Reminders

HOMEWORK 12 / MILESTONE 4 ARE DUE DUE WEDNESDAY.

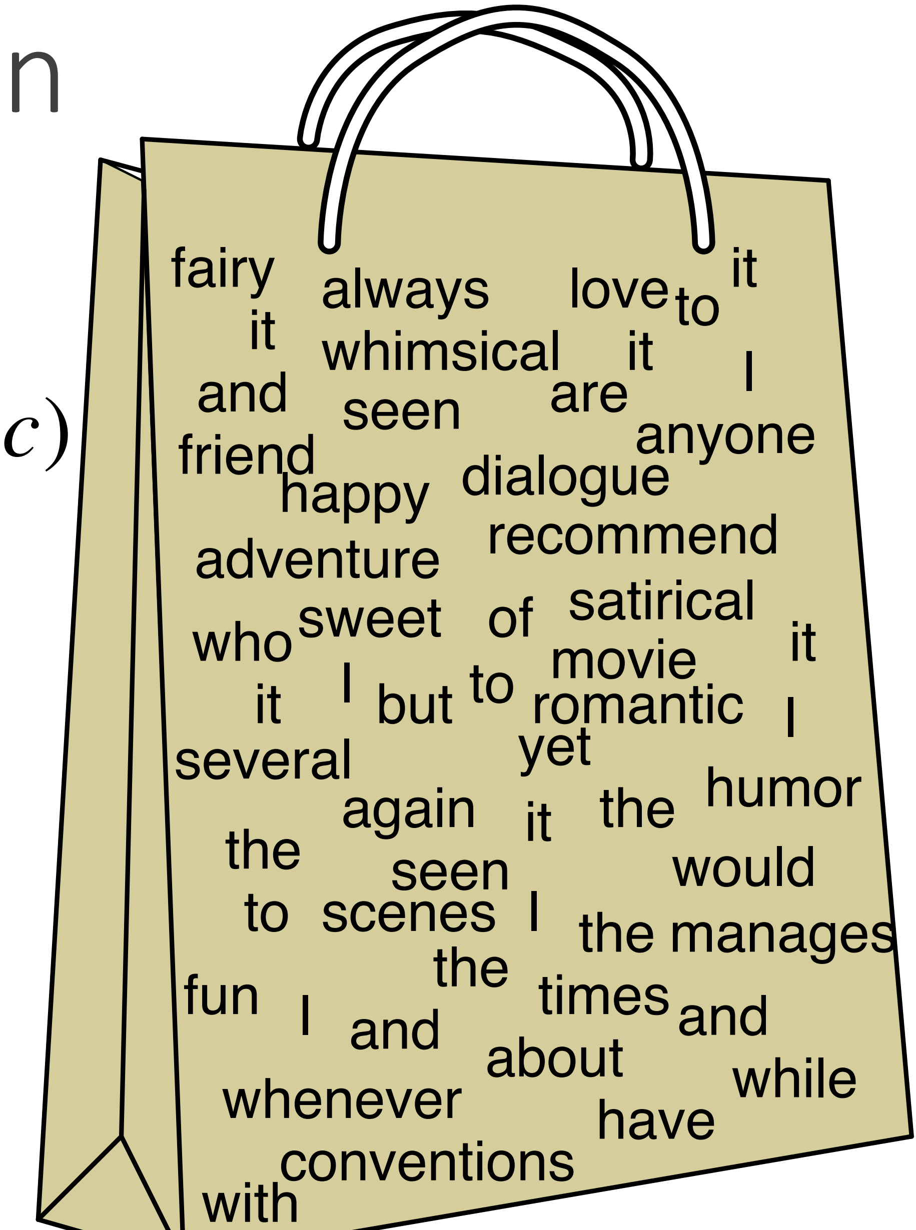
SIGN-UP FOR A 30-MINUTE SLOT TO PRESENT YOUR FINAL PROJECT AT ccb.youcanbook.me

What have we learned?

FINAL LECTURE OF
CIS 530

Text Classification

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}}\, P(c_j) \prod_{x \in X} P(x \mid c)$$



Regular Expressions

and Hearst Patterns

The bow lute, **such as** the Bambara ndang, is plucked and has an individual curved neck for each string

Morphology

Morphemes:

- The small meaningful units that make up words
- **Stems**: The core meaning-bearing units
- **Affixes**: Bits and pieces that adhere to stems
- Often with grammatical functions

Stemming

Reduce terms to their stems in information retrieval

Stemming is crude chopping of affixes

- language dependent
- e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.

for example compressed and compression are both accepted as equivalent to compress.

for exampl compress and compress ar both accept as equival to compress

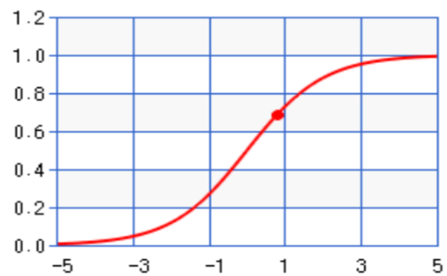
Word Pieces via Byte Pair Encoding

Merge	**Current Vocabulary**
(n, ew)	_, d, e, i, l, n, o, r, s, t, w, r_, er_, ew, new
(l, o'	_, d, e, i, l, n, o, r, s, t, w, r_, er_, ew, new, lo
(lo, w)	_, d, e, i, l, n, o, r, s, t, w, r_, er_, ew, new, lo, low
(new, er_)	_, d, e, i, l, n, o, r, s, t, w, r_, er_, ew, new, lo, low, newer_
(low, _)	_, d, e, i, l, n, o, r, s, t, w, r_, er_, ew, new, lo, low, newer_, low_

Logistic Regression

Var	Definition	Value	Weight	Product
x_1	Count of positive lexicon words	3	2.5	7.5
x_2	Count of negative lexicon words	2	-5.0	-10
x_3	Does no appear? (binary feature)	1	-1.2	-1.2
x_4	Num 1st and 2nd person pronouns	3	0.5	1.5
x_5	Does ! appear? (binary feature)	0	2.0	0
x_6	Log of the word count for the doc	4.15	0.7	2.905
b	bias	1	0.1	.1

$$P(y = positive) = \sigma(w \cdot x + b)$$



$$= \sigma(0.805) = 0.69$$

Cross-entropy loss

Why does minimizing this negative log probability do what we want? We want the **loss** to be **smaller** if the model's estimate is **close to correct**, and we want the **loss** to be **bigger if it is confused.**

It's hokey. There are virtually no surprises , and the writing is second-rate . So why was it so enjoyable? For one thing , the cast is great . Another nice touch is the music . I was overcome with the urge to get off the couch and start dancing . It sucked me in , and it'll do the same to you .

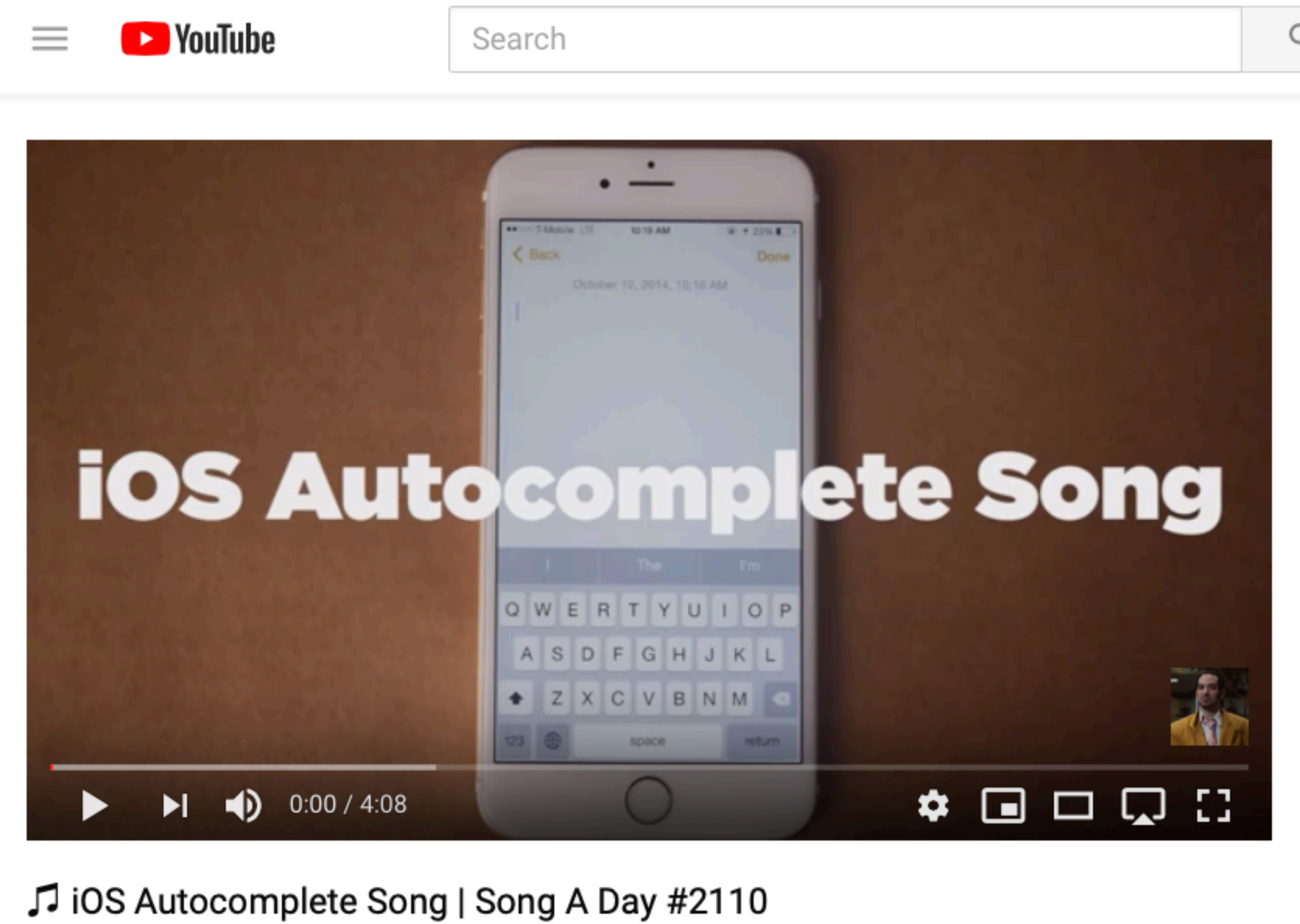
P(sentiment=1|It's hokey...) = 0.69. Let's say y=1.

$$L_{CE}(\hat{y}, y) = -[y \log \sigma(w \cdot x + b) + (1-y) \log(1 - \sigma(w \cdot x + b))]$$

$$= -[\log \sigma(w \cdot x + b)]$$

$$= -\log(0.69) = \mathbf{0.37}$$

Gradient Descent



https://www.youtube.com/watch?v=M8MJFrdfGe0

N-Gram Language Models

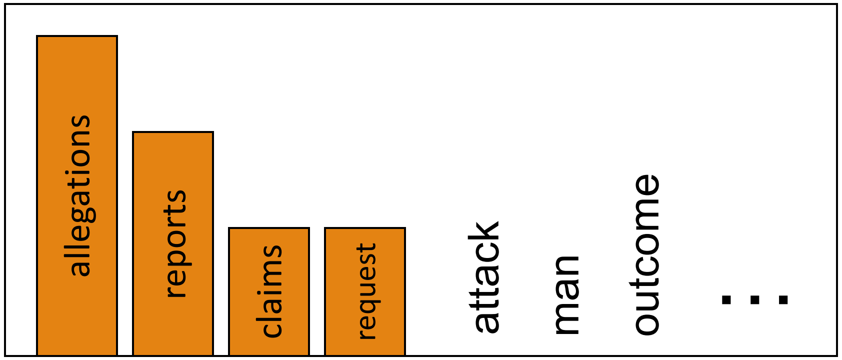
unigram	**no history**	$\prod_{i}^{n} \mathrm{p}(w_i)$	$p(w_i) = \frac{count(w_i)}{all\ words}$
bigram	1 word as history	$\prod_{i}^{n} \mathrm{p}(w_i \vert w_{i-1})$	$p(w_i \vert w_{i-1}) = \frac{count(w_{i-1}w_i)}{count(w_{i-1})}$
trigram	2 words as history	$\prod_{i}^{n} \mathrm{p}(w_i \vert w_{i-2}w_{i-1})$	$p(w_i \vert w_{i-2}w_{i-1}) = \frac{count(w_{i-2}w_{i-1}w_i)}{count(w_{i-2}w_{i-1})}$
4-gram	3 words as history	$\prod_{i}^{n} \mathrm{p}(w_i \vert w_{i-3}w_{i-2}w_{i-1})$	$p(w_i \vert w_{i-3}w_{i-2}w_{i-1}) = \frac{count(w_{i-3}w_{i-2}w_{i-1}w_i)}{count(w_{i-3}w_{i-3}w_{i-1})}$

Smoothing

When we have sparse statistics:

P(w | denied the)
3 allegations
2 reports
1 claims
1 request

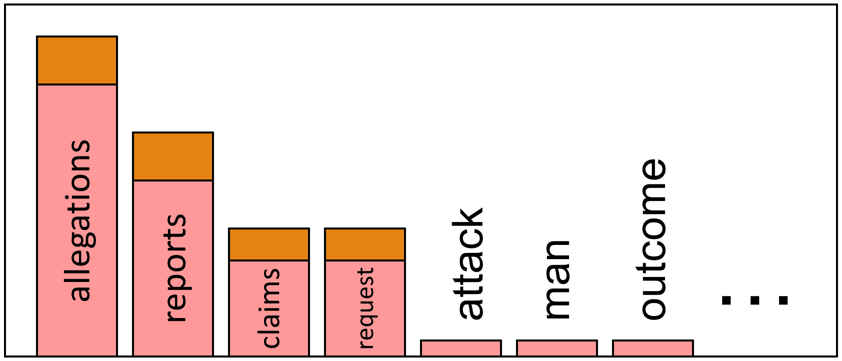
7 total



Steal probability mass to generalize better

P(w | denied the)
2.5 allegations
1.5 reports
0.5 claims
0.5 request
2 other

7 total



Approximating Shakespeare

1 gram	–To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have –Hill he late speaks; or! a more to leg less first you enter
2 gram	–Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. –What means, sir. I confess she? then all sorts, he is trim, captain.
3 gram	–Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, ’tis done. –This shall forbid it should be branded, if renown made it empty.
4 gram	–King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv’d in; –It cannot be but so.

Distributional Hypothesis

If we consider *optometrist* and *eye-doctor* we find that, as our corpus of utterances grows, these two occur in almost the same environments. In contrast, there are many sentence environments in which *optometrist* occurs but *lawyer* does not...

It is a question of the relative frequency of such environments, and of what we will obtain if we ask an informant to substitute any word he wishes for optometrist (not asking what words have the same meaning).

These and similar tests all measure the probability of particular environments occurring with particular elements... If A and B have almost identical environments we say that they are synonyms.
–Zellig Harris (1954)

Term-Document Matrix

	D1	D2	D3	D4	D5
abandon					
abdicate					
abhor					
academic					
…					
zygodactyl					
zymurgy					

We can measure how similar two documents are by comparing their column vectors

Term-Document Matrix

	D1	D2	D3	D4	D5
abandon					
abdicate					
abhor					
academic					
…					
zygodactyl					
zymurgy					

We can measure word similarity by comparing two row vectors

Sparse Representations

Term-Document Matrices are

- **long** (length |V|= 20,000 to 50,000)
- **sparse** (most elements are zero)

Word embeddings

We shifted vectors which are

- **short** (length 50-1000)
- **dense** (most elements are non-zero)
- **learned representations** (not just counts)

Word2Vec Training

Training sentence:

... lemon, a tablespoon of **apricot** jam a pinch ...

c1 c2 t c3 c4

Training data: input/output pairs centering on *apricot*

Assume a +/- 2 word window

Word2Vec Training

Training sentence:

... lemon, a tablespoon of **apricot** jam a pinch ...

c1 c2 t c3 c4

positive examples +

t	c
apricot	tablespoon
apricot	of
apricot	preserves
apricot	or

For each positive example, we'll create *k* negative examples.

Using *noise* words

Any random word that isn't *t*

k-Nearest Neighbors

```
vectors.most_similar("pandemic")

[('flu_pandemic', 0.8552696),
 ('influenza_pandemic', 0.85390663
 ('pandemic_flu', 0.79758835),
 ('pandemic_influenza', 0.77891153
 ('H#N#_pandemic', 0.7507599),
 ('bird_flu_pandemic', 0.73017865)
 ('avian_flu_pandemic', 0.7283541)
 ('swine_flu_pandemic', 0.7262859)
 ('avian_influenza_pandemic', 0.71
 ('pandemics', 0.6915629)]
```

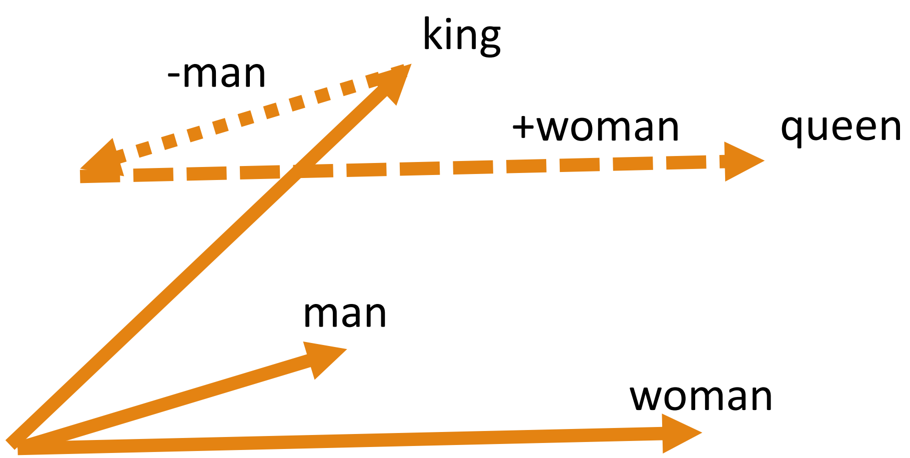
```
vectors.most_similar("coronavirus")

[('corona_virus', 0.7276226),
 ('coronaviruses', 0.7216537),
 ('paramyxovirus', 0.71130025),
 ('SARS_coronavirus', 0.66019076),
 ('arenavirus', 0.64944106),
 ('influenza_virus', 0.64498264),
 ('H#N#_subtype', 0.636014),
 ('H#N#_strain', 0.6324742),
 ('H7_virus', 0.6261192),
 ('flu_virus', 0.62492055)]
```

Word Analogies

a:a∗ as **b:b∗. b∗** is a hidden vector.

b∗ should be similar to the vector b − a + a∗

vector(*‘king’*) - vector(*‘man’*) + vector(*‘woman’*) ≈ vector(‘queen’)



Word Analogies

a:a∗ as **b:b∗. b∗** is a hidden vector.

b∗ should be similar to the vector b − a + a∗

vector(*'king'*) - vector(*'man'*) + vector(*'woman'*) ≈ vector(*'queen'*)

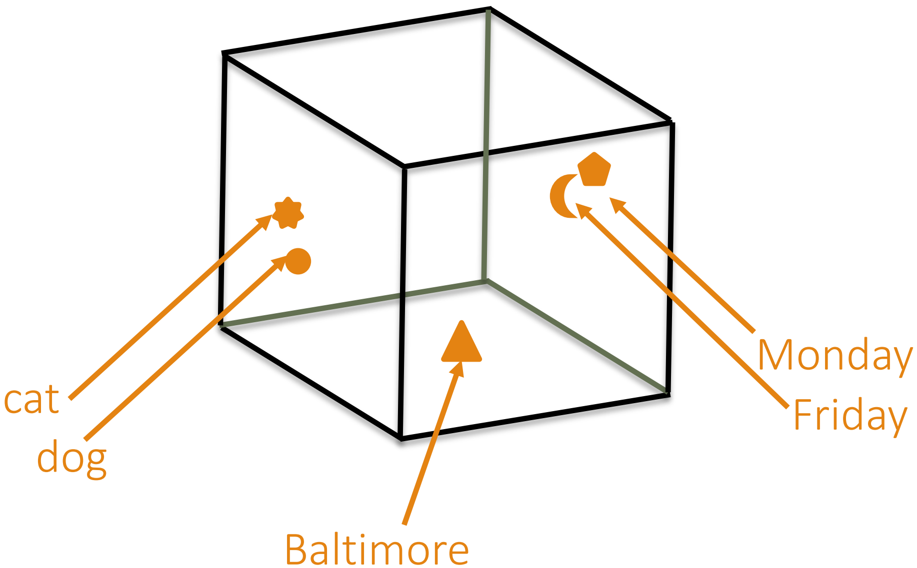
```
vectors.most_

[('queen', 0.
 ('monarch',
 ('princess',
 ('crown_prin
 ('prince', 0
 ('kings', 0.
 ('Queen_Cons
 ('queens', 0
 ('sultan', 0
 ('monarchy',
```

```
vectors.most_similar(negative = ["man"],
                     positive=["woman",
                               "computer_programmer"])

[('homemaker', 0.5627119),
 ('housewife', 0.5105047),
 ('graphic_designer', 0.50518024),
 ('schoolteacher', 0.49794948),
 ('businesswoman', 0.4934892),
 ('paralegal', 0.49255118),
 ('registered_nurse', 0.49079752),
 ('saleswoman', 0.48816282),
 ('electrical_engineer', 0.47977278),
 ('mechanical_engineer', 0.47553998)]
```

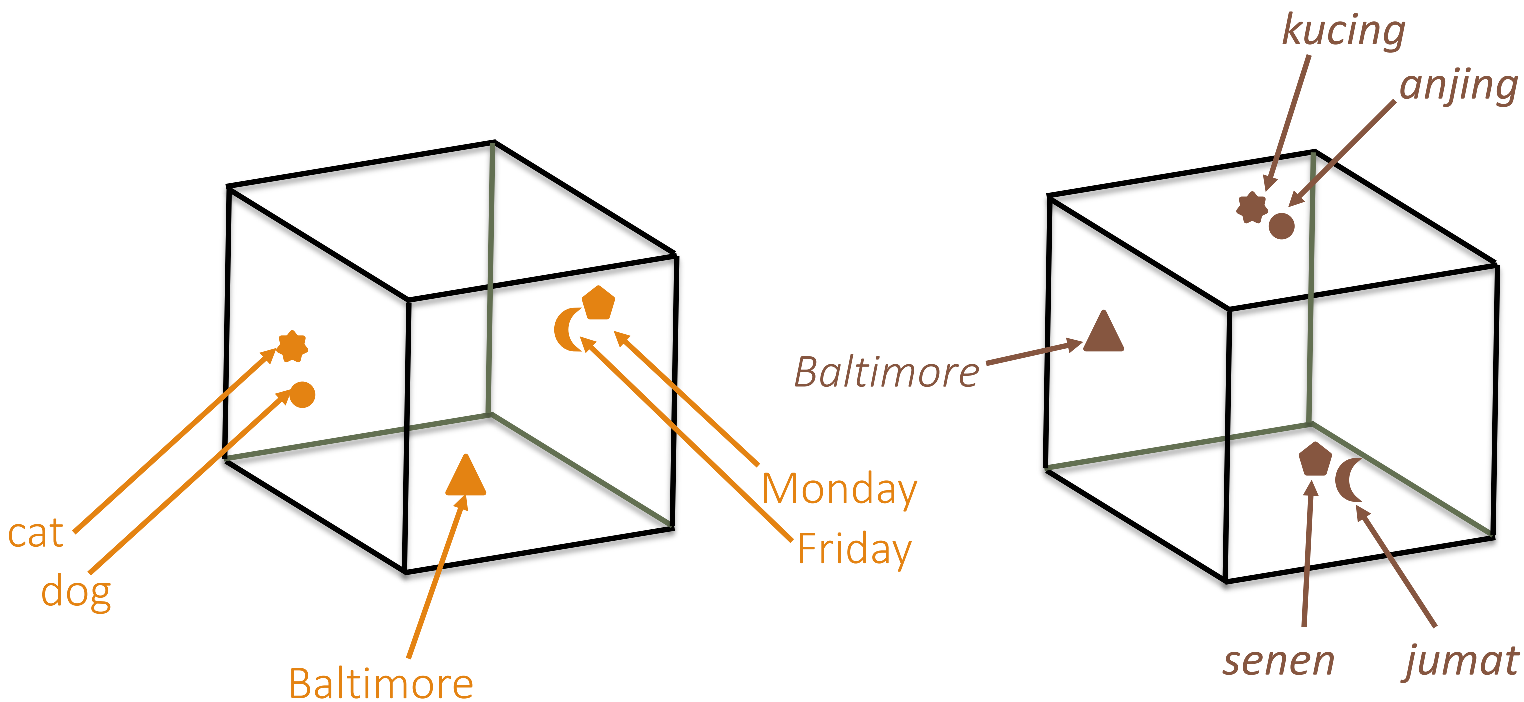
Magnitude: Python Toolkit for Manipulating Embeddings

Plasticity

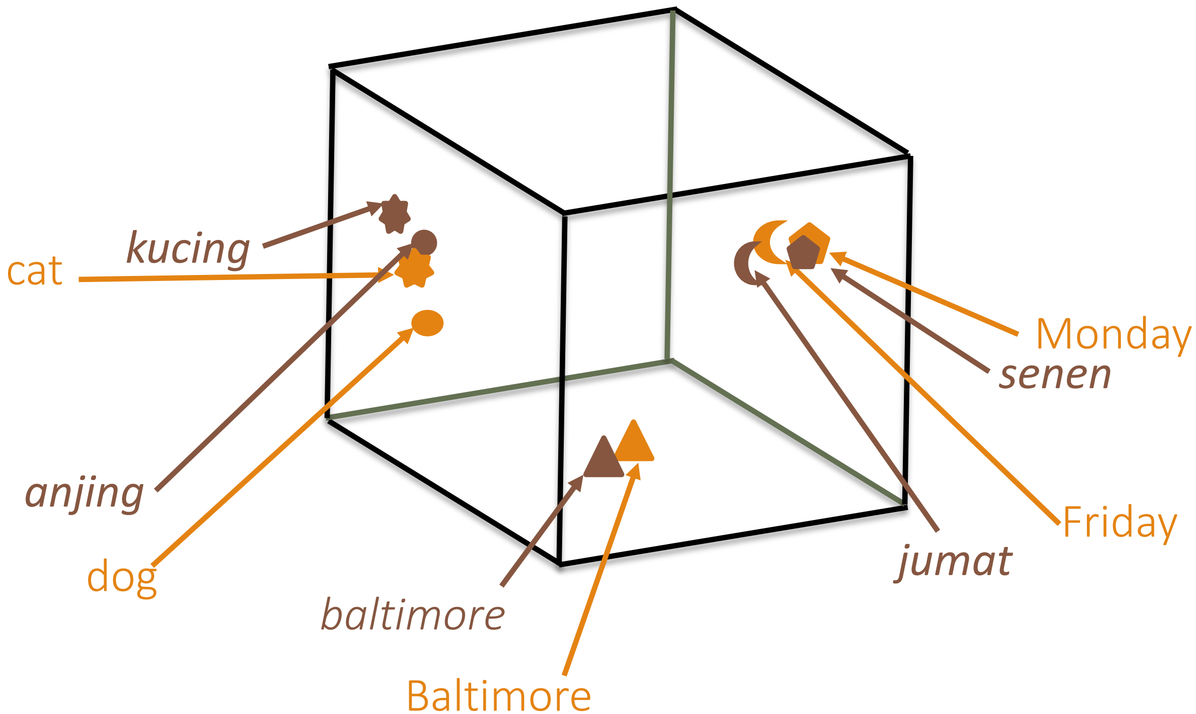
Monolingual Word Embeddings



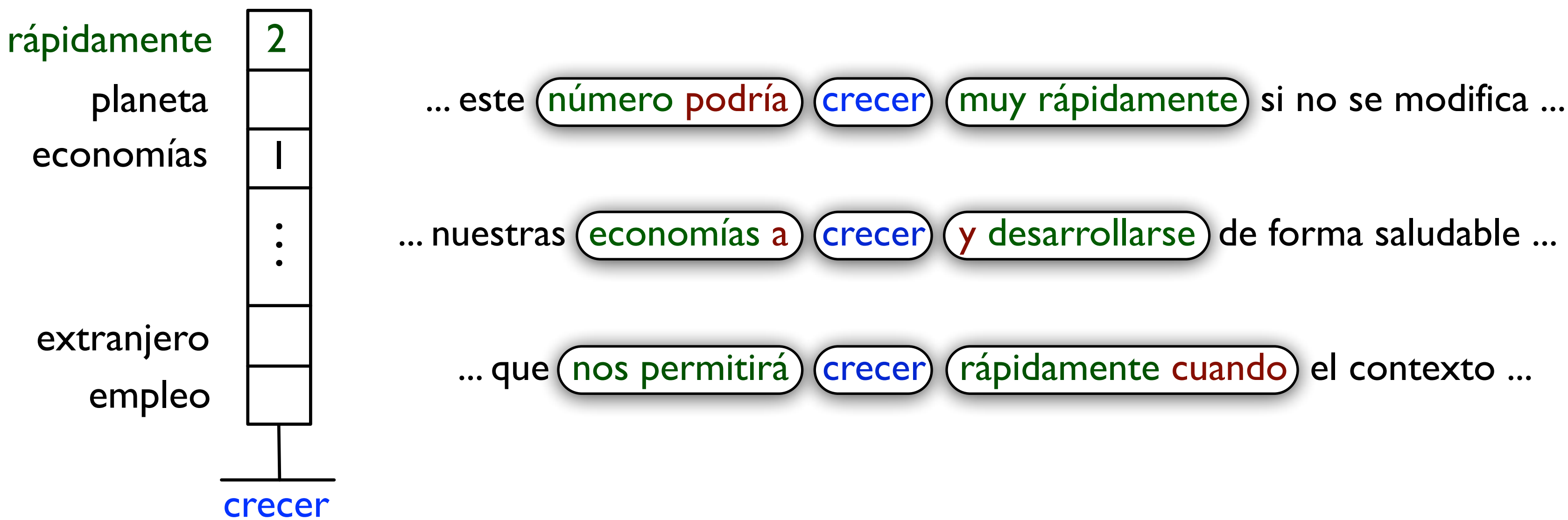
Monolingual Word Embeddings



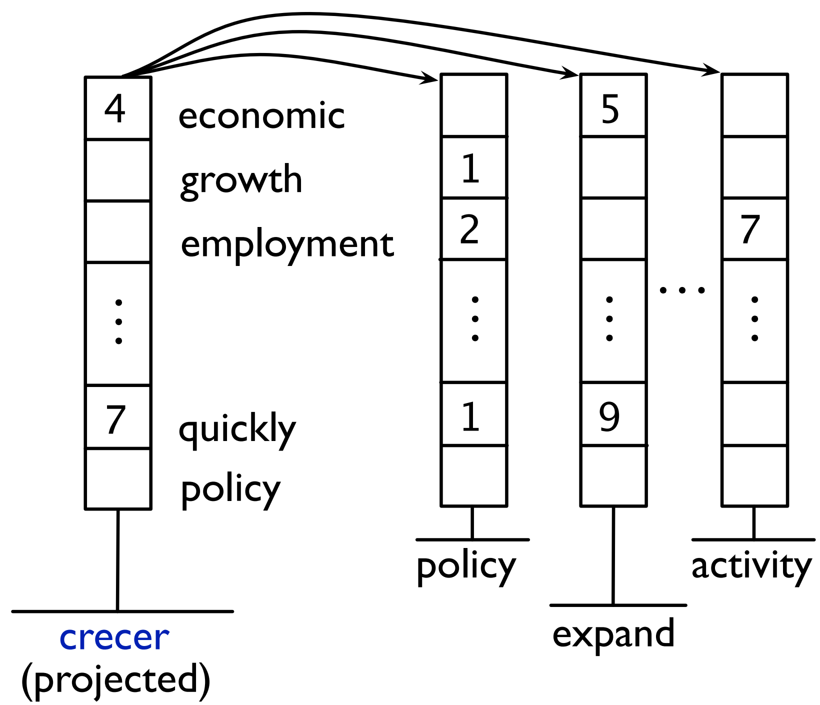
Bilingual Word Embeddings



Projecting Vector Space Models

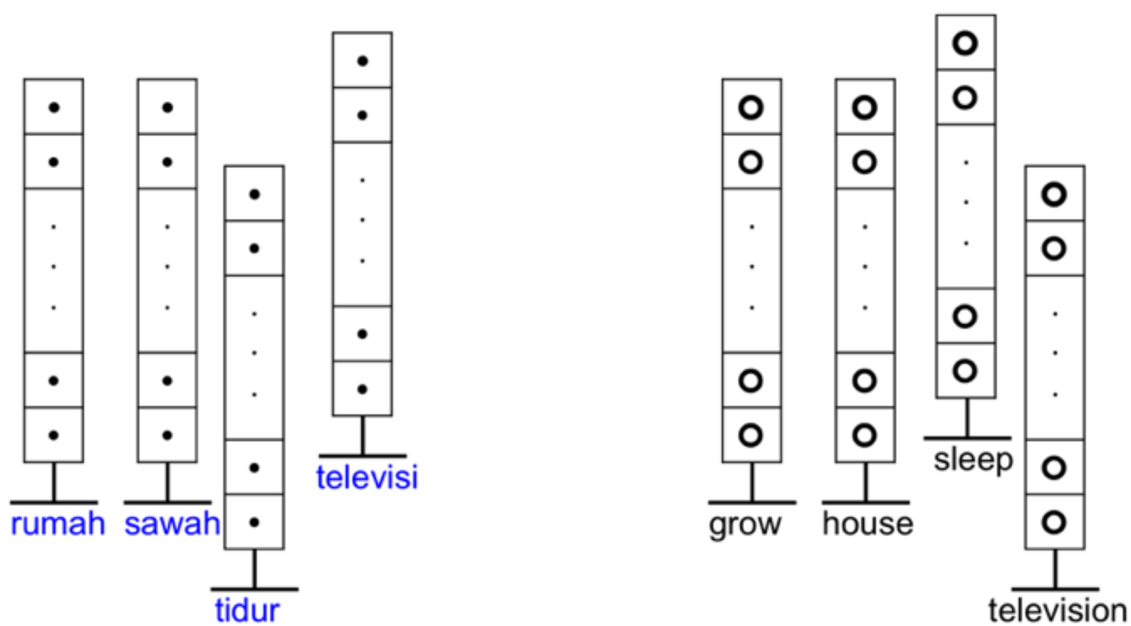


Projecting Vector Space Models



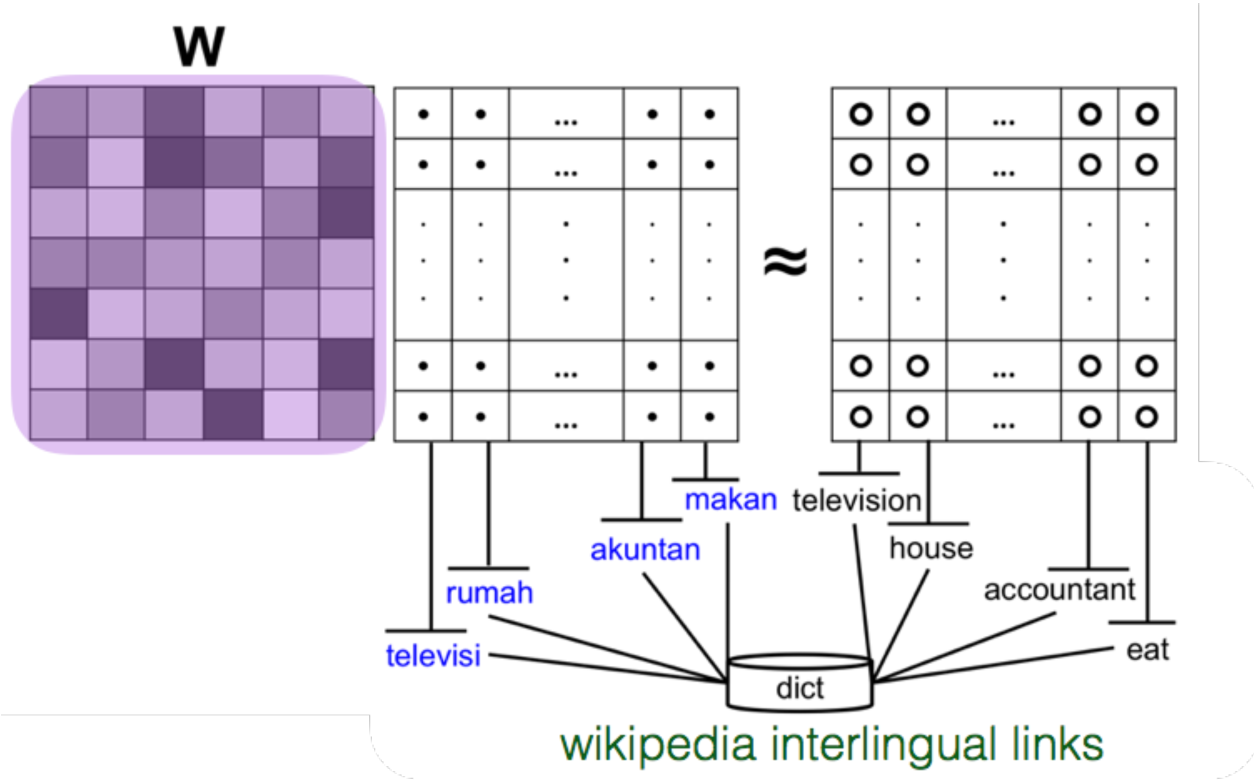
Word Embeddings

Instead of high dimensional vector space models used by Rapp and others in the past, we use low-dimensional word embeddings.

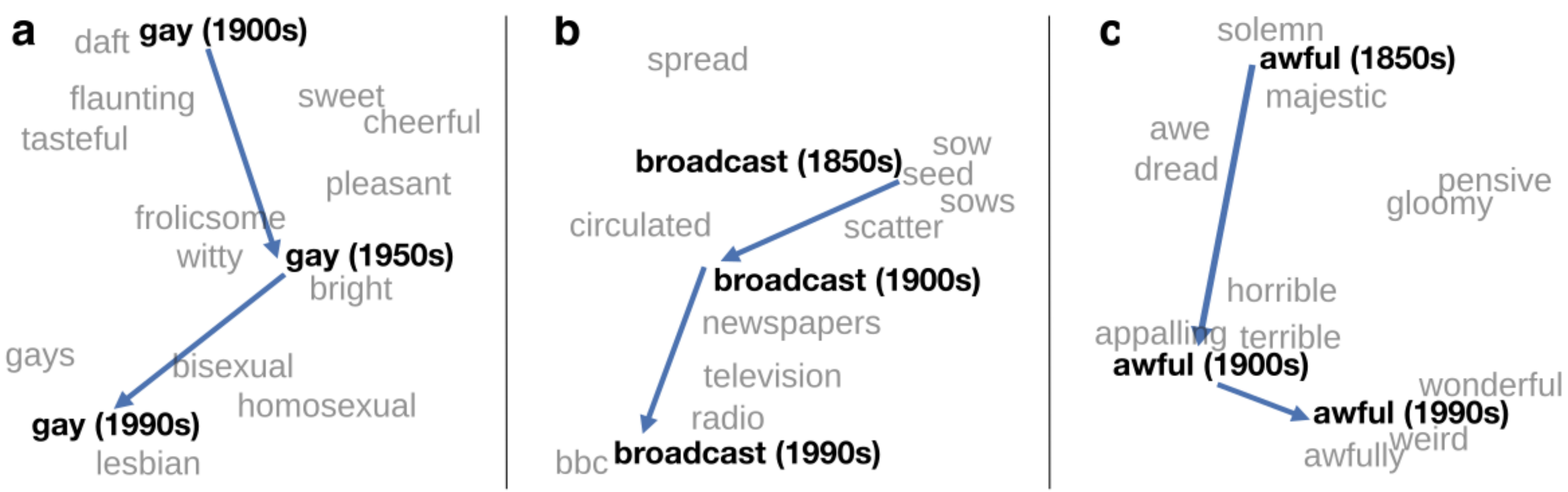


Learning Bilingual Embeddings

mapping function W



Use in Historical Linguistics



~30 million books, 1850-1990, Google Books data

gay | gā |

adjective (**gayer, gayest**)

1 (of a person) homosexual (used especially of a man): *that friend of yours, is he gay?*
- relating to or used by homosexuals: *a gay bar* | *the gay vote can decide an election.*

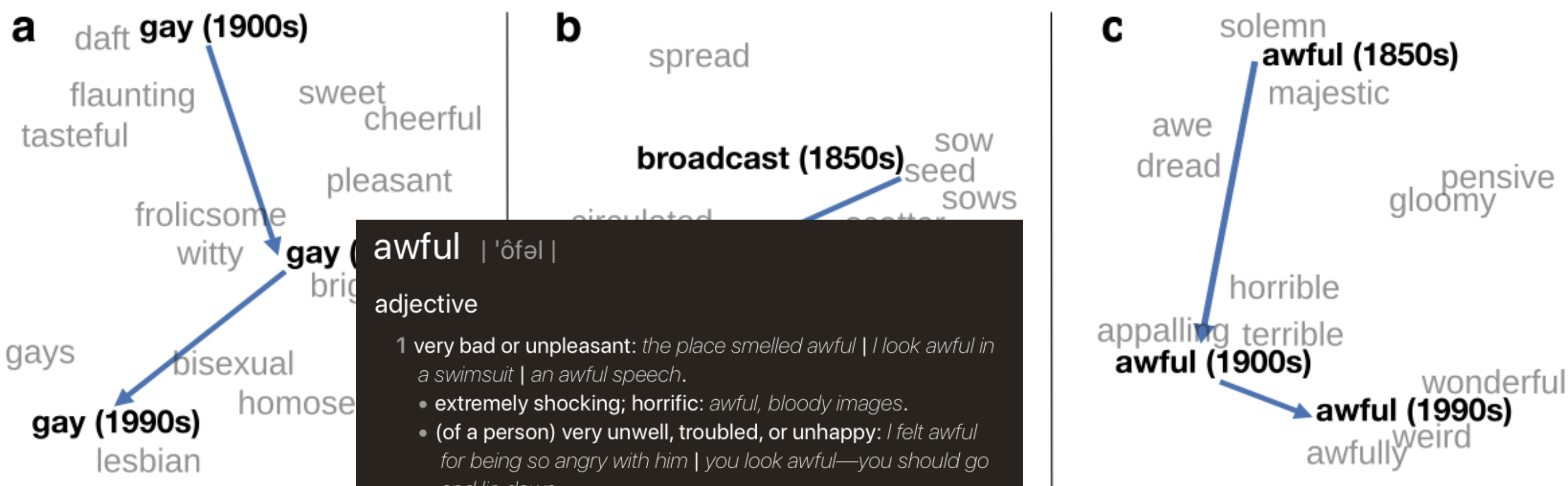
2 *dated* lighthearted and carefree: *Nan had a gay disposition and a very pretty face.*
- brightly colored; showy; brilliant: *a gay profusion of purple and pink sweet peas.*

broadcast | 'brôdˌkast |

verb (past and past participle **broadcast**) *[with object]*

1 transmit (a program or some information) by radio or television: *the announcement was broadcast live* | (as noun **broadcasting**) *: the 1920s saw the dawn of broadcasting.*
- *[no object]* take part in a radio or television transmission: *the station broadcasts 24 hours a day.*
- tell (something) to many people; make widely known: *we don't want to broadcast our unhappiness to the world.*

2 scatter (seeds) by hand or machine rather than placing in drills or rows.



awful | 'ôfəl |

adjective

1 very bad or unpleasant: *the place smelled awful* | *I look awful in a swimsuit* | *an awful speech.*
- extremely shocking; horrific: *awful, bloody images.*
- (of a person) very unwell, troubled, or unhappy: *I felt awful for being so angry with him* | *you look awful—you should go and lie down.*

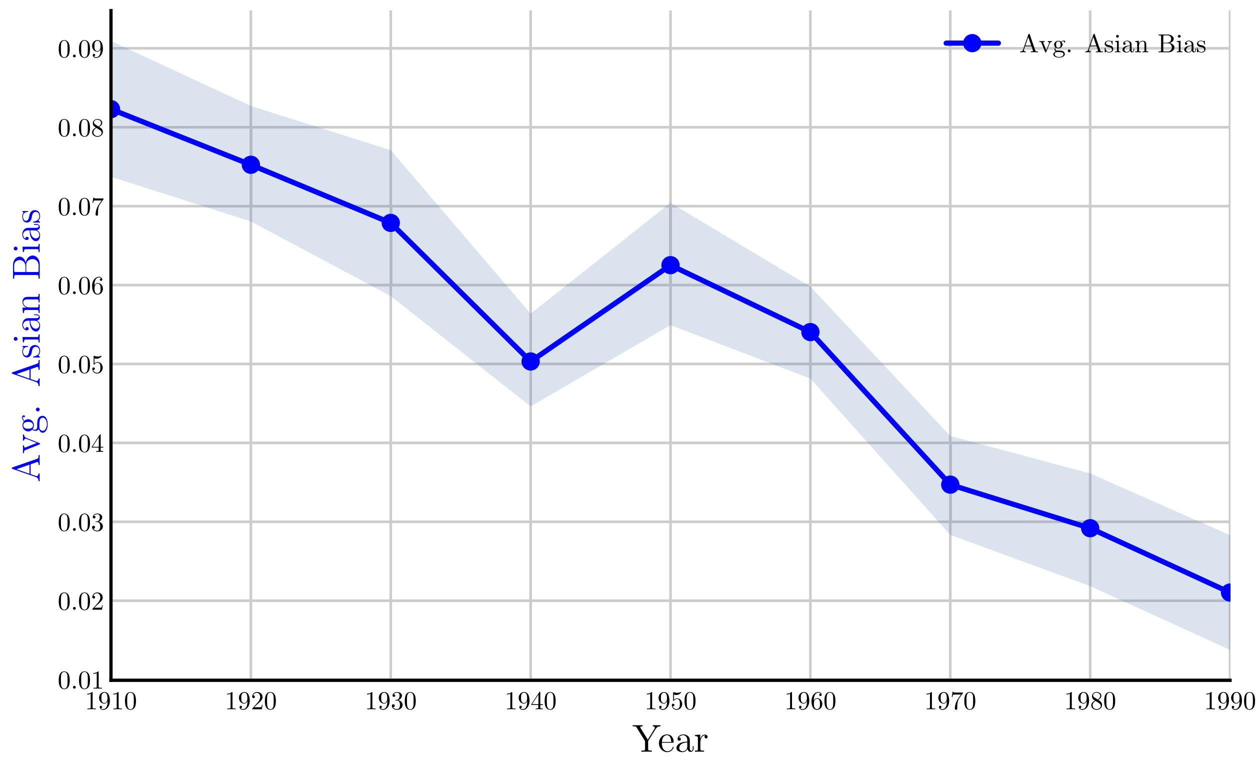
2 *[attributive]* used to emphasize the extent of something, especially something unpleasant or negative: *I've made an awful fool of myself.*

3 *archaic* inspiring reverential wonder or fear.

~30 million boo

Uses in Social Science

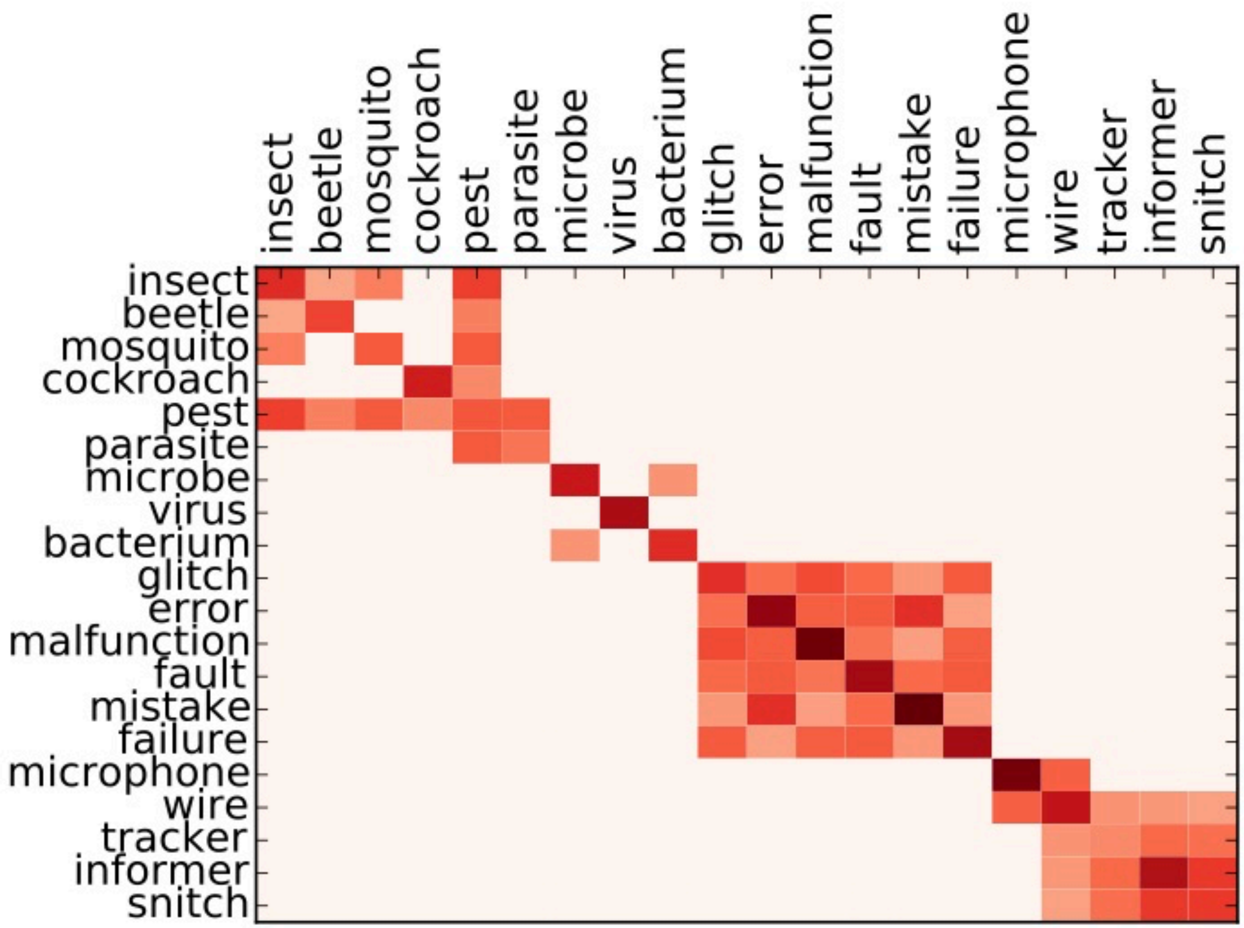
Change in association of Chinese names with adjectives framed as "othering" (*barbaric*, *monstrous*, *bizarre*)



What should a semantic model be able to do?

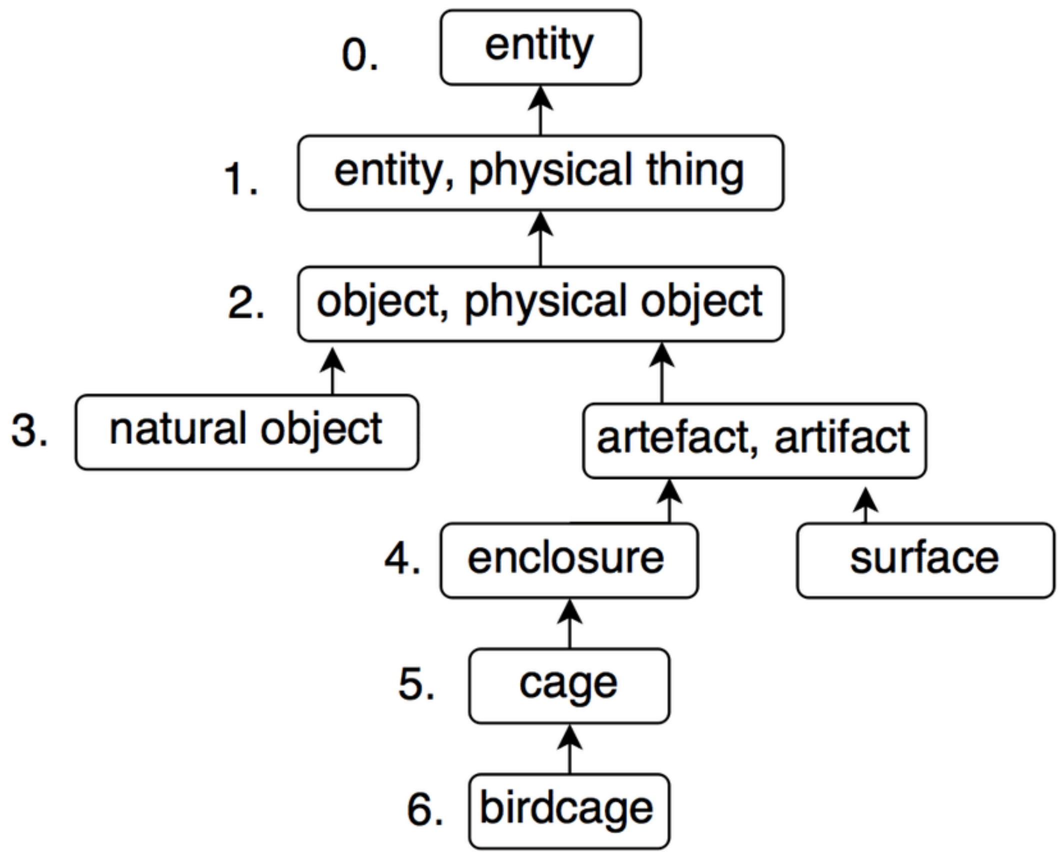
GOALS FOR DISTRIBUTIONAL SEMANTICS

Goal: Word Sense



Goal: Hypernomy

One goal of for a semantic model is to represent the relationship between words. A classic relation is *hypernomy* which describes when one word (the *hypernym)* is more general than the other word (the *hyponym*).



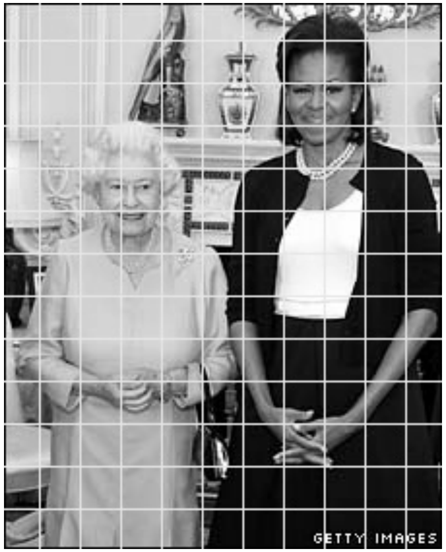
Goal: Compositionality

Language is **productive.** We can understand completely new sentences, as long as we know each word in the sentence. One goal for a semantic model is to be able to **derive** the meaning of a sentence from its parts, so that we can generalize to new combinations. This is known as **compositionality.**

Goal: Grounding

Many experimental studies in language acquisition suggest that word meaning arises not only from exposure to the linguistic environment but also from our interaction with the physical world.

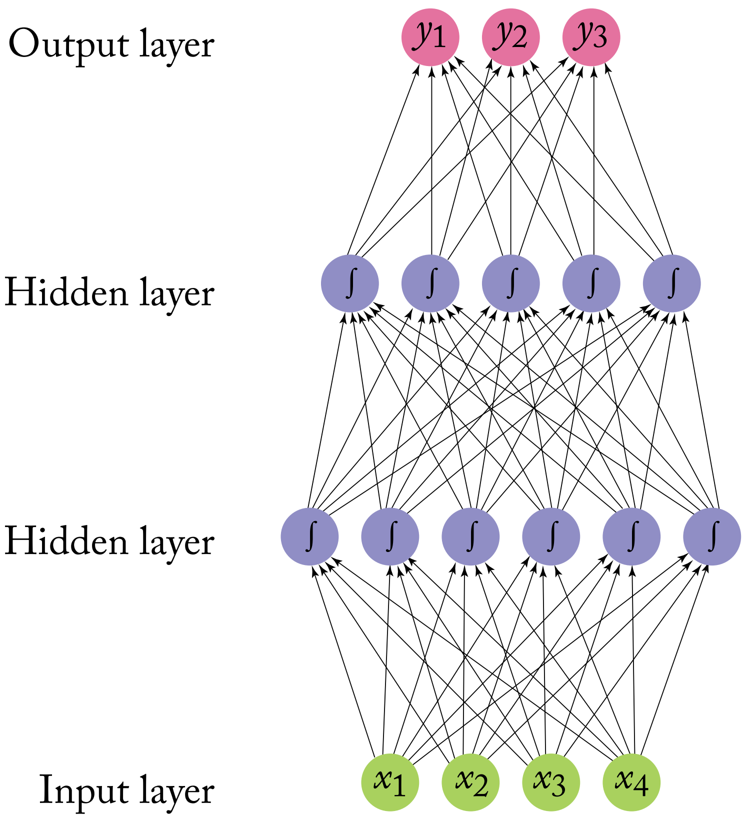
Use collections of documents that contain pictures

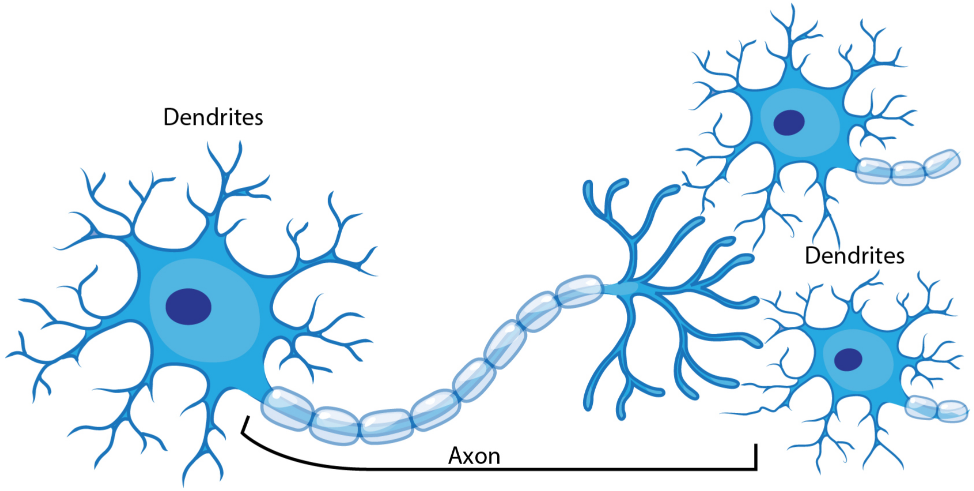


A semantic model should

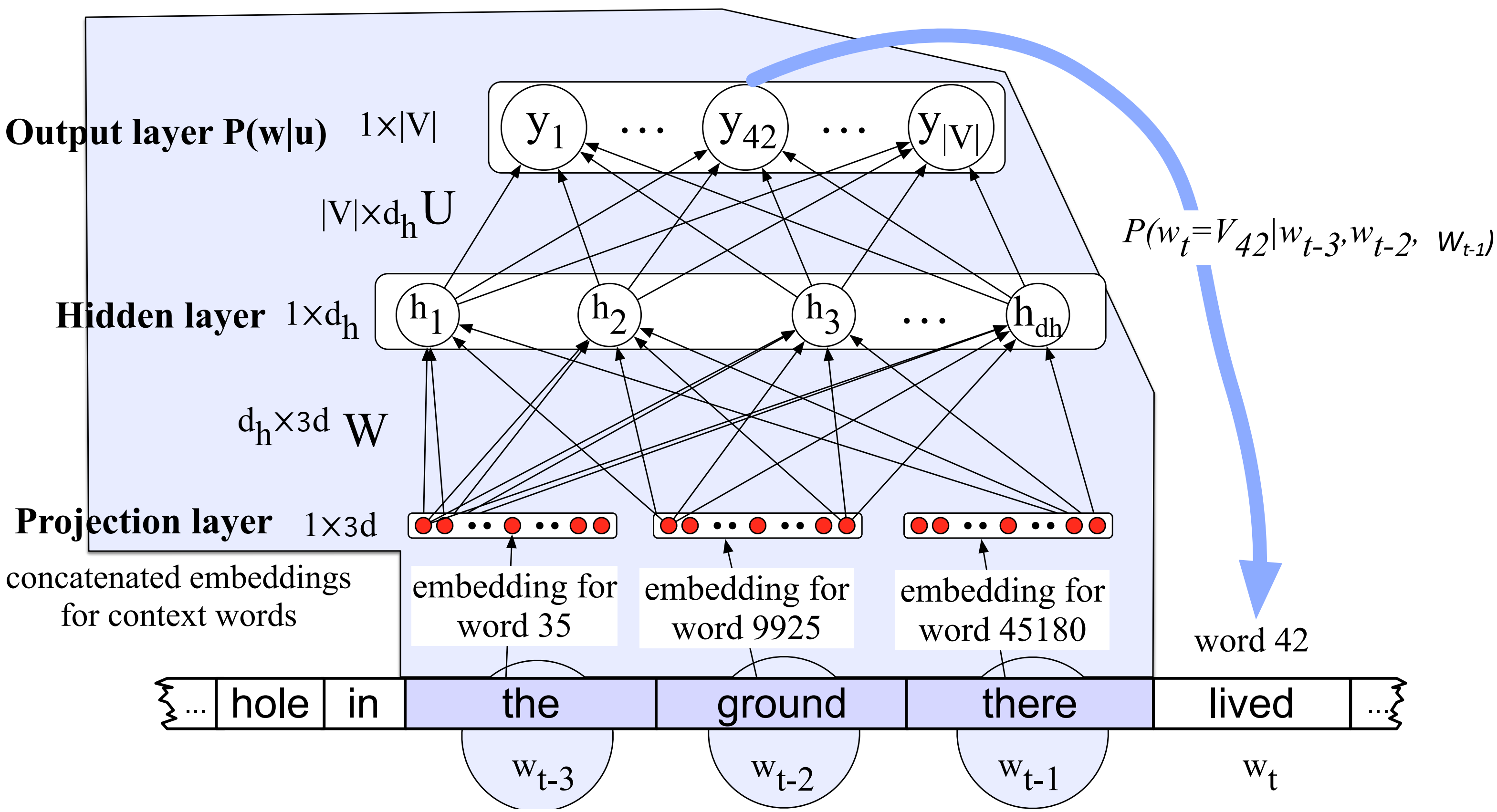
1. Handle words with multiple senses (polysemy) and encode relationships like hyponym between words/word senses
2. Robustly handle vagueness (situations when it is unclear whether an entity is a referent of a concept)
3. Should be able to be combined word representations to encode the meanings of sentences (compositionally)
4. Capture how word meaning depends on context.
5. Support logical notions of truth and entailment
6. Generalize to new situations (connecting concepts and referents)
7. Capture how language relates to the world via sensory perception (grounding)

Neural networks

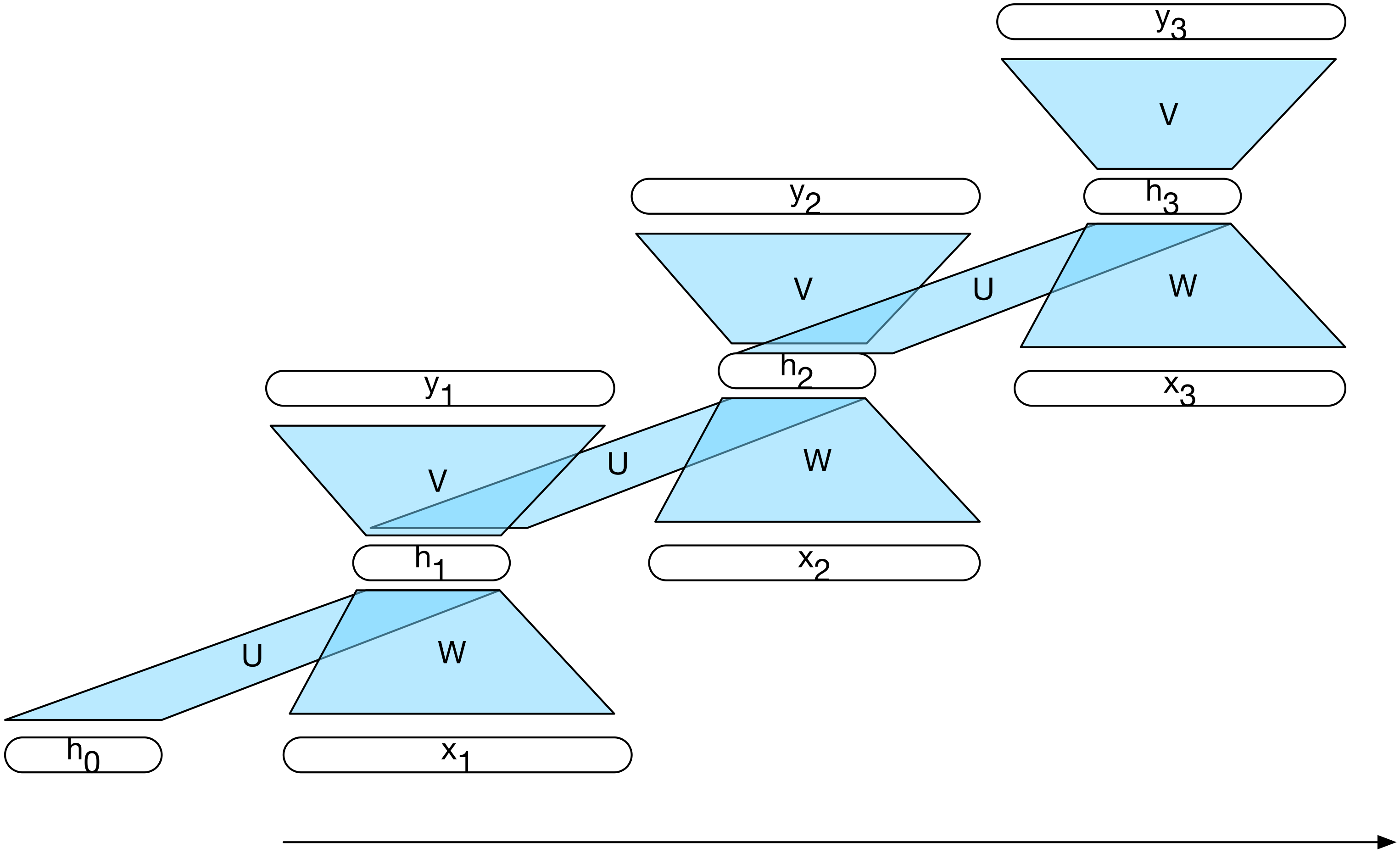




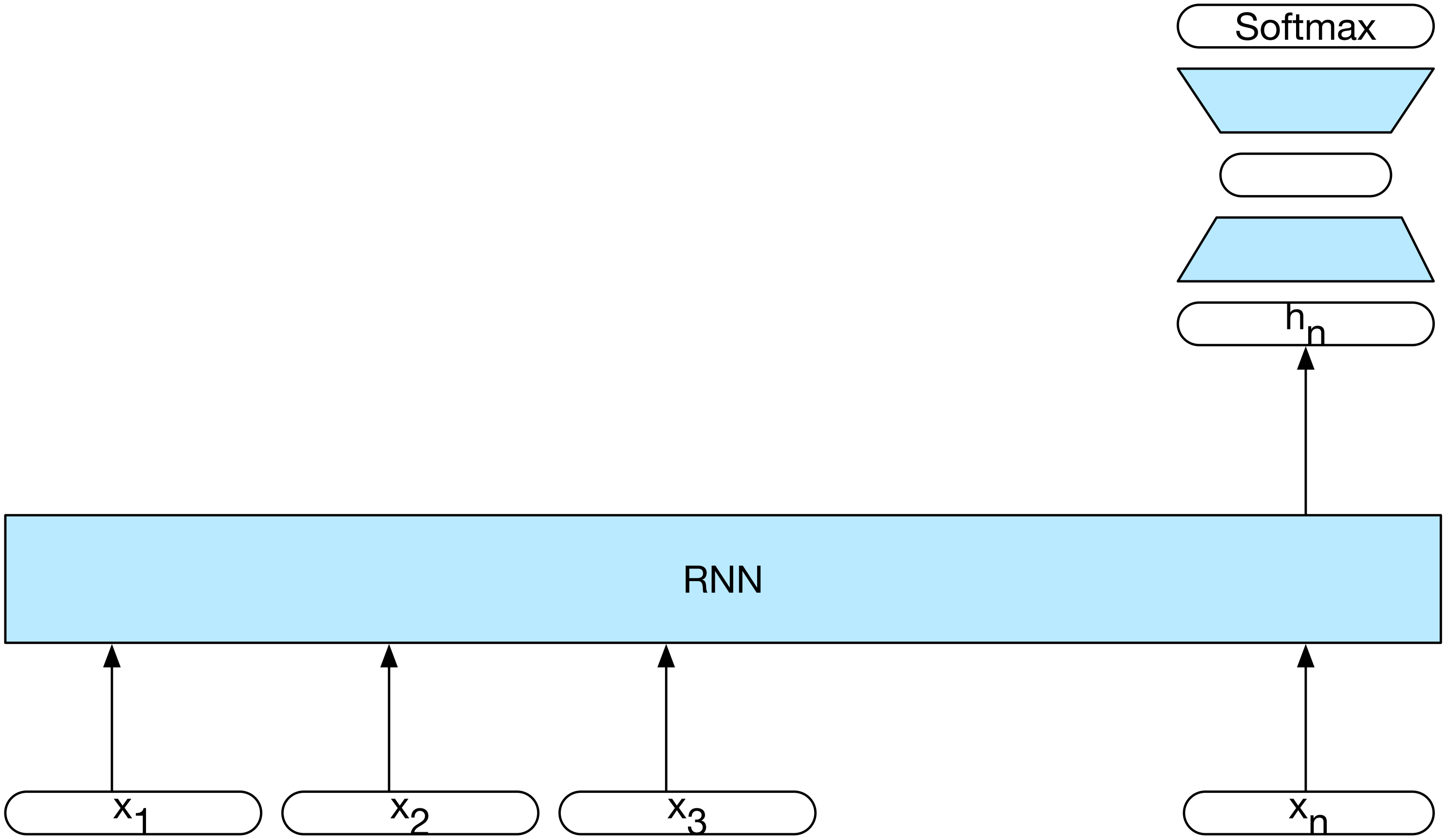
Neural network LMs



Recurrent Neural Networks

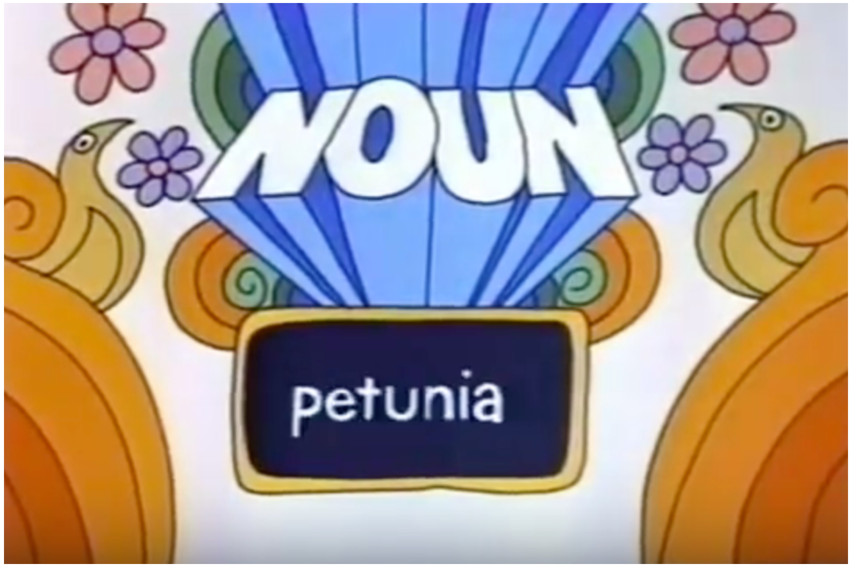


Sequence Classifiers



Sequence Models

A **sequence model** or **sequence classifier** is a model whose job is to assign a label or class to each unit in a sequence, thus mapping a sequence of observations to a sequence of labels.



Noun	Adverb
Verb	Conjunction
Pronoun	Adjective
Preposition	Interjection

Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	**SYM**	symbol	*+, %, &*
CD	cardinal number	*one, two*	**TO**	“to”	*to*
DT	determiner	*a, the*	**UH**	interjection	*ah, oops*
EX	existential “there”	*there*	**VB**	verb base form	*eat*
FW	foreign word	*mea culpa*	**VBD**	verb past tense	*ate*
IN	proposition/sub-conj	*of, in, by*	**VBG**	verb gerund	*eating*
JJ	adjective	*yellow*	**VBN**	verb past participle	*eaten*
JJR	comparative adjective	*bigger*	**VBP**	verb non-3sg pres	*eat*
JJS	superlative adjective	*wildest*	**VBZ**	verb 3sg pres	*eats*
LS	list item marker	*1, 2, One*	**WDT**	wh-determiner	*which, that*
MD	modal	*can, should*	**WP**	wh-pronoun	*what, who*
NN	noun, singular or mass	*llama*	**WP$**	possessive wh-	*whose*
NNS	noun, plural	*llamas*	**WRB**	wh-adverb	*how, where*
NNP	proper noun, sing.	*IBM*	**$**	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	**#**	pound sign	*#*
PDT	predeterminer	*all, both*	**“**	left quote	*‘ or “*
POS	possessive ending	*‘s*	**”**	right quote	*’ or ”*
PRP	personal pronoun	*I, you, we*	**(**	left parenthesis	*[, (, {, <*
PRP$	possessive pronoun	*your, one’s*	**)**	right parenthesis	*],), }, >*

POS Tagging

Words are ambiguous, so tagging must resolve disambiguate.

Types:	WSJ	Brown
Unambiguous (1 tag)	44,432 (**86%**)	45,799 (**85%**)
Ambiguous (2+ tags)	7,025 (**14%**)	8,050 (**15%**)
Tokens:		
Unambiguous (1 tag)	577,421 (**45%**)	384,349 (**33%**)
Ambiguous (2+ tags)	711,780 (**55%**)	786,646 (**67%**)

The amount of tag ambiguity for word types in the Brown and WSJ corpora from the Treebank-3 (45-tag) tagging. These statistics include punctuation as words, and assume words are kept in their original case.

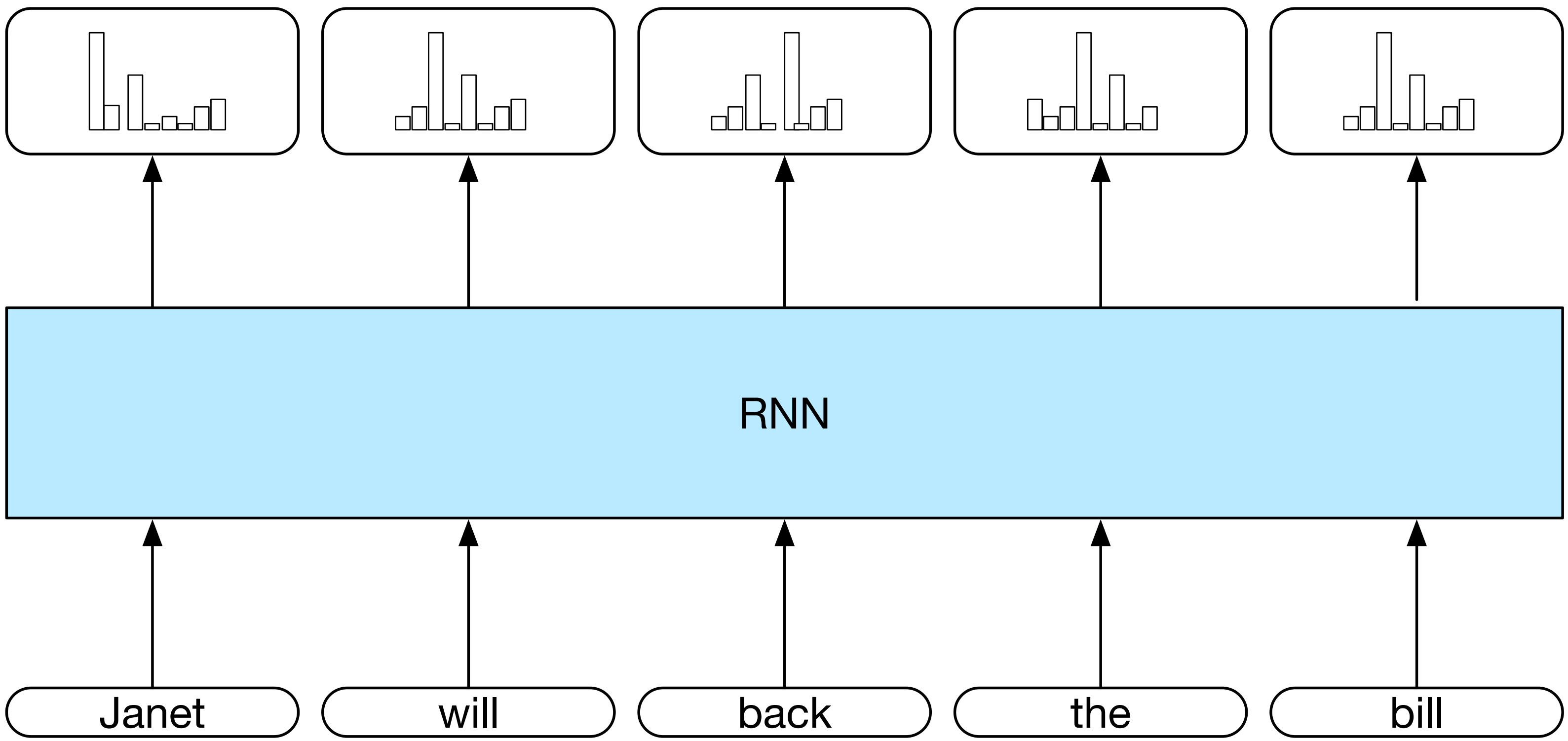
Most frequent class baseline

Many words are easy to disambiguate, because their different tags aren't equally likely.

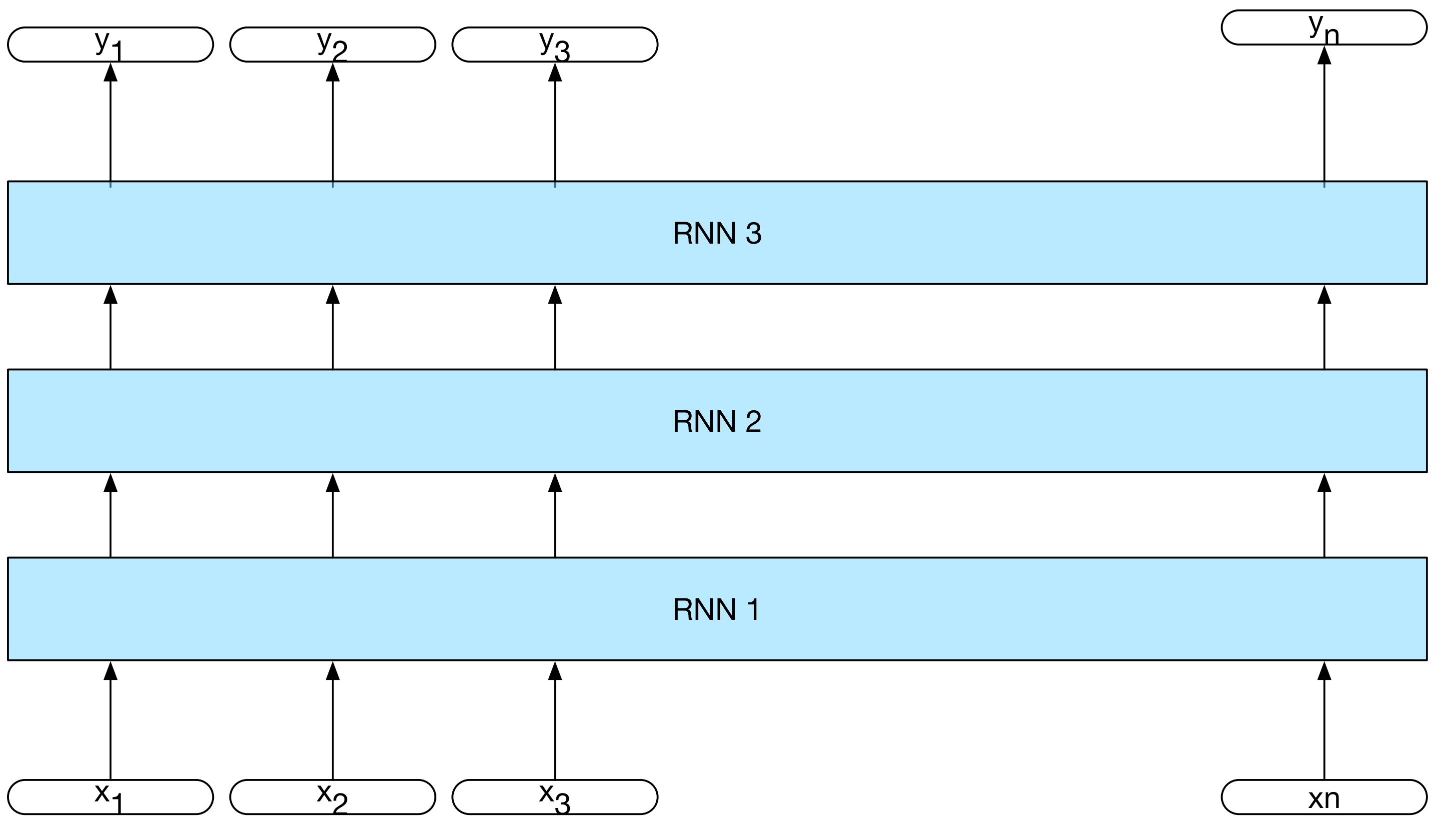
Simplistic baseline for POS tagging: given an ambiguous word, choose the tag which is most frequent in the training corpus.

Most Frequent Class Baseline: Always compare a classifier against a baseline at least as good as the most frequent class baseline (assigning each token to the class it occurred in most often in the training set).

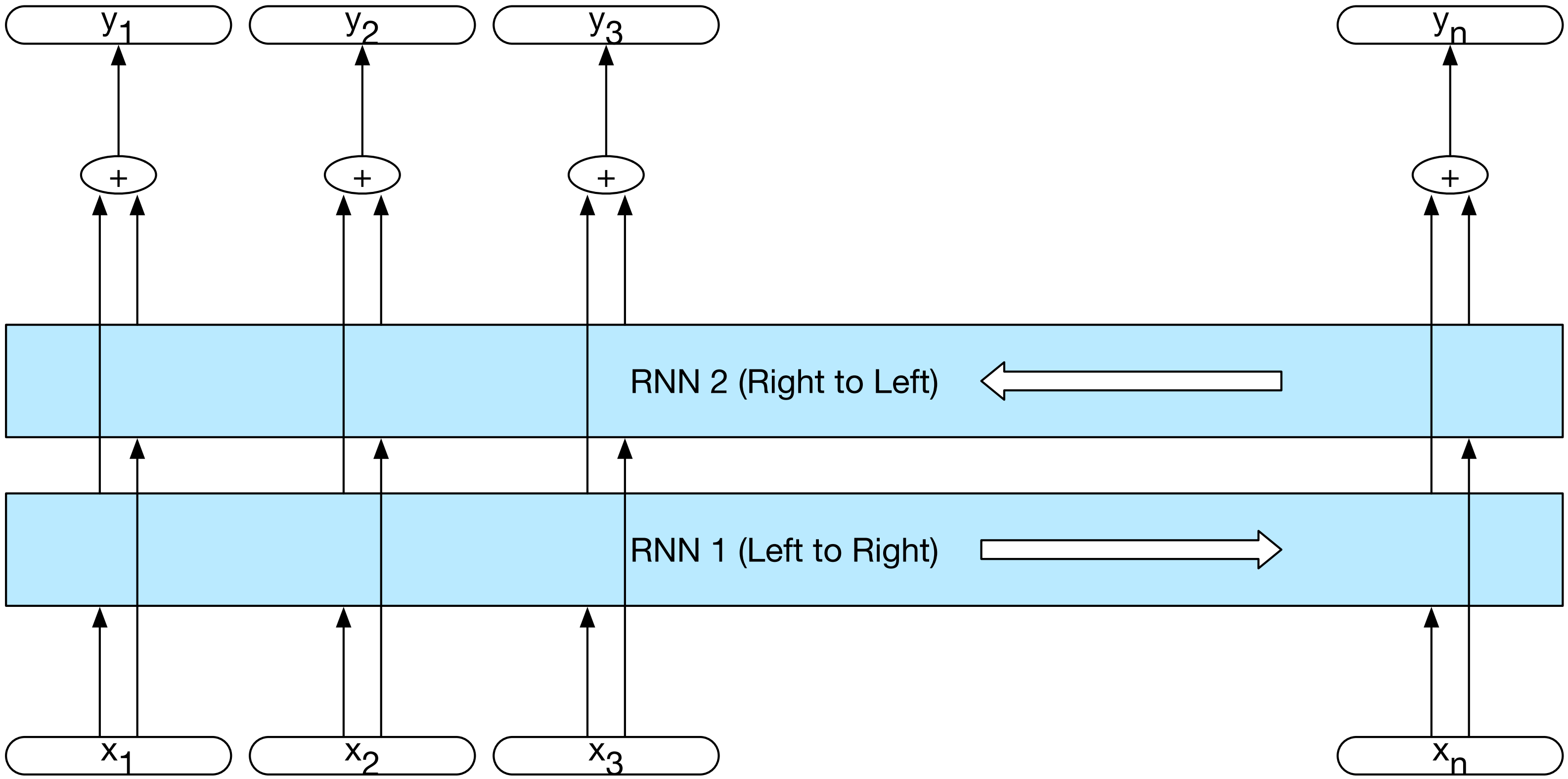
Tag Sequences



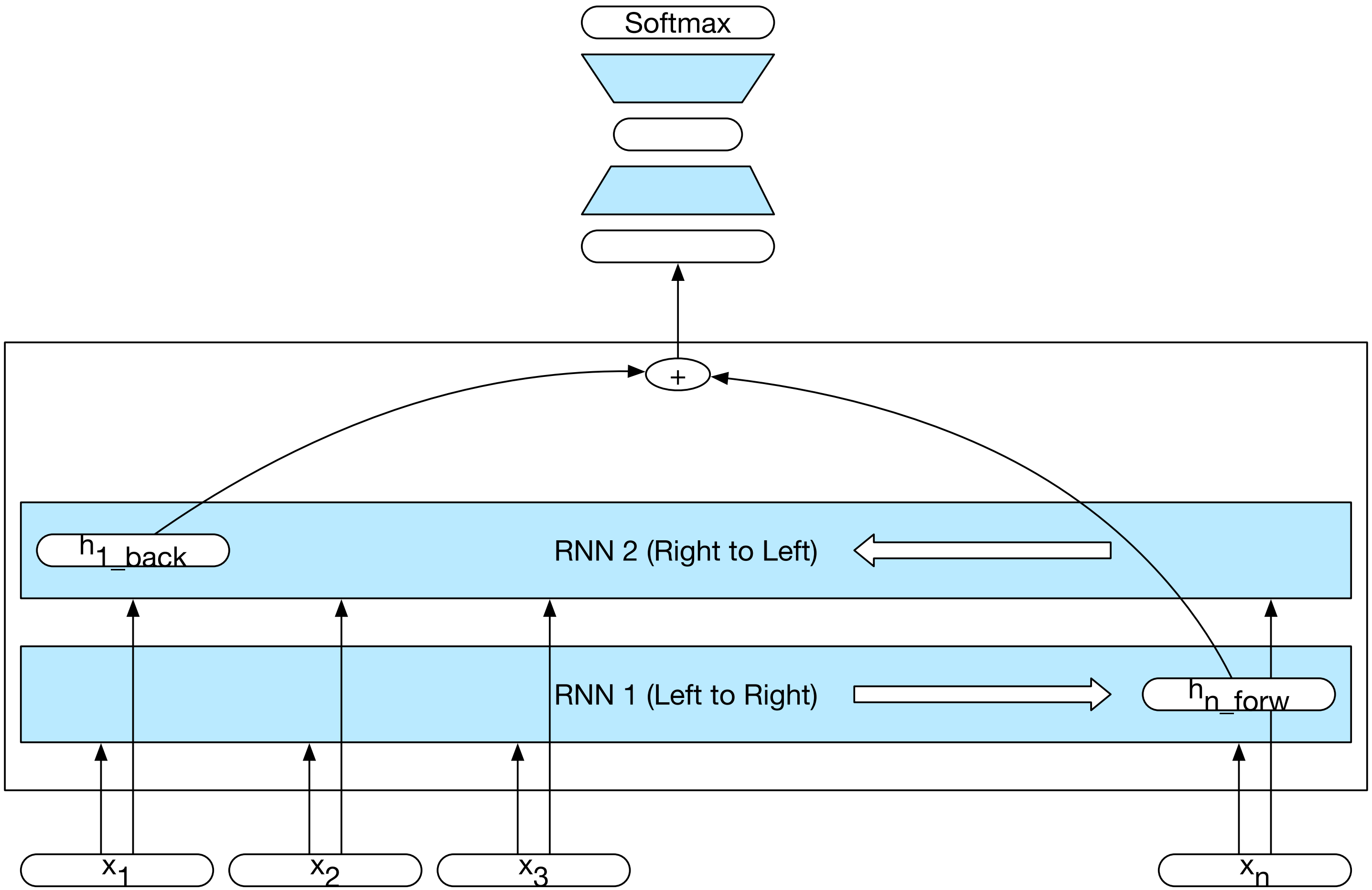
Stacked RNNs



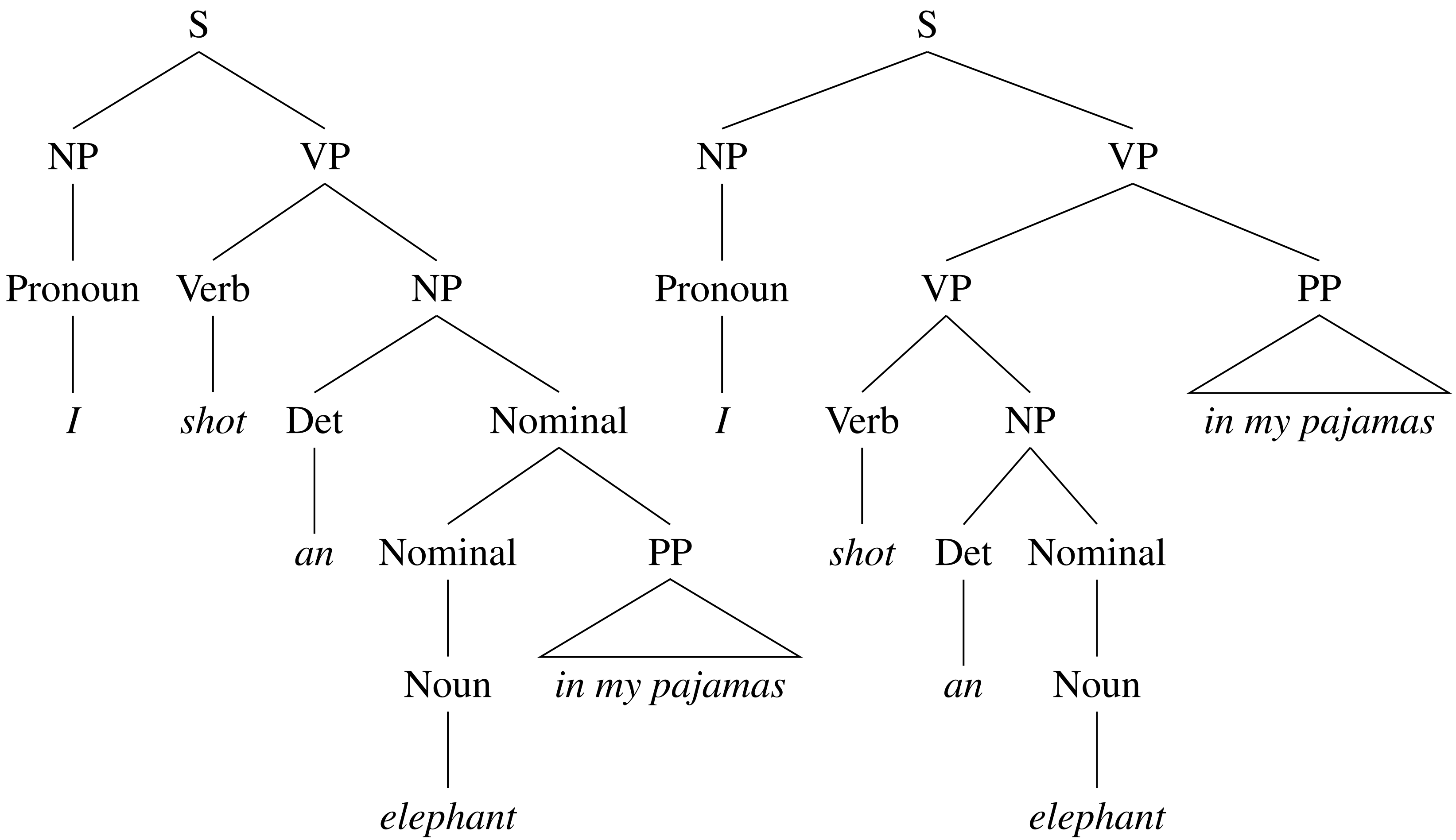
Bidirectional RNNs



Bidirectional RNNs for sequence classification



Syntactic Parsing



Ambiguity

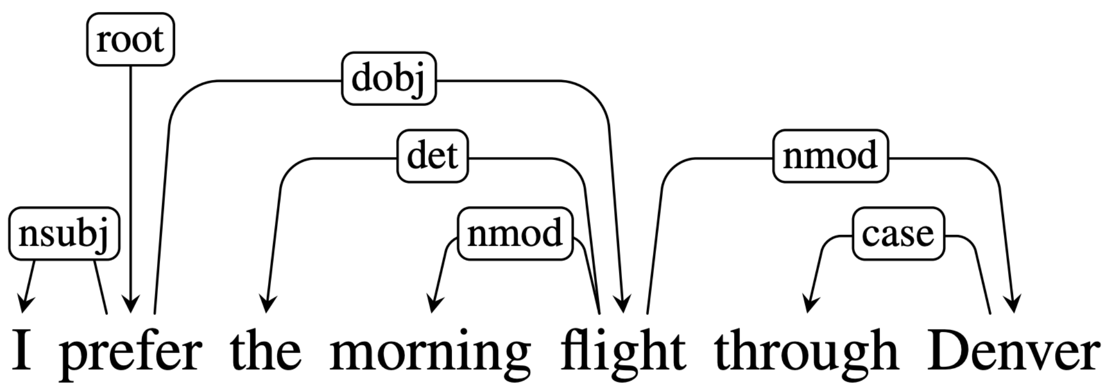
Ambiguity can arise because of words with **multiple senses or POS tags**. Many kinds of ambiguity are also structural.

"One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know."

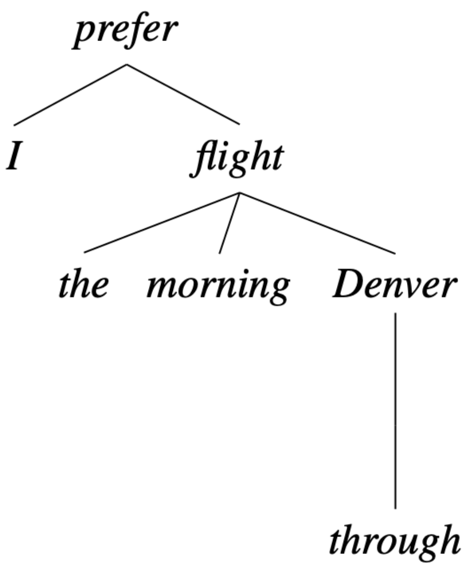
~Groucho Marx
American comedian
1890-1977

Dependency Grammars

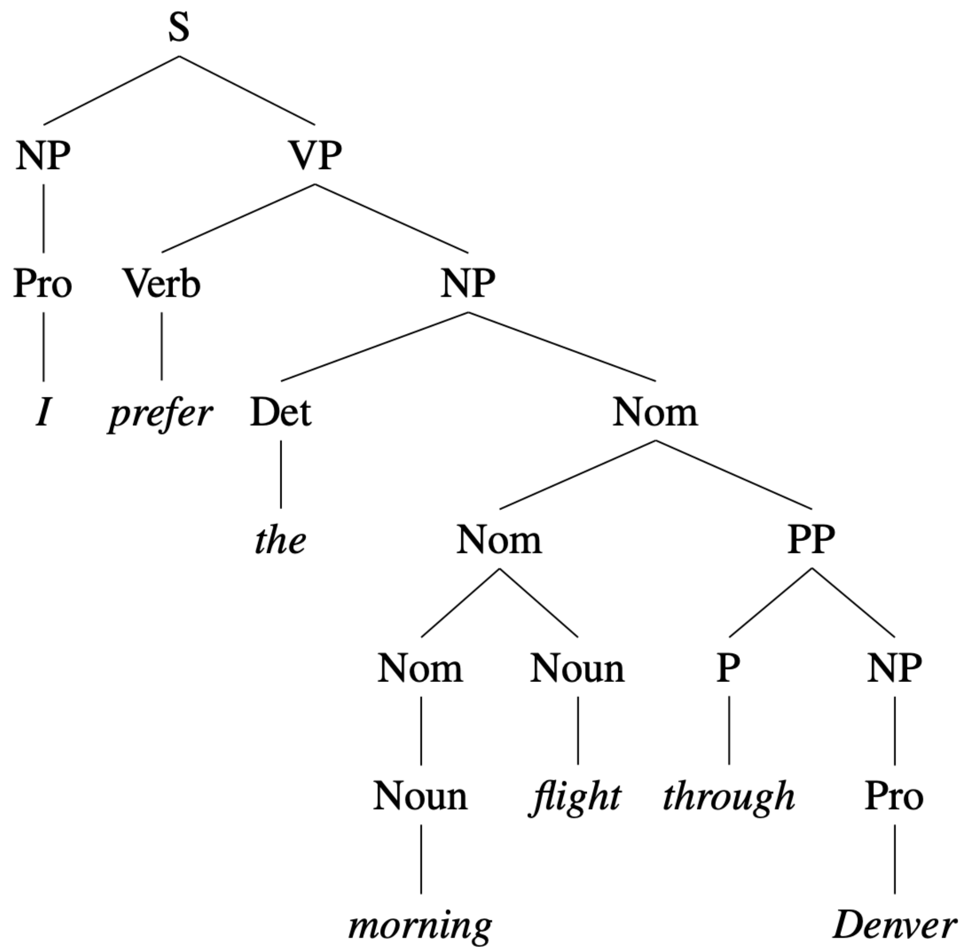
Dependency grammars depict the syntactic structure of sentences solely in terms of the **words in a sentence** and an **associated set of directed head-dependent grammatical relations** that hold among these words.



Dependency – based



Constituent– based



Open Information Extraction

Unsupervised relation extraction

Find all strings of words that satisfy the tripe relation.

United has a hub in Chicago, which is the headquarters of United Continental Holdings.

r1: <United, has a hub in, Chicago>

r2: <Chicago, is the headquarters of, United Continental Holdings>

Template Filling

Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

FARE-RAISE ATTEMPT:	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

Temporal Expression Extraction

Absolute	Relative	Durations
April 24, 1916	yesterday	four hours
The summer of ’77	next semester	three weeks
10:15 AM	two weeks from yesterday	six days
The 3rd quarter of 2006	last quarter	the last three quarters

Lexical triggers for temporal expressions:

Category	Examples
Noun	*morning, noon, night, winter, dusk, dawn*
Proper Noun	*January, Monday, Ides, Easter, Rosh Hashana, Ramadan, Tet*
Adjective	*recent, past, annual, former*
Adverb	*hourly, daily, monthly, yearly*

- Temporal expression recognition
- Temporal normalization
  - mapping a temporal expression to either normalization a specific point in time or to a duration

Event Extraction

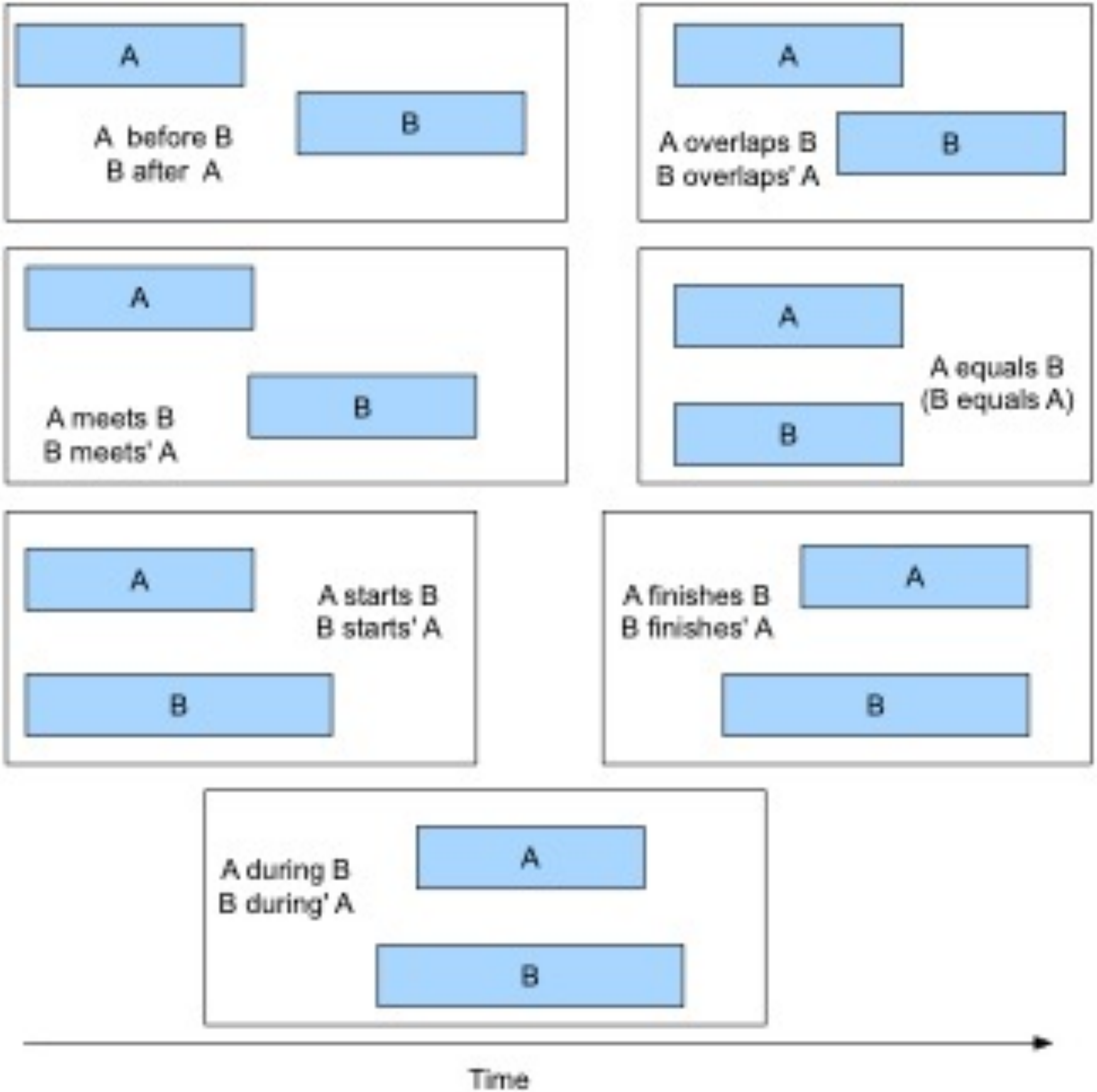
[$_{\text{EVENT}}$ Citing] high fuel prices, United Airlines [$_{\text{EVENT}}$ said] Friday it has [$_{\text{EVENT}}$ increased] fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately [$_{\text{EVENT}}$ matched] [$_{\text{EVENT}}$ the move], spokesman Tim Wagner [$_{\text{EVENT}}$ said]. United, a unit of UAL Corp., [$_{\text{EVENT}}$ said] [$_{\text{EVENT}}$ the increase] took effect Thursday and [$_{\text{EVENT}}$ applies] to most routes where it [$_{\text{EVENT}}$ competes] against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

Events can be classified as **actions**, **states**, **reporting events, perception events**, etc. The aspect, tense, and modality of each event also needs to be extracted.

Temporal ordering of events

Delta Air Lines earnings soared 33% to a record in the fiscal first quarter, bucking the industry trend toward declining profits.

- Soaring$_{e1}$ is **included in** the fiscal first quarter$_{t58}$
- Soaring$_{e1}$ is **simultaneous with** the bucking$_{e3}$
- Declining$_{e4}$ **includes** soaring$_{e1}$

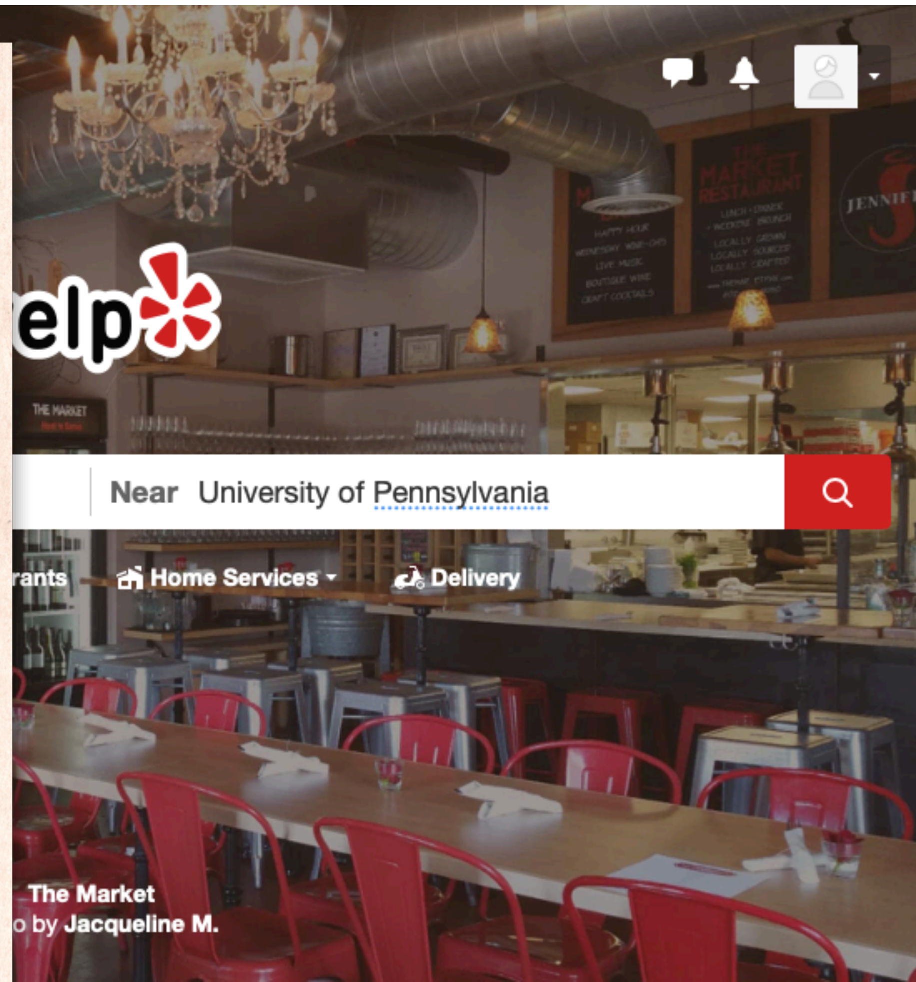

A
B
A before B
B after A
A
B
A overlaps B
B overlaps' A
A
B
A meets B
B meets' A
A
B
A equals B
(B equals A)
A
B
A starts B
B starts' A
A
B
A finishes B
B finishes' A
A
B
A during B
B during' A
Time

Desirable Properties for Meaning Representations

1. **Verifiability** – compare some meaning representation (MR) to a representation in a knowledge base (KB).
2. **Unambiguous Representations** – each ambiguous natural language meaning corresponds to a separate MR
3. **Canonical Forms** – paraphrases are collapsed to one MR
4. **Make Inferences** – draw valid conclusions based on the MR of inputs and its background knowledge in KB
5. **Match variables** – variables can be replaced by some object in the KB so an entire proposition will then match

Unambiguous representation

I want to eat someplace that's near Penn's campus.



Model-Theoretic Semantics

A **model** allows us to bridge the gap between a formal representation and the world. The model stands in for a particular state of affairs in the world.

The **domain** of a model is the set of objects that are being represented. Each distinct thing (*person, restaurant, cuisine*) corresponds to a unique element in the domain

Properties of objects (like whether a restaurant is *expensive*) in a model correspond to sets of objects.

Relations between object (like whether a restaurant *serves* a cuisine) are are sets of tuples.

Domain $\mathscr{D} = \{a, b, c, d, e, f, g, h, i, j\}$

Matthew, Franco, Katie and Caroline a, b, c, d

Frasca, Med, Rio e, f, g

Italian, Mexican, Eclectic h, i, j

Properties

Noisy $Noisy = \{e, f, g\}$

Frasca, Med, and Rio are noisy

Relations

Likes $Likes = \{\langle a, f\rangle, \langle c, f\rangle, \langle c, g\rangle, \langle b, e\rangle, \langle d, f\rangle, \langle d, g\rangle\}$

Matthew likes the Med

Katie likes the Med and Rio

Franco likes Frasca

Caroline likes the Med and Rio

Serves $Serves = \{\langle f, j\rangle, \langle g, i\rangle, \langle e, h\rangle\}$

Med serves eclectic

Rio serves Mexican

Frasca serves Italian

Domain $\mathscr{D} = \{a, b, c, d, e, f, g, h, i, j\}$

Matthew, Franco, Katie and Caroline a, b, c, d

Frasca, Med, Rio e, f, g

Italian, Mexican, Eclectic

Properties

Noisy

Frasca, Med, and Rio are noisy

Relations

Likes $Likes = \{\langle a, f\rangle, \langle c, f\rangle, \langle c, g\rangle, \langle b, e\rangle, \langle d, f\rangle, \langle d, g\rangle\}$

Matthew likes the Med

Katie likes the Med and Rio

Franco likes Frasca

Caroline likes the Med and Rio

Serves

Med serves eclectic

Rio serves Mexican

Frasca serves Italian

Katie likes Rio

Katie → c

Rio → g

likes → Likes

<c,g> ∈ Likes

so *Katie likes Rio*

is True

First-Order Logic

FOL is a meaning representation language that satisfies the desirable qualities that we outlined. It provides a computational basis for **verifiability** and **inference**.

It doesn't have many requirements other than the represented world consists of objects, properties of objects, and relations among objects.

Logical Connectives

We can conjoin formula with logical connectives like **and (∧), or (∨), not** (¬), and **implies (⇒)**

Each one has a **truth table:**

P	Q	$P \vee Q$
False	*False*	*False*
False	*True*	*True*
True	*False*	*True*
True	*True*	*True*

Quantifiers

All restaurants in Philly are closed.

$$\forall x Restaurant(x) \land Is((LocationOf(x), Philadelphia) \Rightarrow Closed(x)$$

The ∀ operator states that for the logical formula to be true, the substitution of **any object** in the knowledge base for the **universally quantified variable** should result in a true formula.

Value of Logical Representation of Sentences

Is Barack Obama a US Citizen?

$Citizen\_Of(Barack\_Obama, United\_States)$

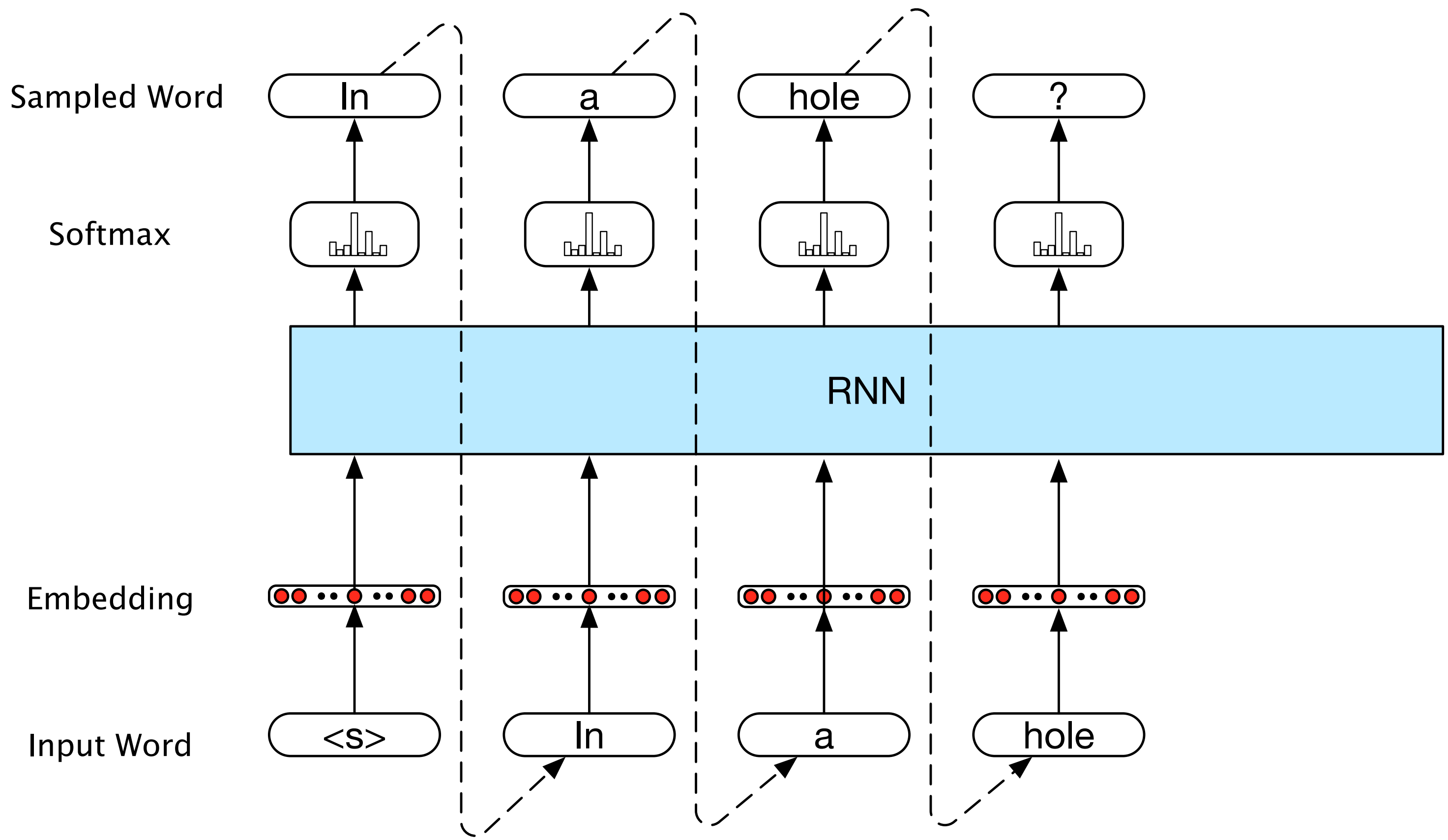
$$\forall x\ Person(x) \land Born\text{-}In(x, y) \land Located\text{-}In(y, United\_States) \Rightarrow Citizen\_Of(x, United\_States)$$

$$\frac{Person(Barack\_Obama) \land Born\text{-}In(Barack\_Obama, Hawaii) \land Located\text{-}In(Hawaii, United\_States)}{Citizen\_Of(Barack\_Obama, United\_States)}$$

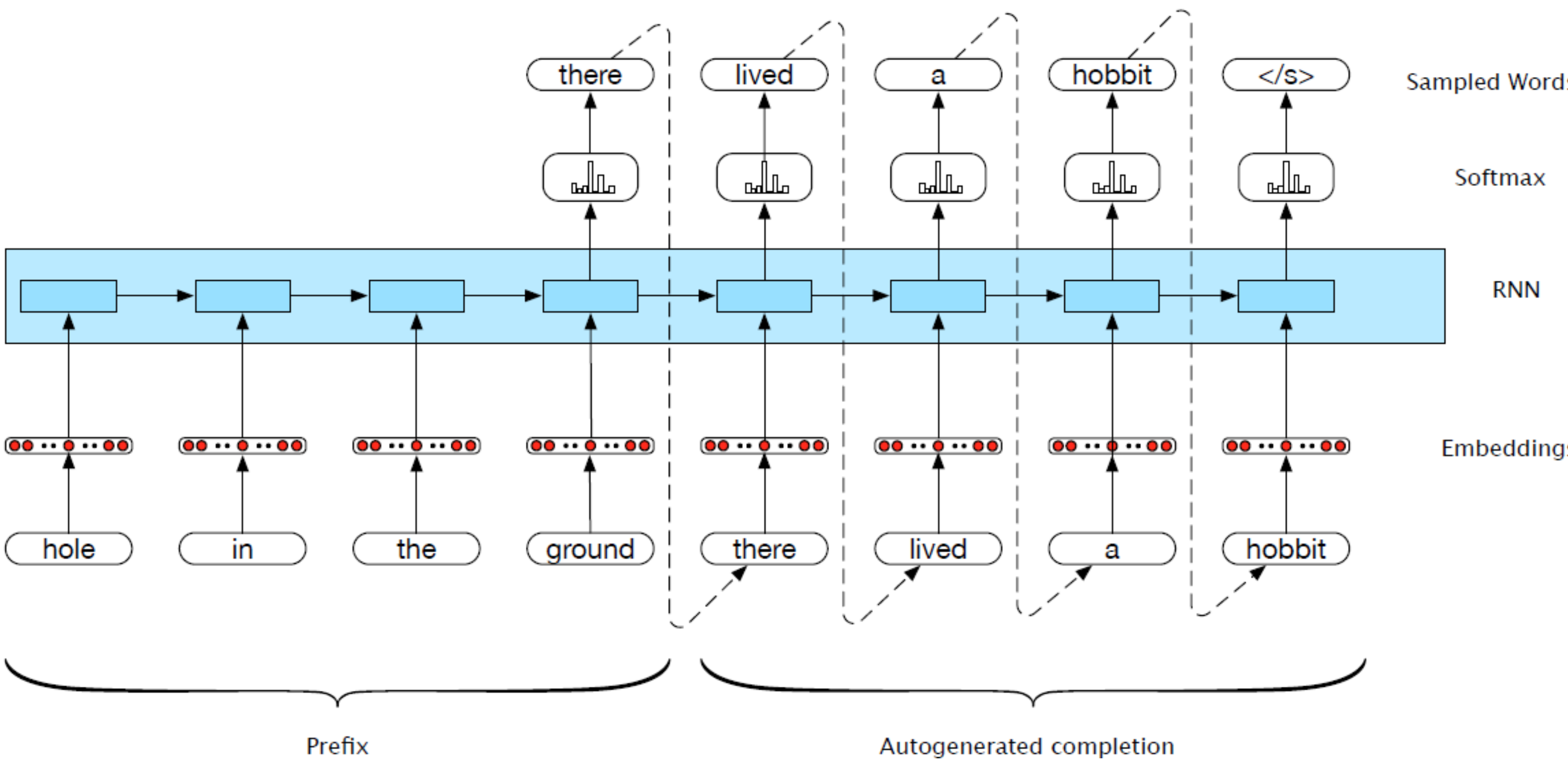
Encoder-Decoder Models

MACHINE TRANSLATION

Generation with an RNN LM



Generation with prefix



Machine Translation

Translation from one language to another

ペンシルベニア大学で講演をしています。

↓

I'm giving a talk at University of Pennsylvania

Conversational Agents aka Dialogue Systems

Digital Assistants

Answering questions on websites

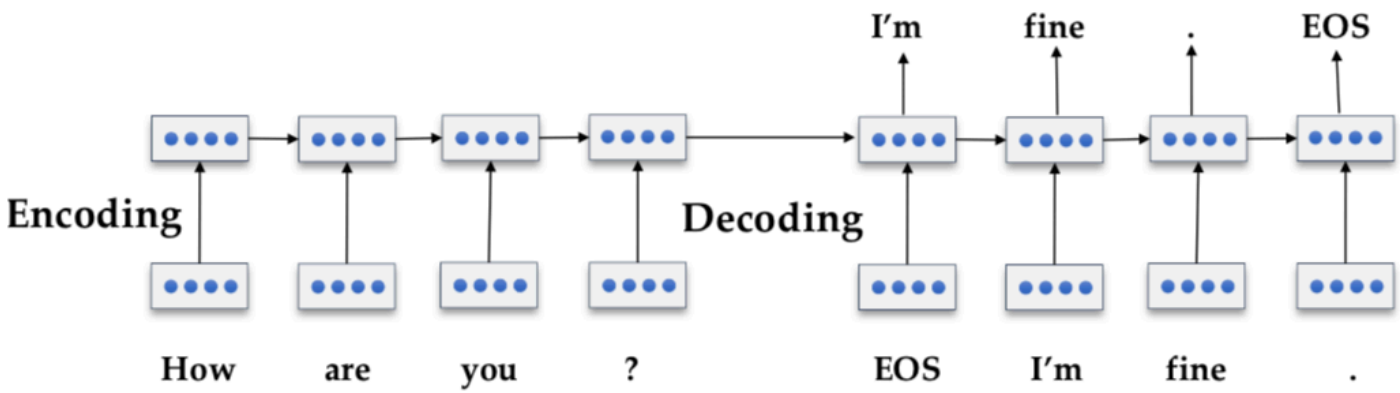
Communicating with robots

Chatting for fun

Clinical uses

Neural Chatbots

- Think of response generation as a task of transducing from the user's prior turn to the system's turn
- Response generation using encoder-decoder models



- Train a deep neural network
  - Map from user1 turn to user2 response

Current state of the art neural LMs

[ELMo]

[GPT]

[BERT]

[GPT-2]

Attention

Weaknesses of the context vector:

- Only directly available at the beginning of the process and its influence will wane as the output sequence is generated
- Context vector is a function (e.g. last, average, max, concatenation) of the hidden states of the encoder. This approach loses useful information about each of the individual encoder states

Potential solution: **attention mechanism**

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

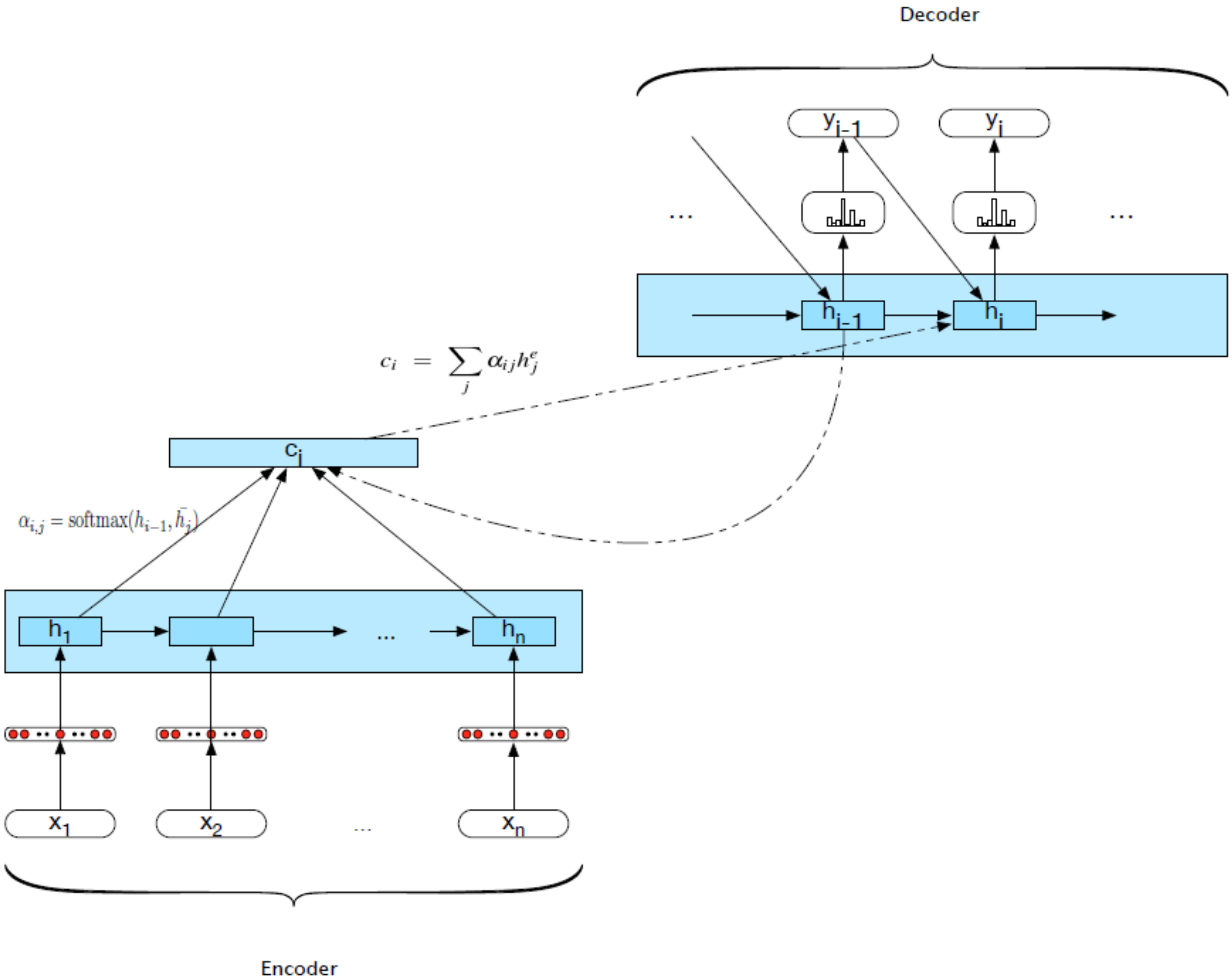
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

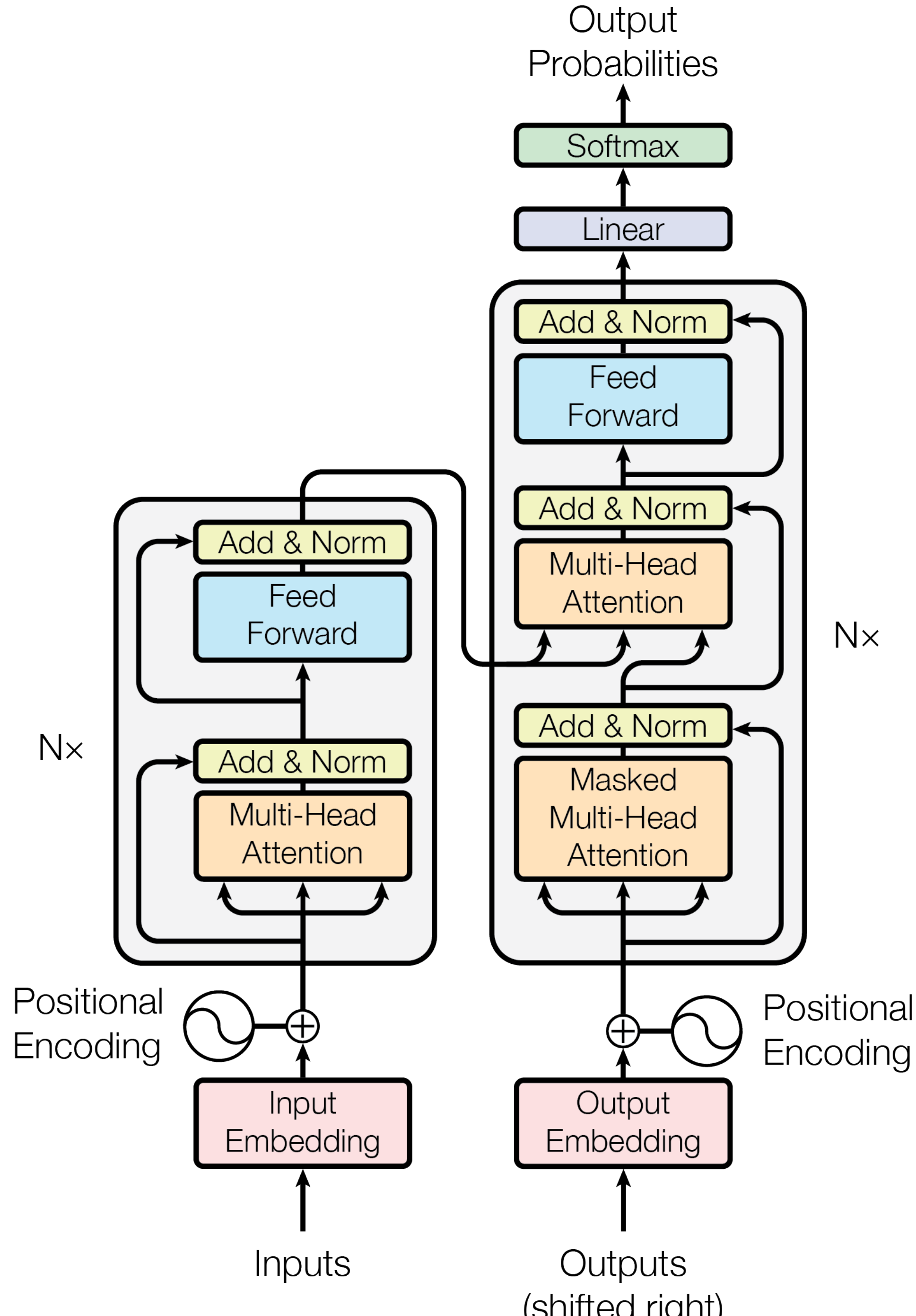
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

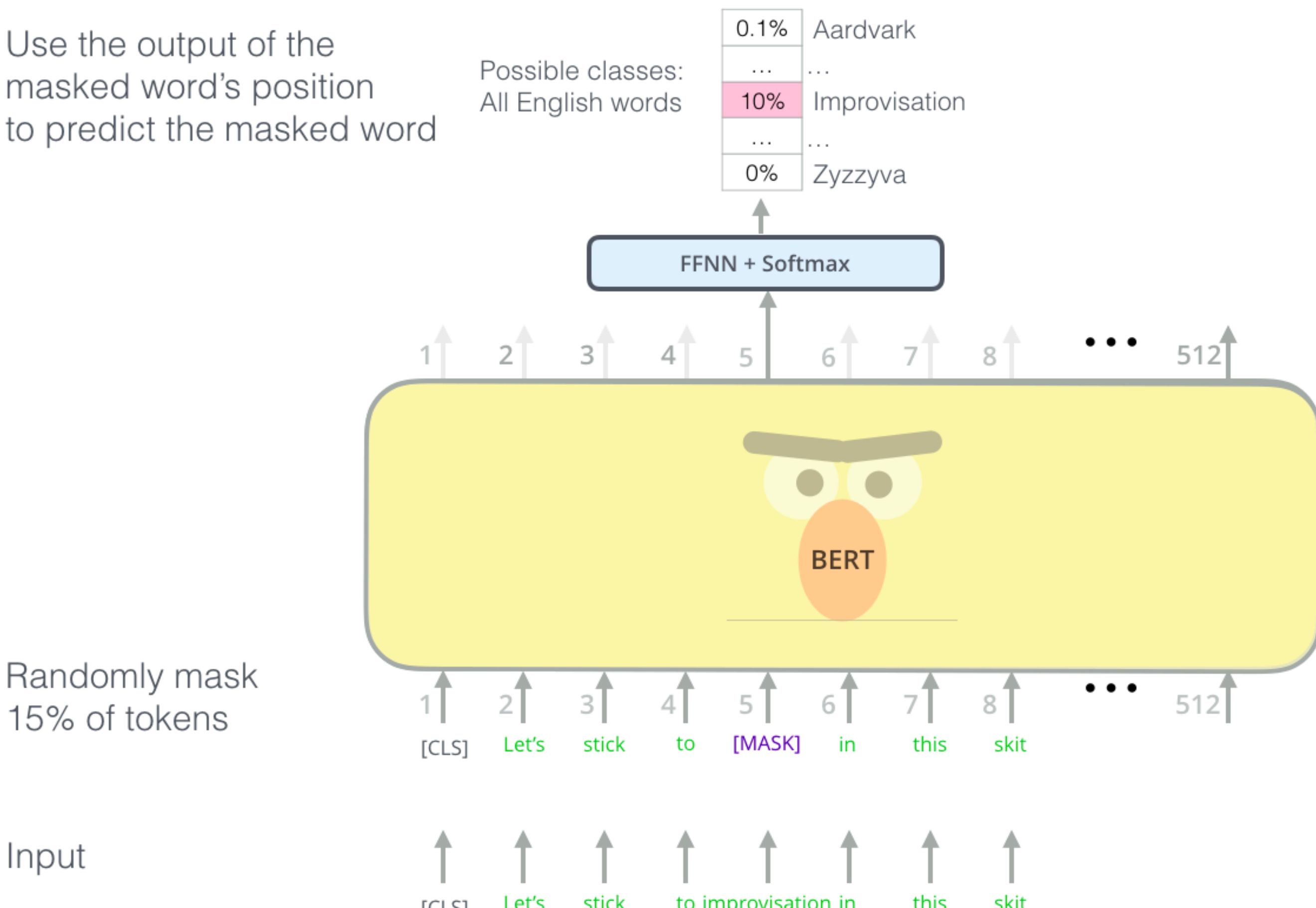
Attention mechanism



Transformer Architecture



Bidirectional Encoder Representations from Transformers (BERT)



Credit: http://jalammar.github.io/illustrated-bert/

Question Answering

what temperature is recommended for salmon?

All | Shopping | News | Videos | Images | More | Settings | Tools

About 80,100,000 results (0.58 seconds)

145 degrees Fahrenheit

The United States Food and Drug Administration recommends cooking salmon to an internal temperature of **145 degrees Fahrenheit**. Push the tip of the meat thermometer gently into the middle of the salmon fillet at its thickest part. Nov 27, 2018

healthyeating.sfgate.com › Nutrition › Nutrition in Foods

How Hot Should Salmon Cook To? | Healthy Eating | SF Gate

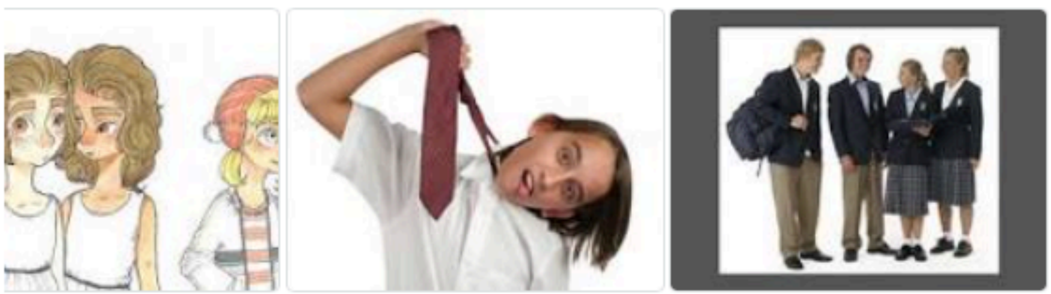
About Featured Snippets   Feedback

Question Answering

should uniforms be required in school?

All | News | Images | Videos | Shopping | More | Settings | Tools

About 44,800,000 results (0.52 seconds)



View all

Public **school** students are not **required** to wear **uniforms**, but in many religious and private **schools**, **uniforms** are **required**. ... Some positives about wearing a **uniform** in **school** are that you don't have to worry about picking out an outfit or be bullied for your choice of clothes. Sep 30, 2017

www.newsday.com › Lifestyle › Family › Kidsday

Should public schools require uniforms? | Newsday

HW10 - NMT	
1	Yue, Yuezhan
2	Pengrui, Yinhong
3	Ji-Eun, Rajalakshmi

HW8 - Hypernyms	
1	Jundong, Zitong
2	Pedro, Suyog
3	Bowen, Keyu

HW7 - NER	
1	Yuan
2	Pengrui, Yinhong
3	Bowen, Keyu

HW6 - Neural LMs	
1	Pengrui, Tien
2	Pengrui, Nupur
3	Weichen, Yinuo

HW5 - Clustering	
1	Sai, Rutuja
2	Fang, Bo
3	Shubham, Nupur

HW5 - without k	
1	Bo, Hang
2	Sai, Rutuja
3	Aayush, Shiping

HW3 – N-Gram LMs	
1	Pengrui, Tien
2	Worthan, Joseph
3	Hanbang

HW2 - Text Classif.	
1	Yue, Yuezhan
2	Ashish, Vikas
3	Sri, Simmi

HW2 - extra data	
1	Ashish, Vikas
2	Worthan, Joseph
3	Megha, Sadhana

What can you do next?

Artificial Intelligence: CIS 421/521

Machine Learning: CIS 419/519 or CIS 520

Deep Learning: CIS 522

Computer Vision: CIS 580 Machine Perception

CIS 700 courses

Independent Studies / Master Thesis

Be a TA!!


We Want
You To TA!

Thank you to our awesome TAs!








Thank you!