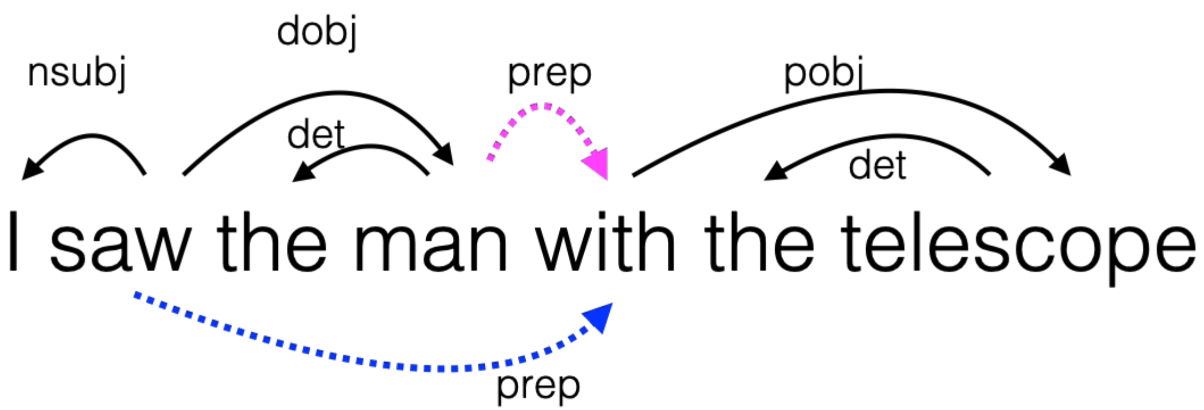
Slide Credits: Diyi Yang (Georgia Tech)

Language, People and the Web



Slide Credits: Diyi Yang (Georgia Tech)

Scope



NLP often operates on text divorced from the context in which it is uttered.

It's now being used more and more to reason about human behavior.

Slide Credit: David Bamman (UC Berkeley)

Learning to Assess Systems Adversarially

- Who could benefit from such a technology?
- Who can be harmed by such a technology?

Representativeness of training data

- Could sharing this data have major effect on people's lives?

- What are confounding variables and corner cases to control for?
- Does the system optimize for the "right" objective?
- Could prediction errors have major effect on people's lives?

Slide Credits: Diyi Yang (Georgia Tech)

Privacy Concerns

- Demographic factors prediction (gender, age, etc)
- Sexual orientation prediction

Dual Use NLP Applications

- E.g., Persuasive language generation
- Socially Beneficial Applications
  - Hate speech detection
  - Monitoring disease outbreaks
  - Psychological monitoring/counseling

  \+ many more

Bias and Fairness Concerns

- Is my NLP model capturing social stereotypes?
- Are my classifiers' predictions fair?

Slide Credits: Diyi Yang (Georgia Tech)

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012

Slide Credit: David Bamman (UC Berkeley)

Facebook fiasco: was Cornell's study of 'emotional contagion' an ethics breach?

A covert experiment to influence the emotions of more than 600,000 people. A major scientific journal behaving like a rabbit in the headlights. A university in a PR tailspin

Slide Credit: David Bamman (UC Berkeley)

Dual Use and Adversarial NLP

Authorship attribution (author of *Federalist Papers* vs. author of ransom note vs. author of political dissent)

Fake review detection vs. fake review generation

Censorship evasion vs. enabling more robust censorship

Slide Credit: David Bamman (UC Berkeley)

Overgeneralization

Managing and communicating the uncertainty of our predictions

Algorithmic Bias: deferring to an automated response.

"The system said so"

Is a false answer worse than no answer?

Slide Credit: David Bamman (UC Berkeley)

Exclusion

Focus on data from one domain/demographic

State-of-the-art models perform worse for young (Hovy and Søgaard, 2015) and minorities (Blodgett *et al.*, 2016)

	AAE	White-Aligned
langid.py	13.2%	7.6%
Twitter-1	8.4%	5.9%
Twitter-2	24.4%	17.6%

Table 3: Proportion of tweets in AA- and white-aligned corpora classified as non-English by different classifiers. (§4.1)

Parser	AA	Wh.	Difference
SyntaxNet	64.0 (2.5)	80.4 (2.2)	16.3 (3.4)
CoreNLP	50.0 (2.7)	71.0 (2.5)	21.0 (3.7)

Language identification

Dependency Parsing

Slide Credit: David Bamman (UC Berkeley)

Biased NLP Technologies

- Bias in Word Embeddings *(Bolukbasi et al. 2017; Caliskan et al. 2017; Garg et al. 2018)*
- Bias in Language ID *(Blodgett & O'Connor. 2017; Jurgens et al. 2017)*
- Bias in Visual Semantic Role Labeling *(Zhao et al. 2017)*
- Bias in Natural Language Inference *(Rudinger et al. 2017)*
- Bias in Coreference Resolution *(Rudinger et al. 2018; Zhao et al. 2018)*
- Bias in Automated Essay Scoring *(Amorim et al. 2018)*

Slide Credits: Diyi Yang (Georgia Tech)

Science AAAS

REPORT

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan[1,*], Joanna J. Bryson[1,2,*], Arvind Narayanan[1,*]

+ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230

Slide Credit: David Bamman (UC Berkeley)

Humans are the "Natural" in NLP

Natural language data and annotations will reflect social/cognitive biases

ML algorithms will replicate biases present in their training data

NLP *is* human subject research! (in a way)

Human subject: a living individual about whom a researcher obtains
(1) data through intervention or interaction with the individual or
(2) identifiable private information.

Mitigating Word Embedding Bias

Slide Credits: Diyi Yang (Georgia Tech)

Language Identification: Solved!

"This paper describes … how even the most simple of these methods using data obtained from the World Wide Web achieve accuracy approaching 100% on a test suite comprised of ten European languages"

…or not?

Slide Credits: Diyi Yang (Georgia Tech)

World Englishes

Slide Credits: Diyi Yang (Georgia Tech)

Bias in Word Embeddings

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "*Semantics derived automatically from language corpora contain human-like biases.*" Science 356, no. 6334 (2017): 183-186.

Slide Credits: Diyi Yang (Georgia Tech)

$$\min cos(he - she,\ x - y)\ s.t.\ ||x - y||_2 < \delta$$

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 1: **Left** The most extreme occupations as projected on to the *she*−*he* gender direction on w2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded. **Right** Automatically generated analogies for the pair *she-he* using the procedure described in text. Each automatically generated analogy is evaluated by 10 crowd-workers to whether or not it reflects gender stereotype.

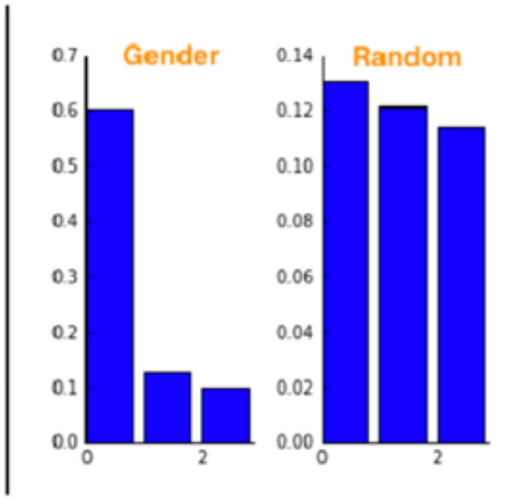
Slide Credits: Diyi Yang (Georgia Tech)

Towards Debiasing

1. Identify gender subspace (direction): B

Bolukbasi et al. (2016) *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*

$$\overrightarrow{\text{she}} - \overrightarrow{\text{he}}$$
$$\overrightarrow{\text{her}} - \overrightarrow{\text{his}}$$
$$\overrightarrow{\text{woman}} - \overrightarrow{\text{man}}$$
$$\overrightarrow{\text{Mary}} - \overrightarrow{\text{John}}$$
$$\overrightarrow{\text{herself}} - \overrightarrow{\text{himself}}$$
$$\overrightarrow{\text{daughter}} - \overrightarrow{\text{son}}$$
$$\overrightarrow{\text{mother}} - \overrightarrow{\text{father}}$$
$$\overrightarrow{\text{gal}} - \overrightarrow{\text{guy}}$$
$$\overrightarrow{\text{girl}} - \overrightarrow{\text{boy}}$$
$$\overrightarrow{\text{female}} - \overrightarrow{\text{male}}$$



The top PC captures the gender subspace

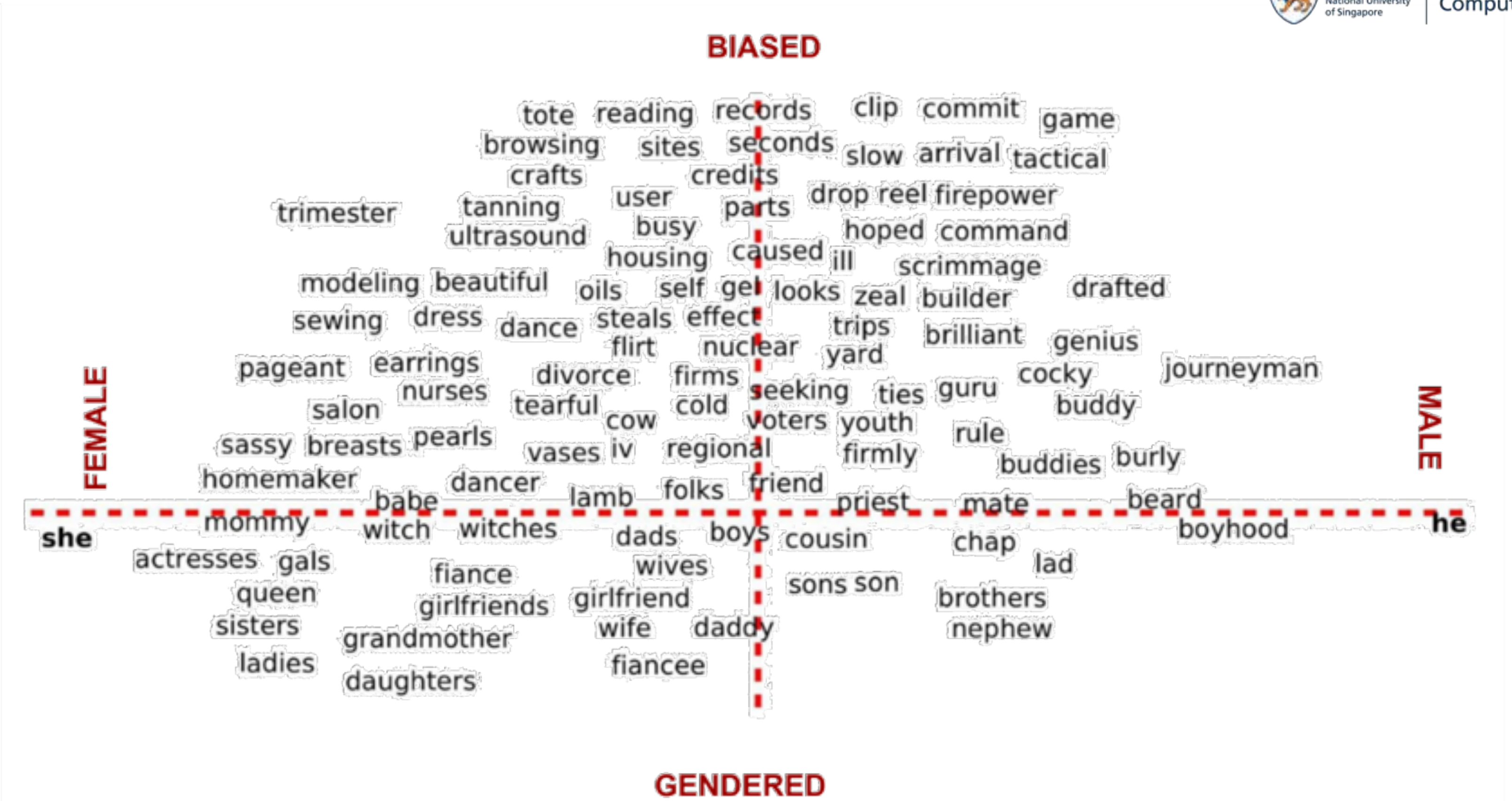
Slide Credits: Diyi Yang (Georgia Tech)

Towards Debiasing

1. Identify gender subspace (direction): B
2. Identify gender-definitional (S) and gender-neutral words (N)

Plus
Bootstrapping

Slide Credits: Diyi Yang (Georgia Tech)



Slide Credits: Diyi Yang (Georgia Tech)

Towards Debiasing

1. Identify gender subspace (direction): B
2. Identify gender-definitional (S) and gender-neutral words (N)
3. Apply matrix transformation (T) to the embedding matrix (W) such that:
   - Project away the gender subspace B from the gender-neutral N
   - While not overly changing the embeddings

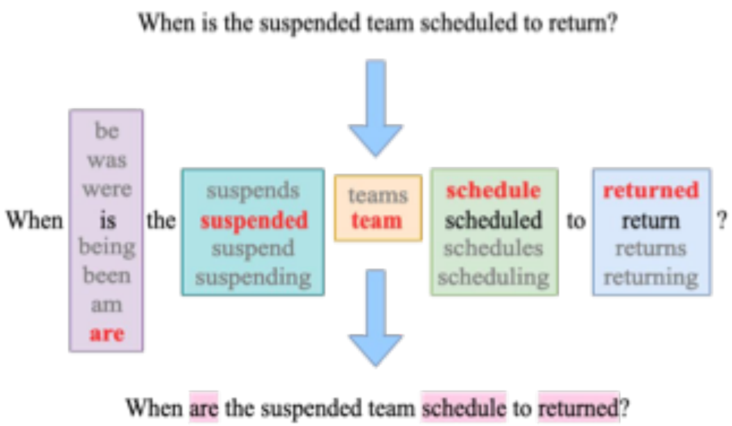
$$min_T \underbrace{||(TW)^T(TW) - W^TW||_F^2}_{\text{Don't modify embeddings too much}} + \lambda \underbrace{||(TN)^T(TB)||_F^2}_{\text{Minimize gender component}}$$

T - the desired debiasing transformation
W - embedding matrix
B - biased space
N - embedding matrix of gender neutral words

Slide Credits: Diyi Yang (Georgia Tech)

Augment the Training Data: Morpheus

Tan et al. (2020) *[It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations]()*



Algorithm 1 Morpheus

Require: Original instance x, Label y, Model f

Ensure: Adversarial example $\hat{x}$

$T \leftarrow \text{TOKENIZE}(x)$

for all $t_i \in T$ **do**

  if $\text{POS}(t_i) \in \{\text{NOUN}, \text{VERB}, \text{ADJ}\}$ **then**

    $I \leftarrow \text{GETINFLECTIONS}(t_i)$

    $t_i \leftarrow \text{MAXINFLECTED}(I, y, f)$

  end if

end for

$\hat{x} \leftarrow \text{DETOKENIZE}(T)$

Ethics Summary

- Who could benefit from your technology?
- Who can be harmed by your technology?

Representativeness of your data

- Could sharing your data have major effect on people's lives?

- What are confounding variables and corner cases for you to control for?
- Does your system optimize for the "right" objective?
- Could prediction errors of your technology have major effect on people's lives?

Course Revision

CS4248 Natural Language Processing

Week 13

Min-Yen KAN

13

Slide Credit: David Bamman (UC Berkeley)

How do you go out and solve new problems involving text?

Slide Credit: David Bamman (UC Berkeley)

1. Language has structure

Slide Credit: David Bamman (UC Berkeley)

"… is a film which still causes real, not figurative, chills to run along my spine, and it is certainly the bravest and most ambitious fruit of Coppola's genius"

Roger Ebert, *Apocalypse Now*

"I hated this movie. Hated hated hated hated hated this movie. Hated it. Hated every simpering stupid vacant audience-insulting moment of it. Hated the sensibility that thought anyone would like it."

Roger Ebert, *North*

Slide Credit: David Bamman (UC Berkeley)

Bag of Words

Representation of text only as the counts of words that it contains

	Apocalypse Now	North
the	1	1
of	0	0
hate	0	9
genius	1	0
bravest	1	0
stupid	0	1
like	0	1
…		

Slide Credit: David Bamman (UC Berkeley)

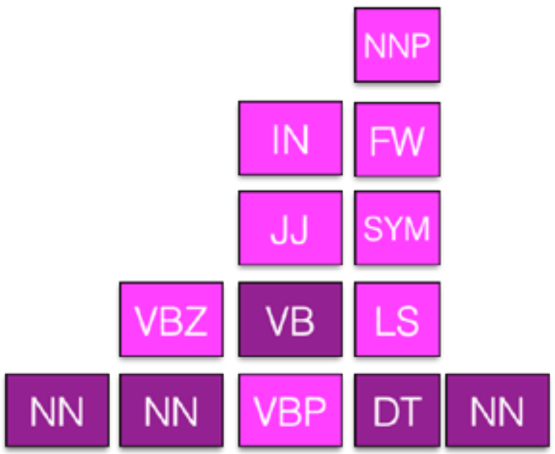
Language Model

"Hillary Clinton seemed to add Benghazi to her already-long list of culprits to blame for her upset loss to Donald ________"

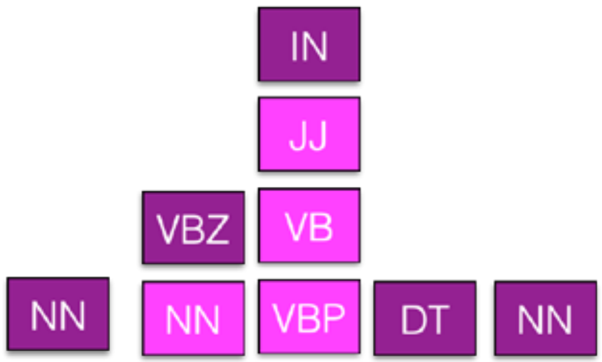
Slide Credit: David Bamman (UC Berkeley)
Source Text: http://www.foxnews.com/politics/2017/09/13/clinton-laments-how-benghazi-tragedy-hurt-her-politically.html

POS Tagging

Labeling the tag that is correct **for the context**.



Fruit flies like a banana.



Time flies like an arrow.

Slide Credit: David Bamman (UC Berkeley)

Word Senses

original	It urged that the city take steps to remedy this problem
lemma sense	It urge^{1} that the city^{2} take^{1} step^{1} to remedy^{1} this problem^{2}
synset number	It $\text{urge}^{2:32:00}$ that the $\text{city}^{1:15:01}$ $\text{take}^{2:41:04}$ $\text{step}^{1:04:02}$ to $\text{remedy}^{2:30:00}$ this $\text{problem}^{1:10:00}$

Slide Credit: David Bamman (UC Berkeley)

Thematic Roles

The thematic roles for verbs generally are predictable by the syntactic position of the argument (specific to each verb class). Some allow for consistent alternations:

Agent Theme Goal

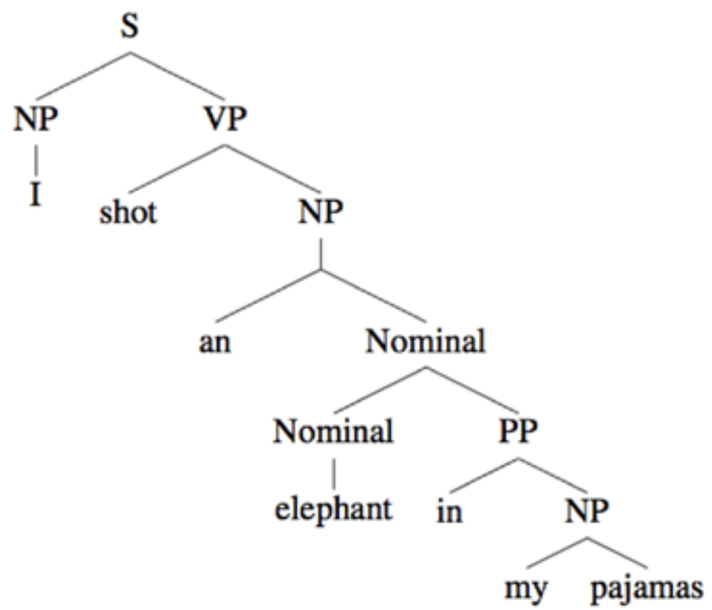
Doris gave the book to Cary

Agent Goal Theme

Doris gave Cary the book

Slide Credit: David Bamman (UC Berkeley)

Phrase Structure Syntax



Every internal node is a phrase

my pajamas

in my pajamas

elephant in my pajamas

an elephant in my pajamas

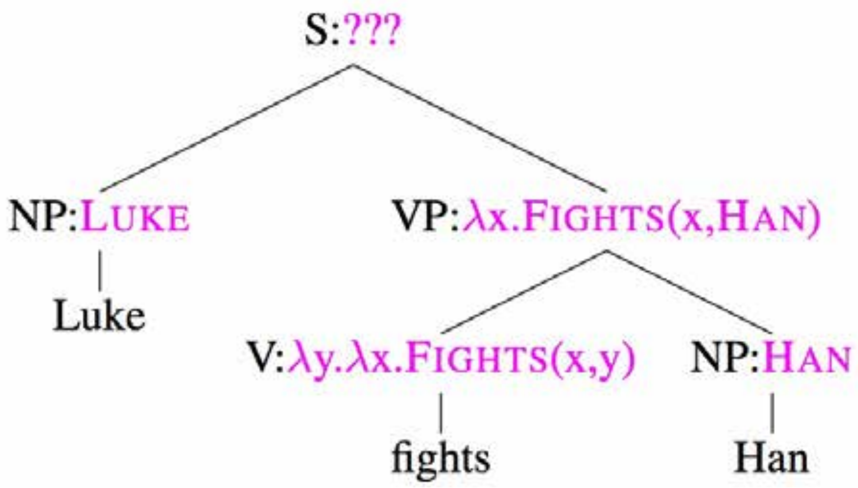
shot an elephant in my pajamas

I shot an elephant in my pajamas

Each phrase could be replaced by another of the same type of constituent

Slide Credit: David Bamman (UC Berkeley)

Compositional Semantics



Slide Credit: David Bamman (UC Berkeley)

Coreference

LUKE
I'll never join you!

VADER
If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE
He told me enough! It was you who killed him.

VADER
No. I am your father.

LUKE
No. No. That's not true! That's impossible!

VADER
Search your feelings. You know it to be true.

LUKE
No! No! No!

Slide Credit: David Bamman (UC Berkeley)

2. Most new problems can be solved with a familiar class of algorithms

Slide Credit: David Bamman (UC Berkeley)

NUS National University of Singapore | School of Computing

- Classification
- Sequence Labeling
- Trees
- Graphs

Counting and normalizing (NB, PCFG, HMM)

Loglinear (logistic regression, MEMM, CRF)

Neural (CNN, RNN, LSTM, seq2seq, attention)

Slide Credit: David Bamman (UC Berkeley)

Classification

Slide Credit: David Bamman (UC Berkeley)

Bayes' Rule

NUS National University of Singapore | School of Computing

Likelihood: How probable is the data given that our document is a member of y?

Prior: How probable is a document to be a member of class y seeing any data?

$$P(y|\mathbf{w}) = \frac{P(\mathbf{w}|y)P(y)}{P(\mathbf{w})}$$

Posterior: How probable is the instance classified as a member of class y?

Marginal: How probable is the evidence under any class?

Slide adapted from CS3244 Machine Learning

Naïve Bayes Classifier

Training a Naïve Bayes classifier consists of estimating these two quantities from training data for all classes Y

At test time, use those estimated probabilities to calculate the posterior probability of each class y and select the class with the highest probability

$$c_{MAP} = \operatorname*{argmax}_{c \in \mathcal{C}} P(c|d)$$

Maximum a posteriori or mostly likely class

$$= \operatorname*{argmax}_{c \in \mathcal{C}} \frac{P(d|c)P(c)}{P(d)}$$

Bayes rule

$$= \operatorname*{argmax}_{c \in \mathcal{C}} P(d|c)P(c)$$

Dropping the P(d) in the denominator

$$= \operatorname*{argmax}_{c \in \mathcal{C}} \overbrace{P(f_1, f_2, ..., f_n|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}$$

Document d represented as features $f_1, ..., f_n$ (such as word counts) BoW assumption

$$= \operatorname*{argmax}_{c \in \mathcal{C}} P(f_1|c)P(f_2|c)...P(f_n)|c)P(c)$$

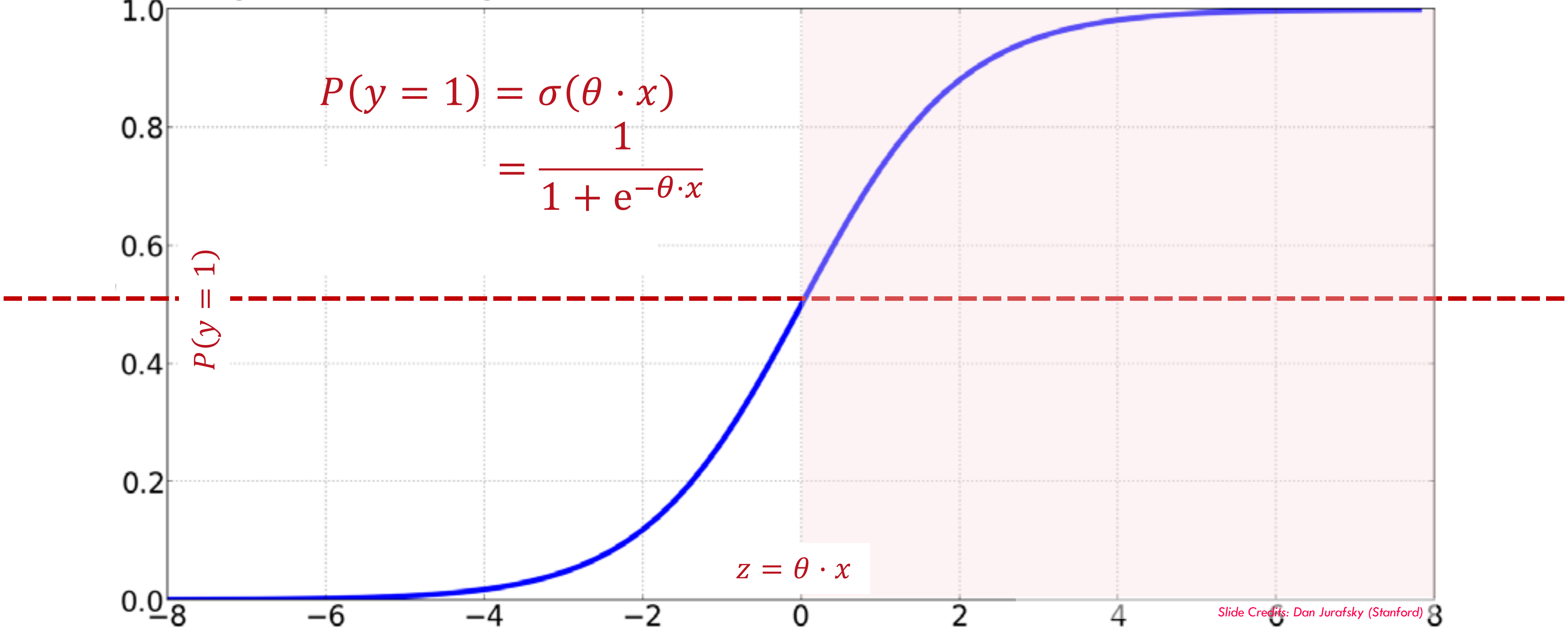
Independence Assumption

$$c_{NB} = \operatorname*{argmax}_{c \in \mathcal{C}} \boxed{P(c)} \prod_{f \in \mathcal{F}} \boxed{P(f|c)}$$

Equation for NB classifier

Slide Credits: David Bamman (UCB)

Logistic Regression



X = feature vector

Feature	Value
the	0
and	0
bravest	0
love	0
loved	0
genius	0
not	0
fruit	1
BIAS	1

W = coefficients

Feature	w
the	0.01
and	0.03
bravest	1.4
love	3.1
loved	1.2
genius	0.5
not	-3.0
fruit	-0.8
BIAS	-0.1

Slide Credit: David Bamman (UC Berkeley)

Features

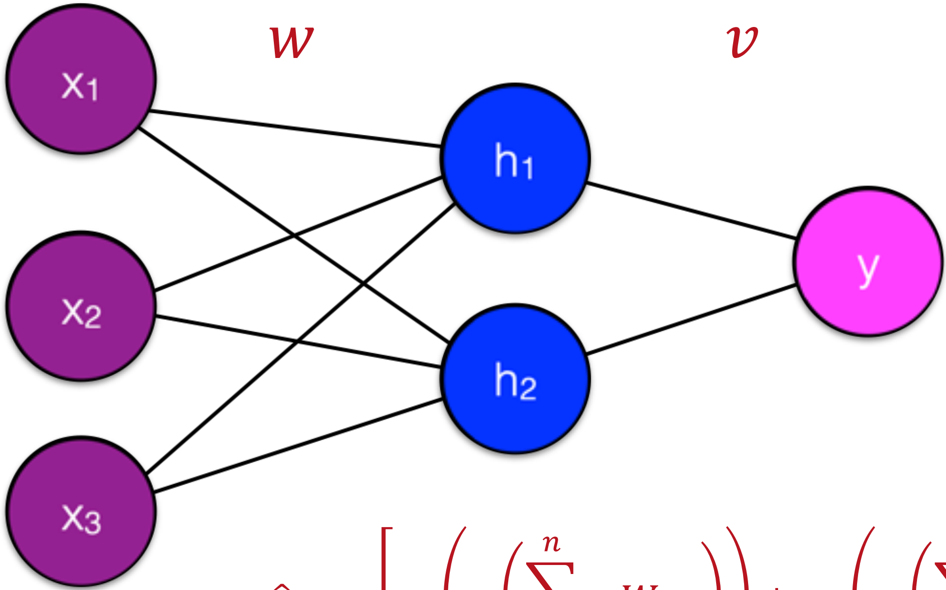
As a discriminative classifier, logistic regression doesn't assume features are independent like Naive Bayes does.

Its power partly comes in the ability to create richly expressive features with out the burden of independence.

We can represent text through features that are not just the identities of individual words, but any feature that is scoped over **the entirety of the input**.

features
contains *like*
has word that shows up in positive sentiment dictionary
review begins with "*I like*"
at least 5 mentions of positive affectual verbs (*like*, *love*, etc.)

Slide Credit: David Bamman (UC Berkeley)



$$\hat{y} = \sigma\left[v_1\left(\sigma\left(\sum_{i}^{n} x_i W_{i,1}\right)\right) + v_2\left(\sigma\left(\sum_{i}^{n} x_i W_{i,2}\right)\right)\right]$$

We can express y as a function only of the input x and the weights W and V

Slide Credit: David Bamman (UC Berkeley)

Sequences

Slide Credit: David Bamman (UC Berkeley)

Sequence Labeling

Sequence labeling problems make a labeling decision at each timestep

B-PER	I-PER	O	O	O	O	B-ORG
Tim	Cook	is	the	CEO	of	Apple

0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
#	b	l	a	c	k	l	i	v	e	s	m	a	t	t	e	r

Slide Credit: David Bamman (UC Berkeley)

Sequence Labeling

$$X = \{x_1, \ldots, x_n\}$$

$$Y = \{y_1, \ldots, y_n\}$$

For a set of inputs x with n sequential time steps, one corresponding label y_i for each x_i

Model the structure that exists between within y

Slide Credit: David Bamman (UC Berkeley)

HMM

$$P(x_1, \ldots, x_n, y_1, \ldots, y_n) \approx \prod_{i=1}^{n+1} P(y_i \mid y_{i-1}) \prod_{i=1}^{n} P(x_i \mid y_i)$$

Slide Credit: David Bamman (UC Berkeley)

Hidden Markov Model

$$P(x \mid y) = P(x_1, \ldots, x_n \mid y_1, \ldots, y_n)$$

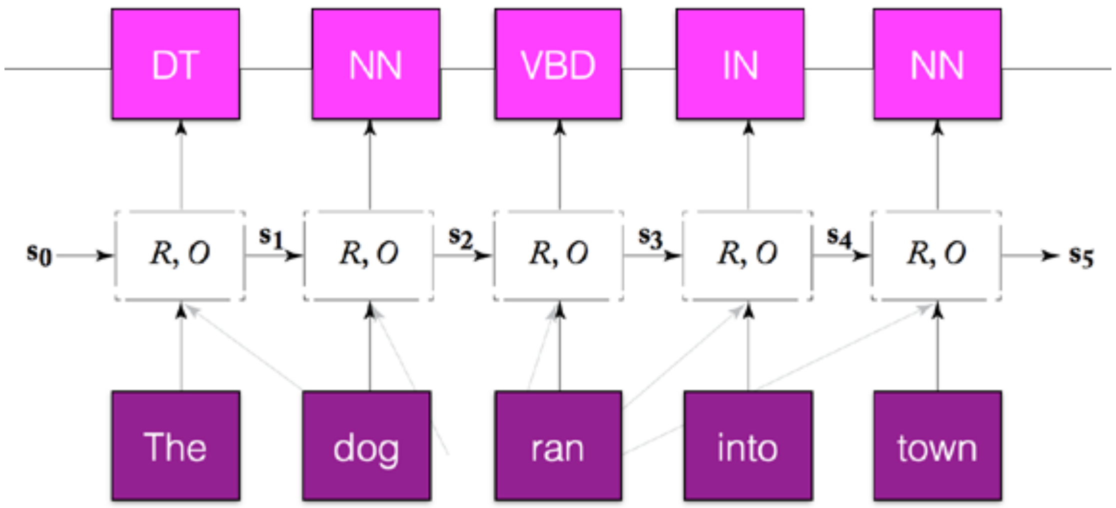
$$P(x_1, \ldots, x_n \mid y_1, \ldots, y_n) \approx \prod_{i=1}^{N} P(x_i \mid y_i)$$

Here again we'll make a strong assumption: the probability of the word we see at a given time step is only dependent on its label

Slide Credit: David Bamman (UC Berkeley)

Recurrent Neural Network

Predict the tag conditioned on the context



Slide Credit: David Bamman (UC Berkeley)

Bidirectional RNN

A powerful alternative is make predictions conditioning both on the past and the future.

Two RNNs

- One running left-to-right
- One right-to-left

Each produces an output vector at each time step, which we concatenate

Slide Credit: David Bamman (UC Berkeley)

Trees

Slide Credit: David Bamman (UC Berkeley)

PCFG

Probabilistic context-free grammar: each production is also associated with a probability.

This lets us calculate the probability of a parse for a given sentence; for a given parse tree T for sentence S comprised of n rules from R (each $A \rightarrow \beta$):

$$P(T, S) = \prod_{i}^{n} P(\beta | A)$$

Slide Credit: David Bamman (UC Berkeley)

Estimating PCFGs

$$\sum_{\beta} P(\beta|A) = \frac{Count(A \rightarrow \beta)}{\sum_i Count(A \rightarrow i)}$$

Or equivalently,

$$\sum_{\beta} P(\beta|A) = \frac{Count(A \rightarrow \beta)}{Count(A)}$$

Slide Credit: David Bamman (UC Berkeley)

A		β	P(β | NP)
NP	→	NP PP	0.092
NP	→	DT NN	0.087
NP	→	NN	0.047
NP	→	NNS	0.042
NP	→	DT JJ NN	0.035
NP	→	NNP	0.034
NP	→	NNP NNP	0.029
NP	→	JJ NNS	0.027
NP	→	QP -NONE-	0.018
NP	→	NP SBAR	0.017
NP	→	NP PP-LOC	0.017
NP	→	JJ NN	0.015
NP	→	DT NNS	0.014
NP	→	CD	0.014
NP	→	NN NNS	0.013
NP	→	DT NN NN	0.013
NP	→	NP CC NP	0.013

Slide Credit: David Bamman (UC Berkeley)

Natural Language Generation

Slide Credit: David Bamman (UC Berkeley)

Language Model

Language modeling is the task of estimating $P(w)$

- Count and normalize
- Featurized
- Neural (RNN)

Slide Credit: David Bamman (UC Berkeley)

Encoder–Decoder Framework

Language modeling: predict a word given its left context.

How about when there's some prior information?

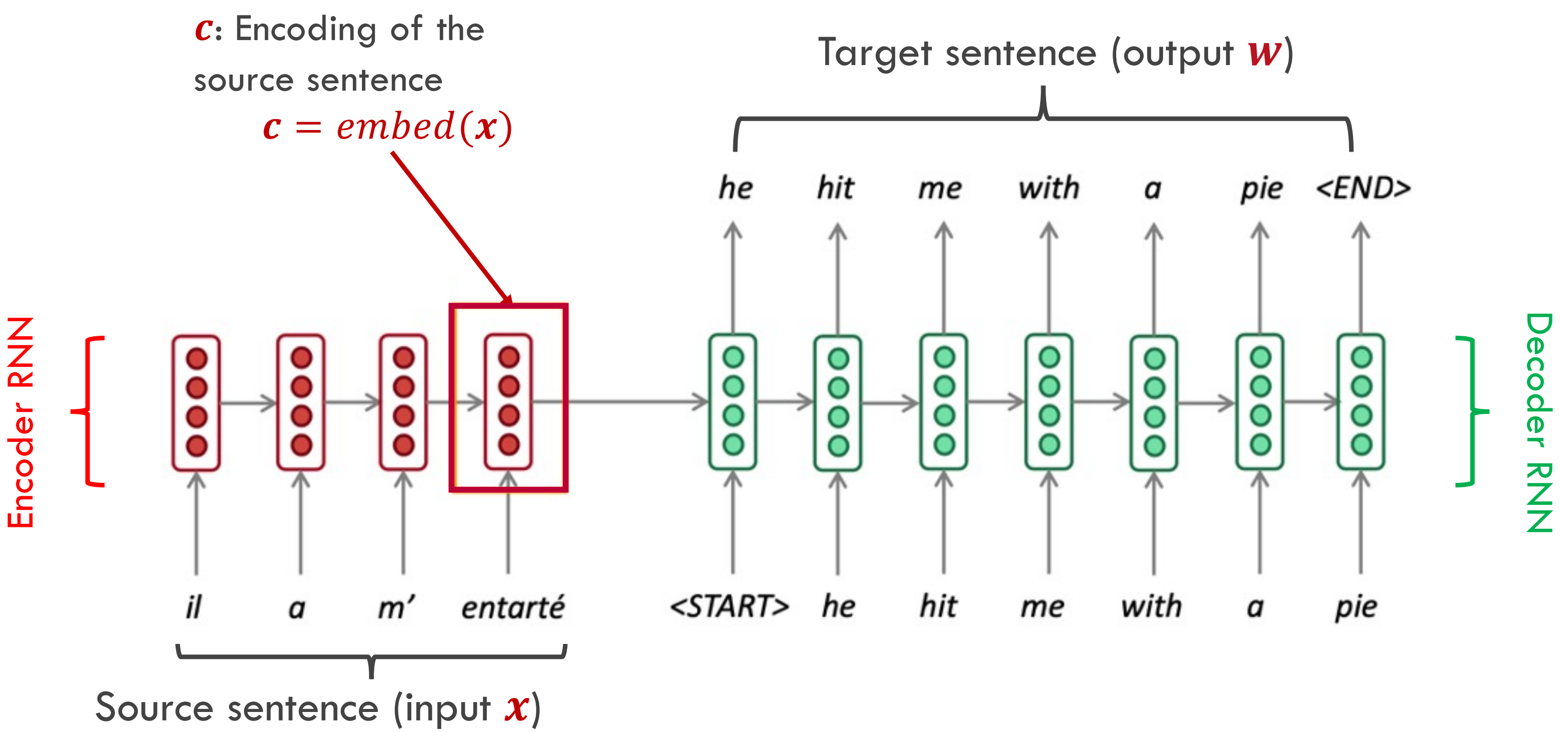
Conditional Language Model

- Question Answering: predict an answer given its left context and the source passage.
- Machine translation: predict a word given its left context and the full text of the source.

Basic idea: encode some context into a fixed vector; and then decode a new sentence from that embedding.

Slide Credit: David Bamman (UC Berkeley)

Encoder–Decoder

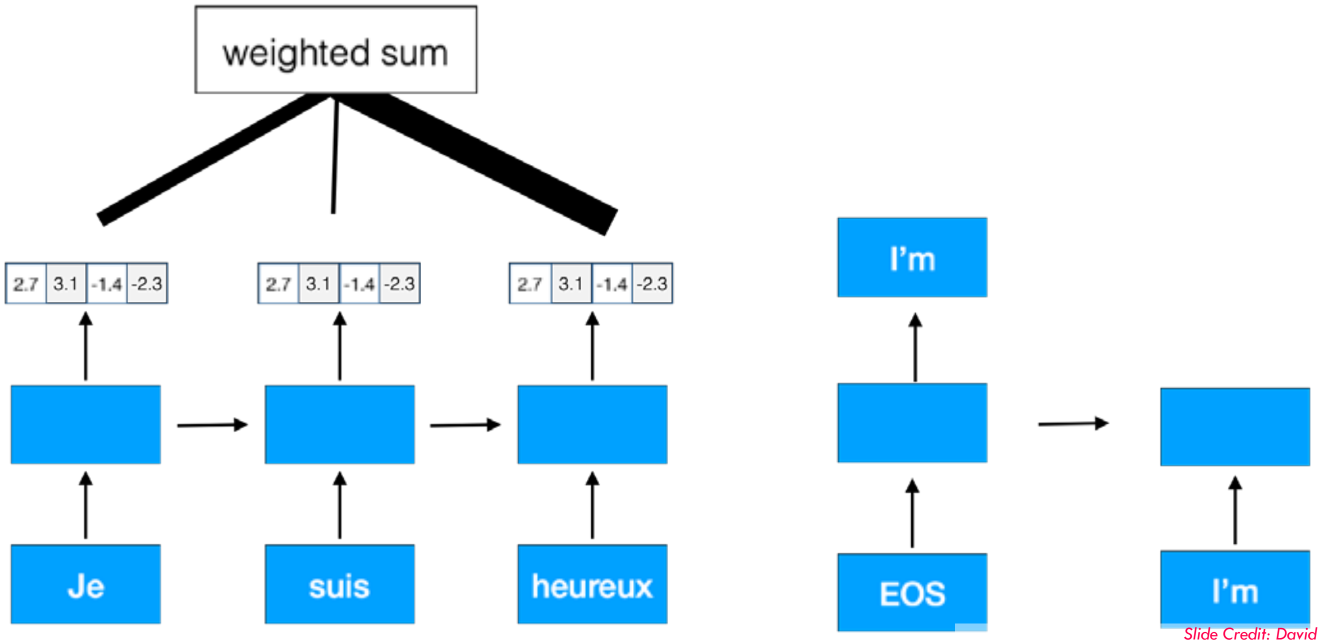


Adaptedd from Chris Manning (Stanford) CS224N

Encoder–Decoder with Attention

$$c = h_1 a_1 + h_2 a_2 + h_3 a_3$$

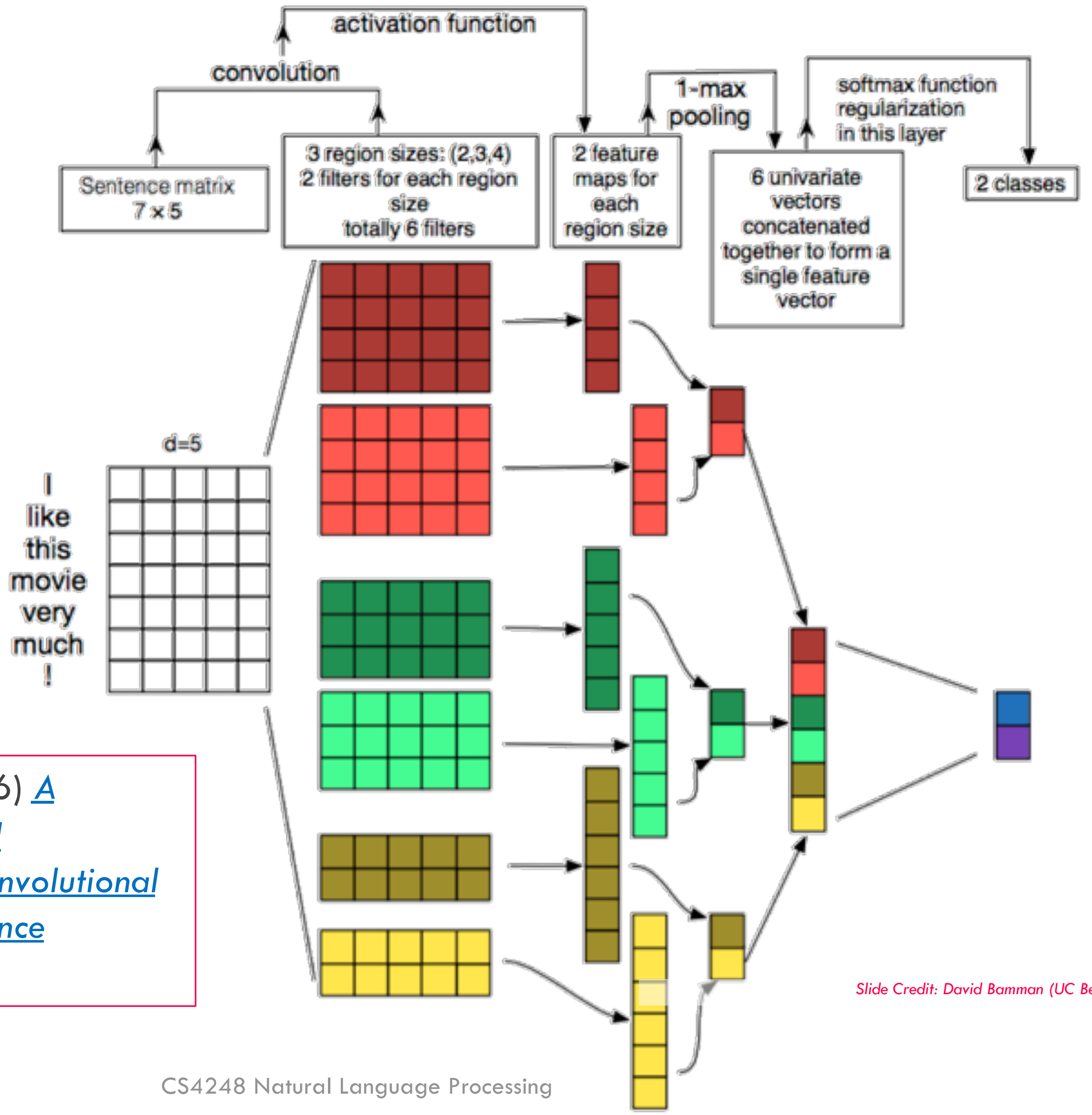
$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$



Slide Credit: David Bamman (UC Berkeley)

3. Neural methods are generally* better.

Slide Credit: David Bamman (UC Berkeley)

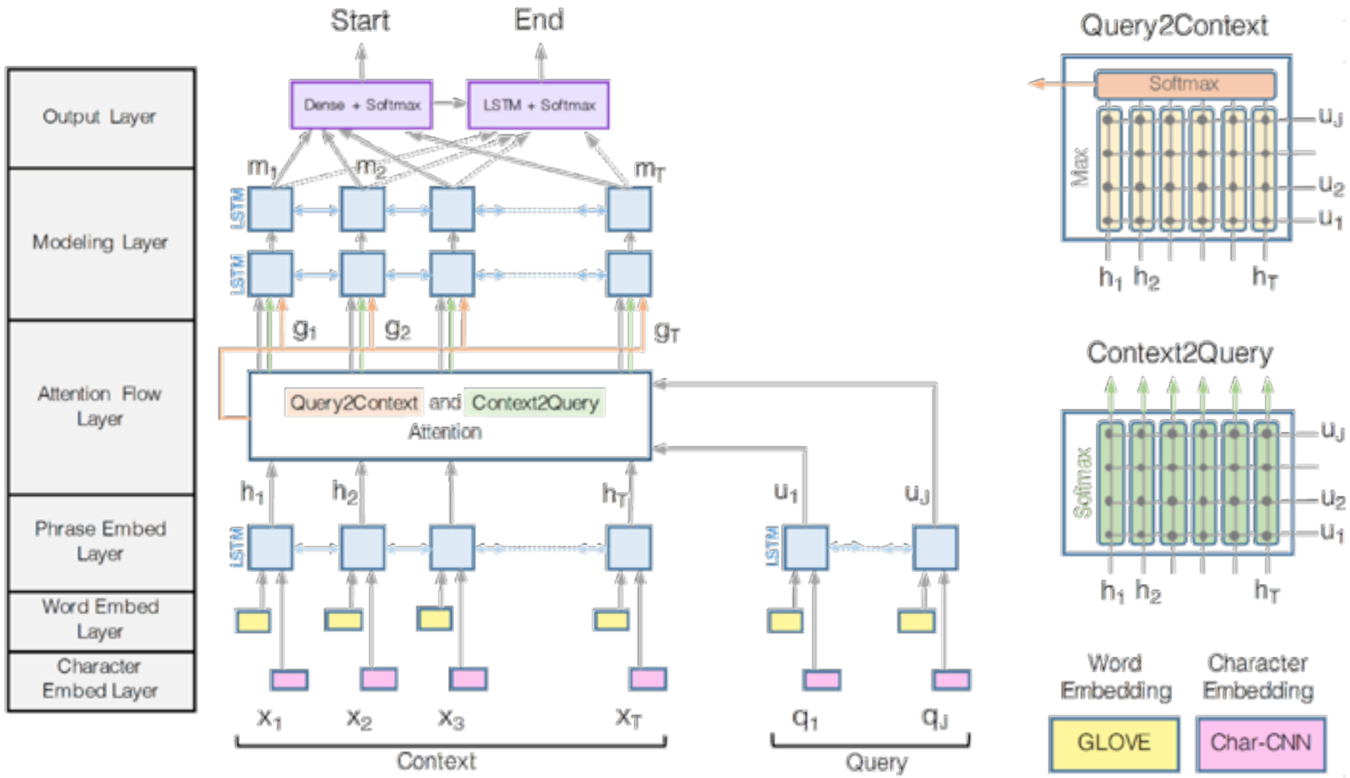


Zhang and Wallace (2016) [A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification]

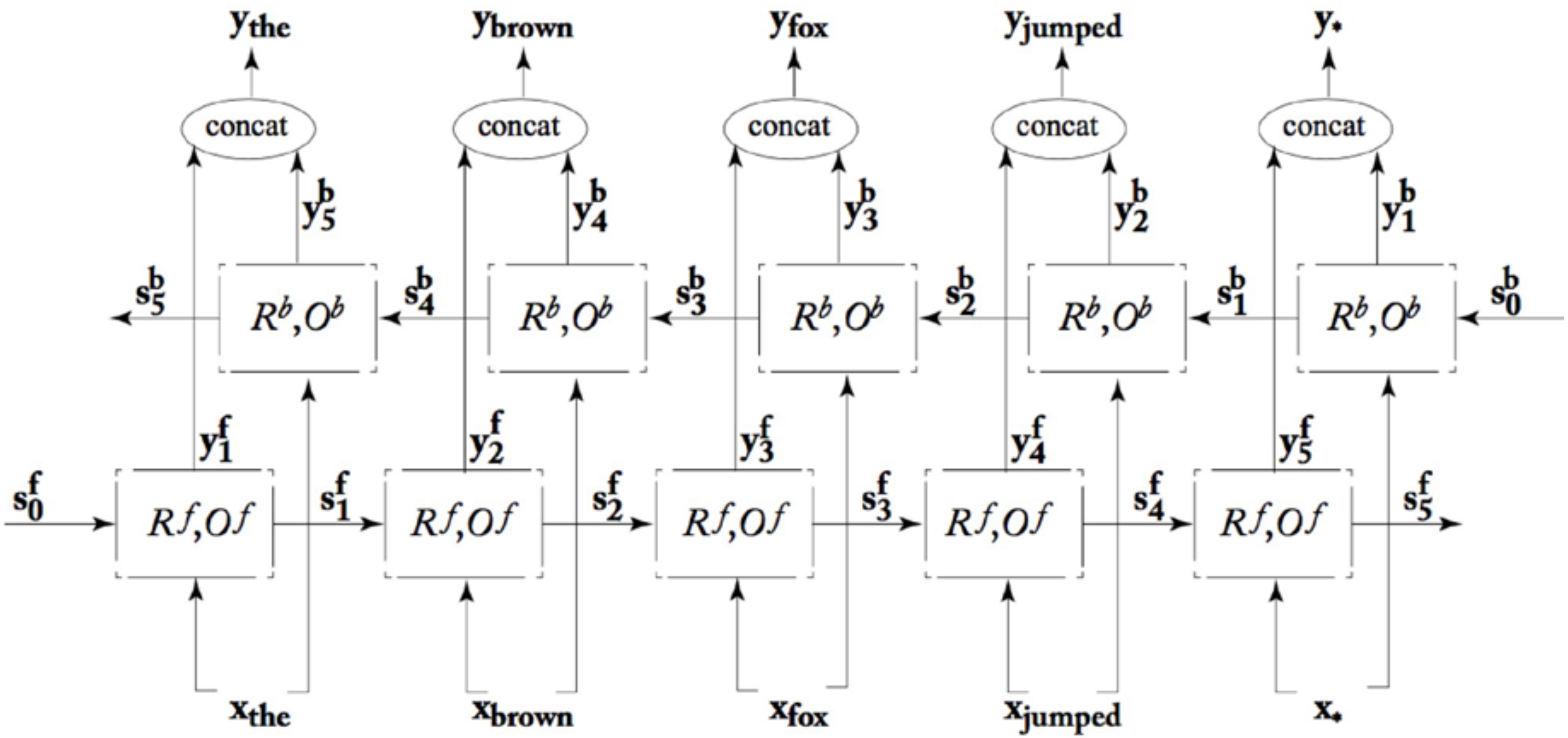
Slide Credit: David Bamman (UC Berkeley)

BiDAF

(Seo et al., 2017): [Bidirectional Attention Flow for Machine Comprehension]



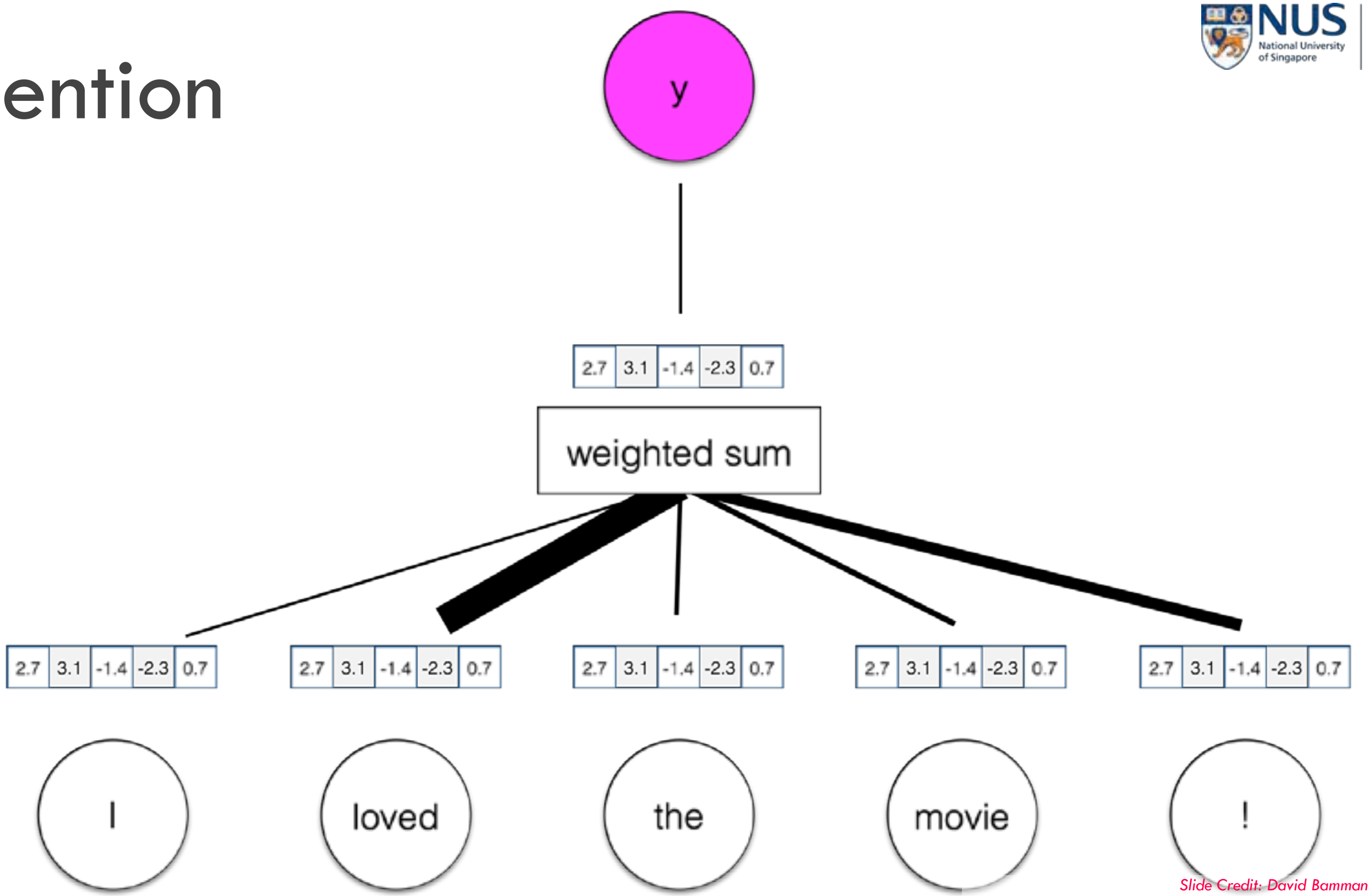
Slide Credits: Diyi Yang (Georgia Tech)



Goldberg (2017)

Slide Credit: David Bamman (UC Berkeley)

Attention



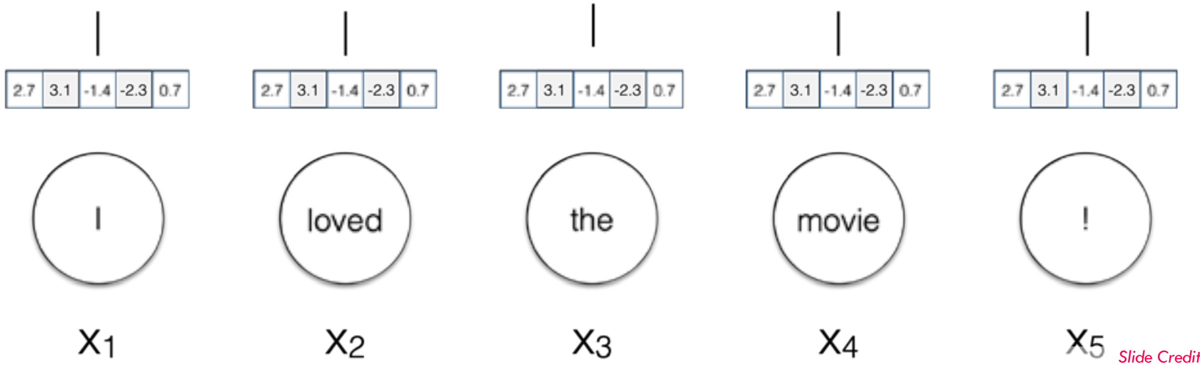
Slide Credit: David Bamman (UC Berkeley)

$$v \in \mathcal{R}^H$$

2.7	3.1	-1.4	-2.3	0.7

Define v to be a vector to be learned; think of it as an "important word" vector. The dot product here measures how similar each input vector is to that "important word" vector.

$r_1 = v^\top x_1 \quad r_2 = v^\top x_2 \quad r_3 = v^\top x_3 \quad r_4 = v^\top x_4 \quad r_5 = v^\top x_5$



Slide Credit: David Bamman (UC Berkeley)

Lexical semantics

"You shall know a word by the company it keeps"

[Firth 1957]

Slide Credit: David Bamman (UC Berkeley)

Distributed Representation

Vector representation that encodes information about the distribution of contexts a word appears in

Words that appear in similar contexts have similar representations (and similar meanings, by the distributional hypothesis).

Slide Credit: David Bamman (UC Berkeley)

4. Evaluation is critical.

Slide Credit: David Bamman (UC Berkeley)

Interannotator Agreement

Annotator 2 \ Annotator 1	puppy	fried chicken
puppy	6	3
fried chicken	2	5

observed agreement = 11/16 = 68.75%

Slide Credit: David Bamman (UC Berkeley)
Source Image: https://twitter.com/teenybiscuit/status/705232709220769792/photo/1

Experiment Design

	training	development	testing
size	80%	10%	10%
purpose	training models	model selection	evaluation; never look at it until the very end

Slide Credit: David Bamman (UC Berkeley)

Metrics

- Perplexity
- Accuracy
- Precision/Recall/F_1
- Parseval (P/R/F_1 over labeled constituents)
- Correlation with human judgments
- BLEU Precision / ROUGE Recall

Slide Credit: David Bamman (UC Berkeley)

5. Text is data.

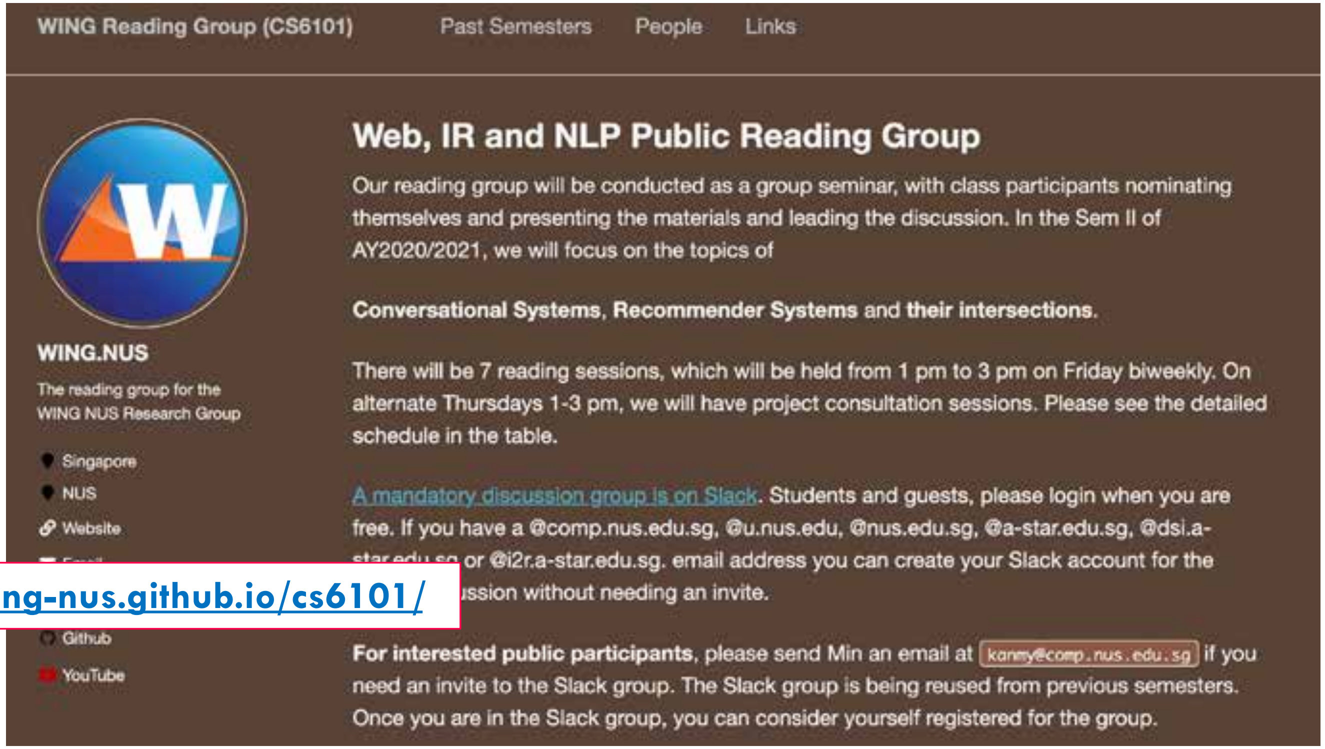
Slide Credit: David Bamman (UC Berkeley)

Slide Credit: David Bamman (UC Berkeley)

Keep in touch with CS4248!

https://www.facebook.com/cs4248

Level up with us!



WING Reading Group (CS6101) Past Semesters People Links

WING.NUS

The reading group for the WING NUS Research Group

- Singapore
- NUS
- Website
- Github
- YouTube

Web, IR and NLP Public Reading Group

Our reading group will be conducted as a group seminar, with class participants nominating themselves and presenting the materials and leading the discussion. In the Sem II of AY2020/2021, we will focus on the topics of

Conversational Systems, **Recommender Systems** and **their intersections**.

There will be 7 reading sessions, which will be held from 1 pm to 3 pm on Friday biweekly. On alternate Thursdays 1-3 pm, we will have project consultation sessions. Please see the detailed schedule in the table.

A mandatory discussion group is on Slack. Students and guests, please login when you are free. If you have a @comp.nus.edu.sg, @u.nus.edu, @nus.edu.sg, @a-star.edu.sg, @dsi.a-star.edu.sg or @i2r.a-star.edu.sg. email address you can create your Slack account for the ssion without needing an invite.

For interested public participants, please send Min an email at `kanmy@comp.nus.edu.sg` if you need an invite to the Slack group. The Slack group is being reused from previous semesters. Once you are in the Slack group, you can consider yourself registered for the group.

https://wing-nus.github.io/cs6101/

NUS National University of Singapore | School of Computing

Goodbye!