CIS 530: Vector Semantics part 3

JURAFSKY AND MARTIN CHAPTER 6

Reminders

HW4 IS DUE ON WEDNESDAY BY 11:59PM

NO CLASS ON WEDNESDAY

HOMEWORK 5 WILL BE RELEASED THEN

Recap: Vector Semantics

Embeddings = vector models of meaning

- More fine-grained than just a string or index
- Especially good at modeling similarity/analogy
- Can use sparse models (tf-idf) or dense models (word2vec, GLoVE)
- Just download them and use cosines!!

Distributional Information is key

What can we do with Distributional Semantics?

HISTORICAL AND SOCIO-LINGUISTICS

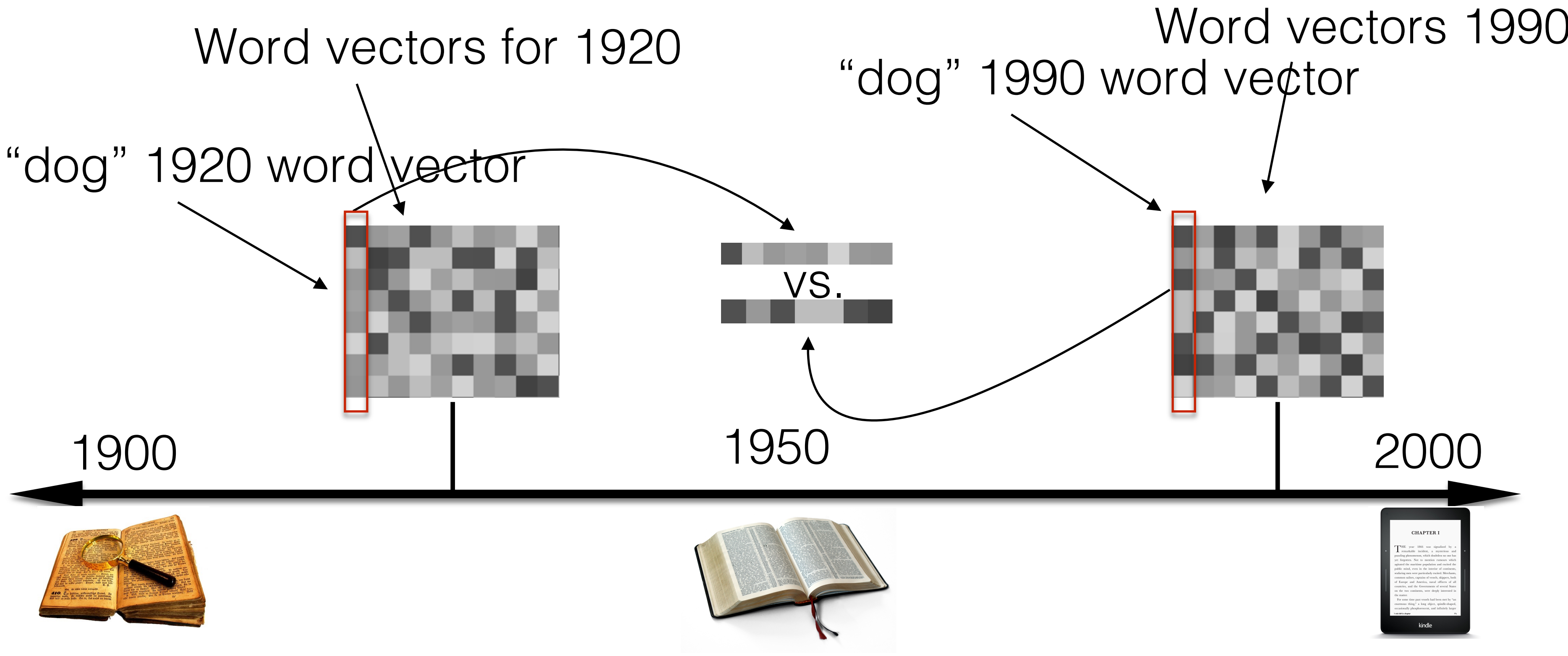
Embeddings can help study word history!

Train embeddings on old books to study changes in word meaning!!

Dan Jurafsky

Will Hamilton

Diachronic word embeddings for studying language change



Word vectors for 1920

Word vectors 1990

“dog” 1990 word vector

“dog” 1920 word vector

vs.

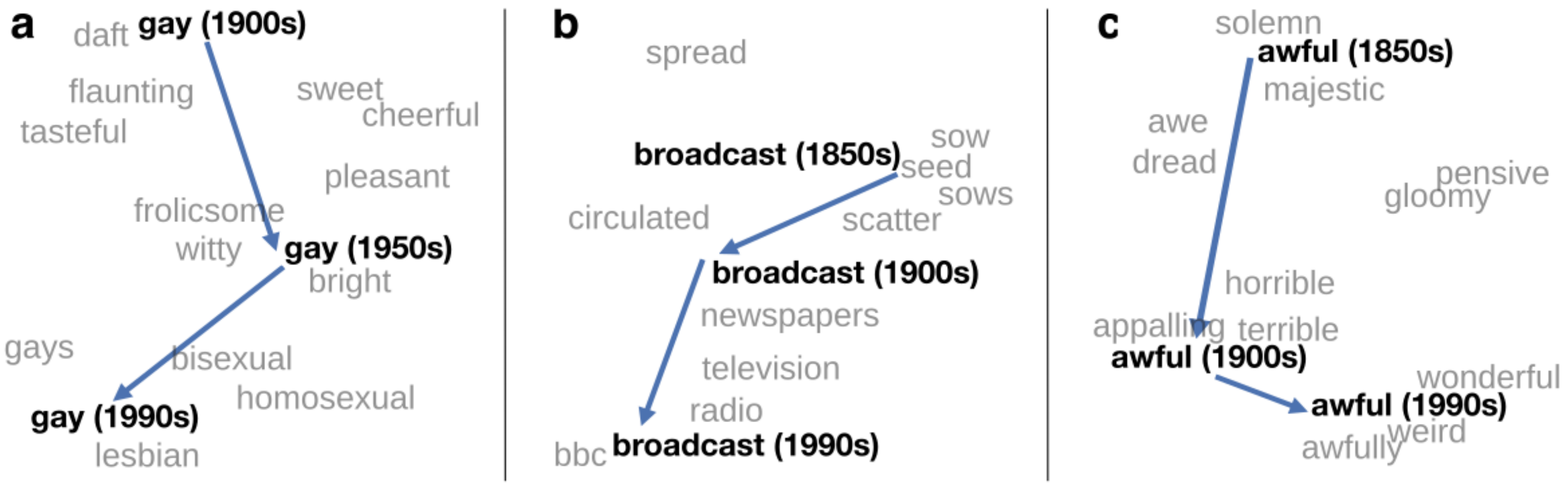
1900

1950

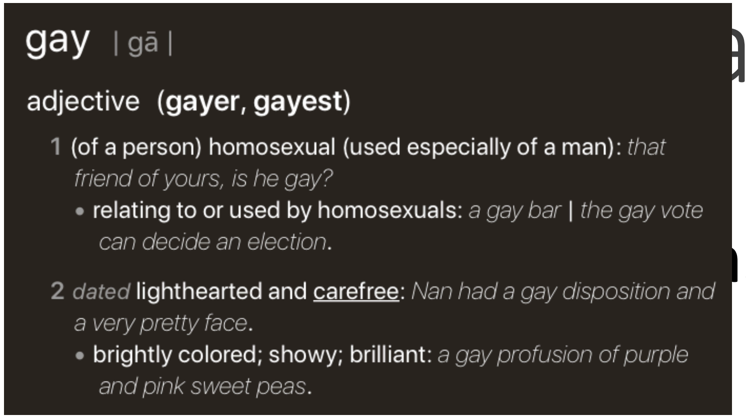
2000

Visualizing changes

Project 300 dimensions down into 2



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

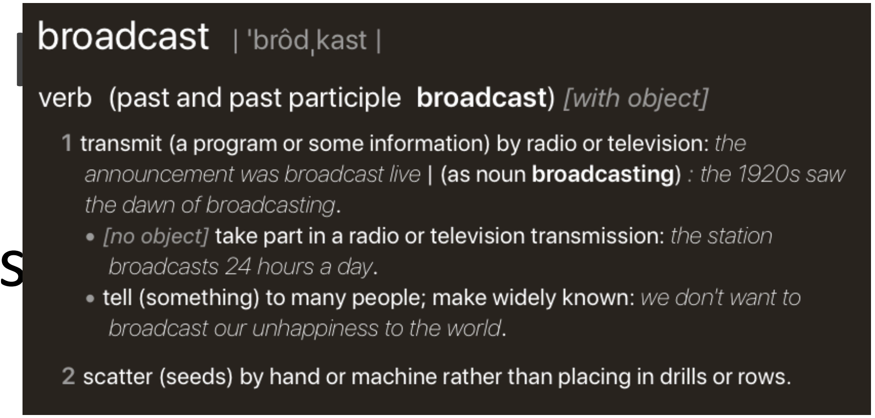


gay | gā |

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.*

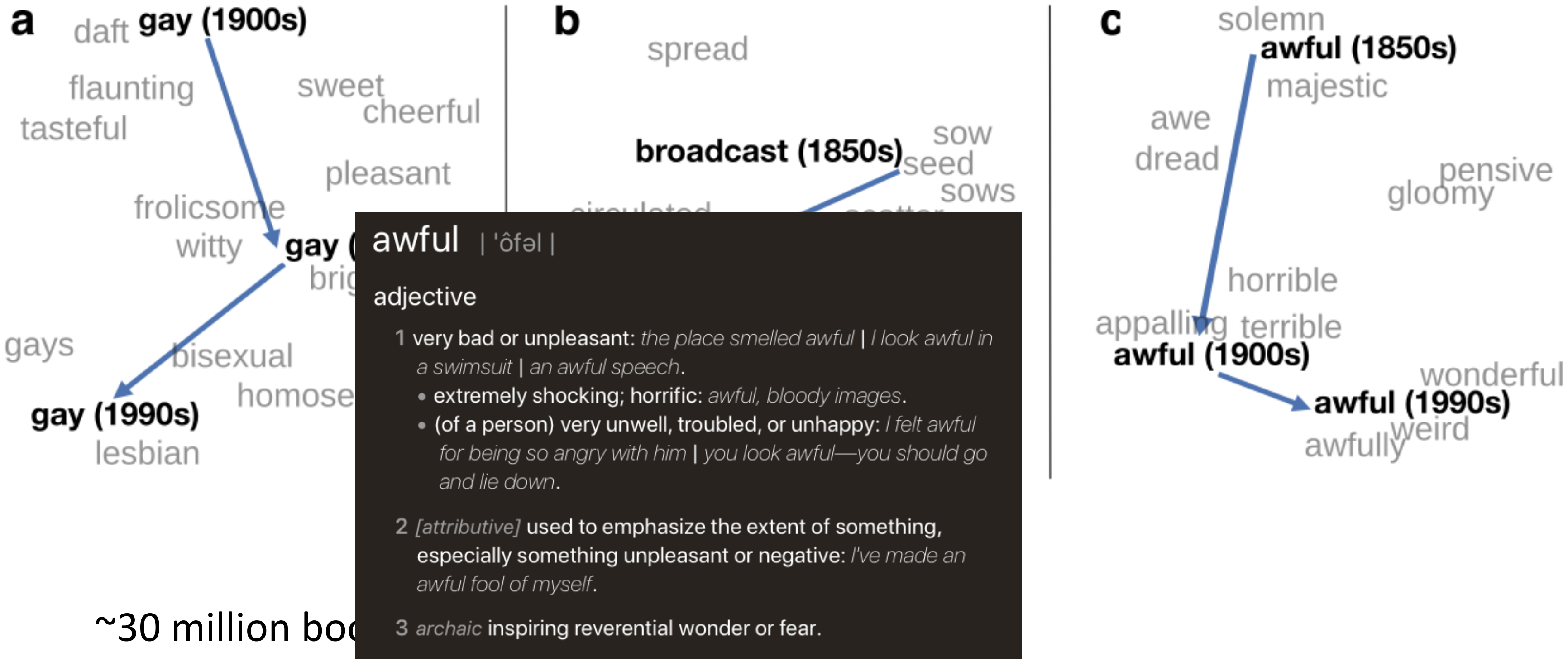


broadcast | 'brôdˌkast |

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

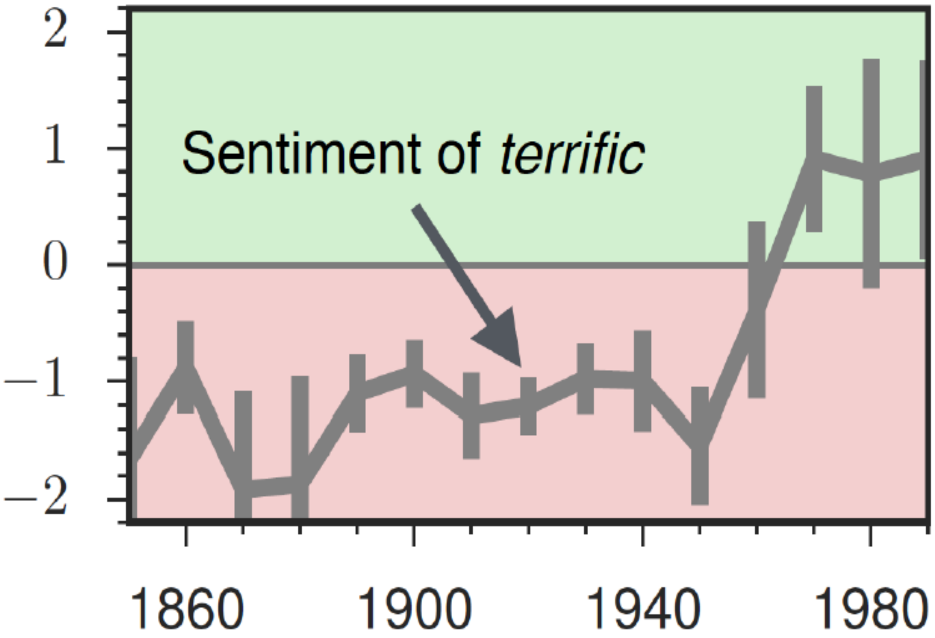
1 very bad or unpleasant: *the place smelled awful* | *I look awful in a swimsuit* | *an awful speech.*
- extremely shocking; horrific: *awful, bloody images.*
- (of a person) very unwell, troubled, or unhappy: *I felt awful for being so angry with him* | *you look awful—you should go and lie down.*

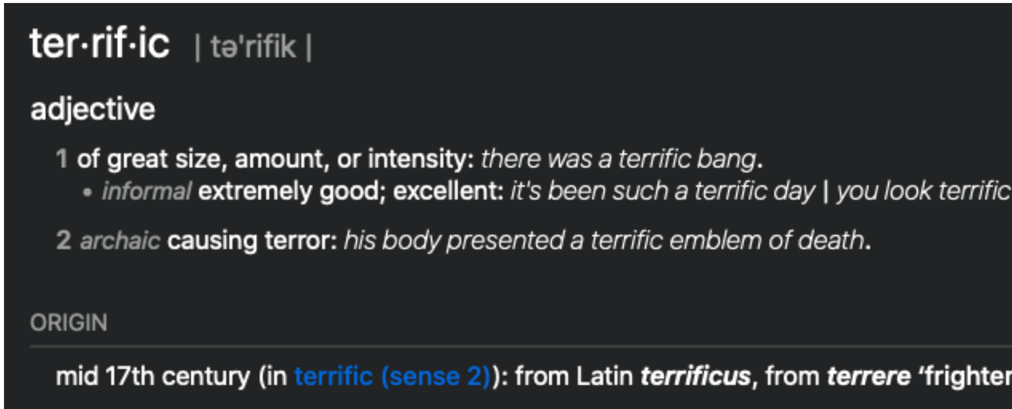
2 *[attributive]* used to emphasize the extent of something, especially something unpleasant or negative: *I've made an awful fool of myself.*

3 *archaic* inspiring reverential wonder or fear.

~30 million boo

The evolution of sentiment words





ter·rif·ic | tə'rifik |

adjective

1 **of great size, amount, or intensity:** *there was a terrific bang*.
- *informal* **extremely good; excellent:** *it's been such a terrific day* | *you look terrific*

2 *archaic* **causing terror:** *his body presented a terrific emblem of death*.

ORIGIN

mid 17th century (in terrific (sense 2)): from Latin ***terrificus***, from ***terrere*** ‘frighter

Embeddings and bias

Embeddings reflect cultural bias

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." In *Advances in Neural Information Processing Systems*, pp. 4349-4357. 2016.

Ask “Paris : France :: Tokyo : x”

- x = Japan

Ask “father : doctor :: mother : x”

- x = nurse

Ask “man : computer programmer :: woman : x”

- x = homemaker

Measuring cultural bias

Implicit Association test (Greenwald et al 1998): How associated are

- concepts (*flowers*, *insects*) & attributes (*pleasantness*, *unpleasantness*)?
- Studied by measuring timing latencies for categorization.

Psychological findings on US participants:

- African-American names are associated with unpleasant words (more than European-American names)
- Male names associated more with math, female names with arts
- Old people's names with unpleasant words, young people with pleasant words.

Embeddings reflect cultural bias

Aylin Caliskan, Joanna J. Bruson and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356:6334, 183-186.

Caliskan et al. replication with embeddings:

- African-American names (*Leroy, Shaniqua*) had a higher GloVe cosine with unpleasant words (*abuse, stink, ugly*)
- European American names (*Brad, Greg, Courtney*) had a higher cosine with pleasant words (*love, peace, miracle*)

Embeddings reflect and replicate all sorts of pernicious biases.

Directions

Debiasing algorithms for embeddings

- Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James Y., Saligrama, Venkatesh, and Kalai, Adam T. (2016). **Man is to computer programmer as woman is to homemaker?** debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357.

Use embeddings as a historical tool to study bias

Embeddings as a window onto history

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

Use the Hamilton historical embeddings

The cosine similarity of embeddings for decade X for occupations (like teacher) to male vs female names

- Is correlated with the actual percentage of women teachers in decade X

History of biased framings of women

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

Embeddings for competence adjectives are biased toward men

- *Smart, wise, brilliant, intelligent, resourceful, thoughtful, logical, etc.*

This bias is slowly decreasing

Princeton Trilogy experiments

Study 1: Katz and Braley (1933)

Investigated whether traditional social stereotypes had a cultural basis

Ask 100 male students from Princeton University to choose five traits that characterized different ethnic groups (for example Americans, Jews, Japanese, Negroes) from a list of 84 word

84% of the students said that Negroes were superstitious and 79% said that Jews were shrewd. They were positive towards their own group.

Study 2: Gilbert (1951)
Less uniformity of agreement about unfavorable traits than in 1933.

Study 3: Karlins et al. (1969)

Many students objected to the task but this time there was greater agreement on the stereotypes assigned to the different groups compared with the 1951 study. Interpreted as a re-emergence of social stereotyping but in the direction more favorable stereotypical images.

Embeddings reflect ethnic stereotypes over time

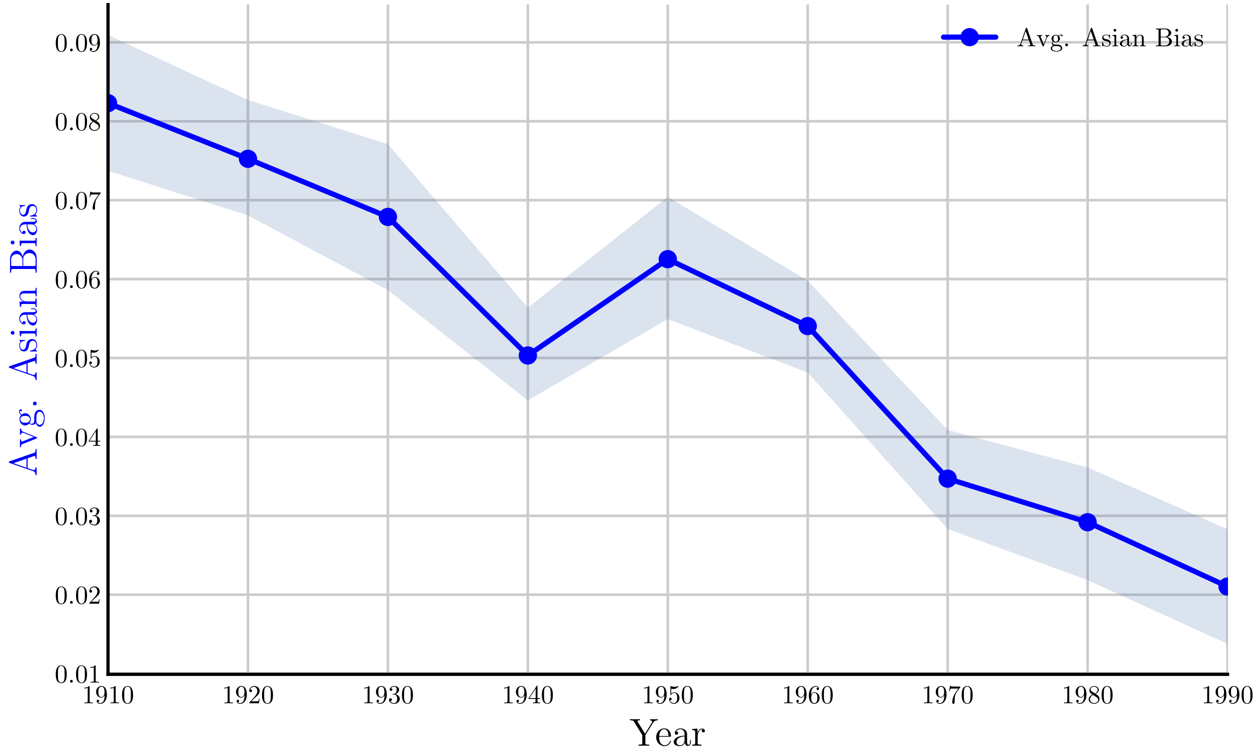
Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

- Princeton trilogy experiments
- Attitudes toward ethnic groups (1933, 1951, 1969) scores for adjectives
  - *industrious, superstitious, nationalistic*, etc
- Cosine of Chinese name embeddings with those adjective embeddings correlates with human ratings.

Change in linguistic framing 1910-1990

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

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



Changes in framing: adjectives associated with Chinese

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

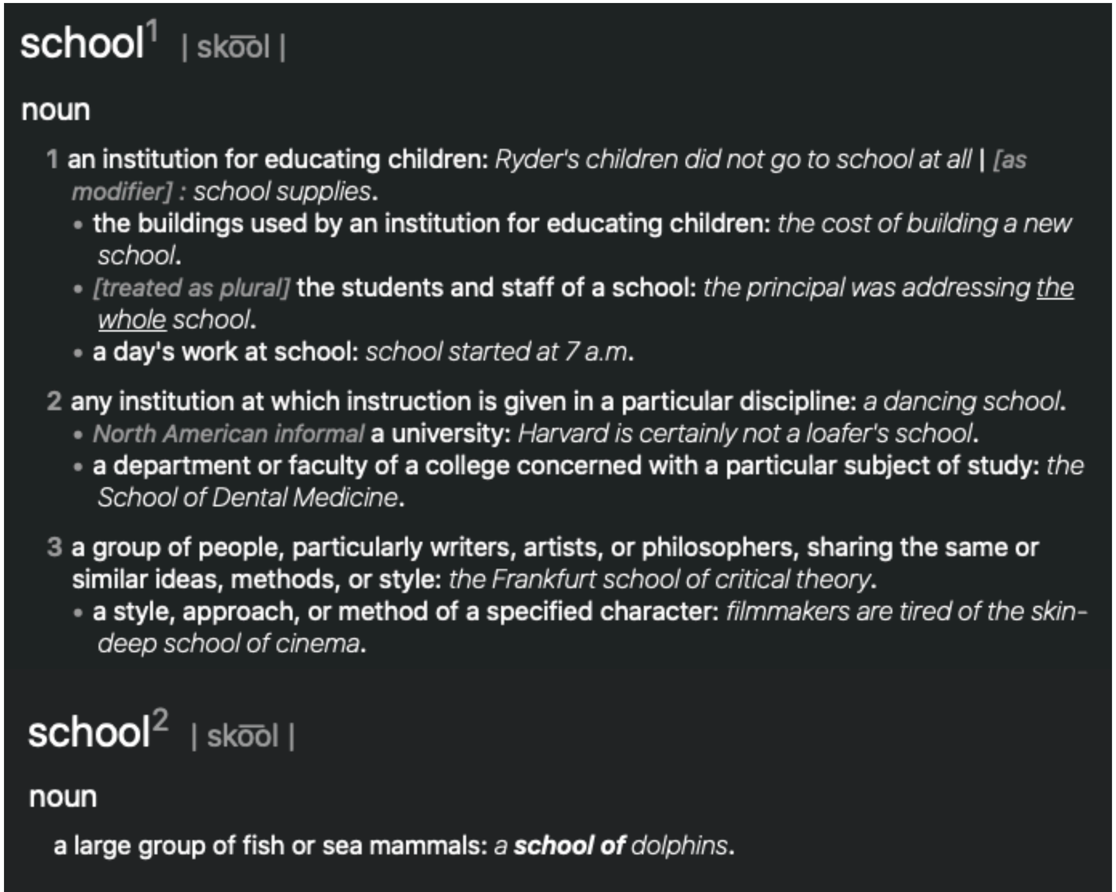
1910	1950	1990
Irresponsible	Disorganized	Inhibited
Envious	Outrageous	Passive
Barbaric	Pompous	Dissolute
Aggressive	Unstable	Haughty
Transparent	Effeminate	Complacent
Monstrous	Unprincipled	Forceful
Hateful	Venomous	Fixed
Cruel	Disobedient	Active
Greedy	Predatory	Sensitive
Bizarre	Boisterous	Hearty

What should a semantic model be able to do?

GOALS FOR DISTRIBUTIONAL SEMANTICS

Goal: Word Sense

The meaning of a word can often be broken up into distinct **senses**. Sometimes we describe these words as **polysemous** or **homonymous**



school[1] | sko͞ol |

noun

1 **an institution for educating children:** *Ryder's children did not go to school at all* | *[as modifier] : school supplies.*

- **the buildings used by an institution for educating children:** *the cost of building a new school.*
- *[treated as plural]* **the students and staff of a school:** *the principal was addressing the whole school.*
- **a day's work at school:** *school started at 7 a.m.*

2 **any institution at which instruction is given in a particular discipline:** *a dancing school.*

- *North American informal* **a university:** *Harvard is certainly not a loafer's school.*
- **a department or faculty of a college concerned with a particular subject of study:** *the School of Dental Medicine.*

3 **a group of people, particularly writers, artists, or philosophers, sharing the same or similar ideas, methods, or style:** *the Frankfurt school of critical theory.*

- **a style, approach, or method of a specified character:** *filmmakers are tired of the skin-deep school of cinema.*

school[2] | sko͞ol |

noun

a large group of fish or sea mammals: *a **school of** dolphins.*

Goal: Word Sense

Do the vector based representations of words that we’ve looked at so far handle word sense well?

Goal: Word Sense

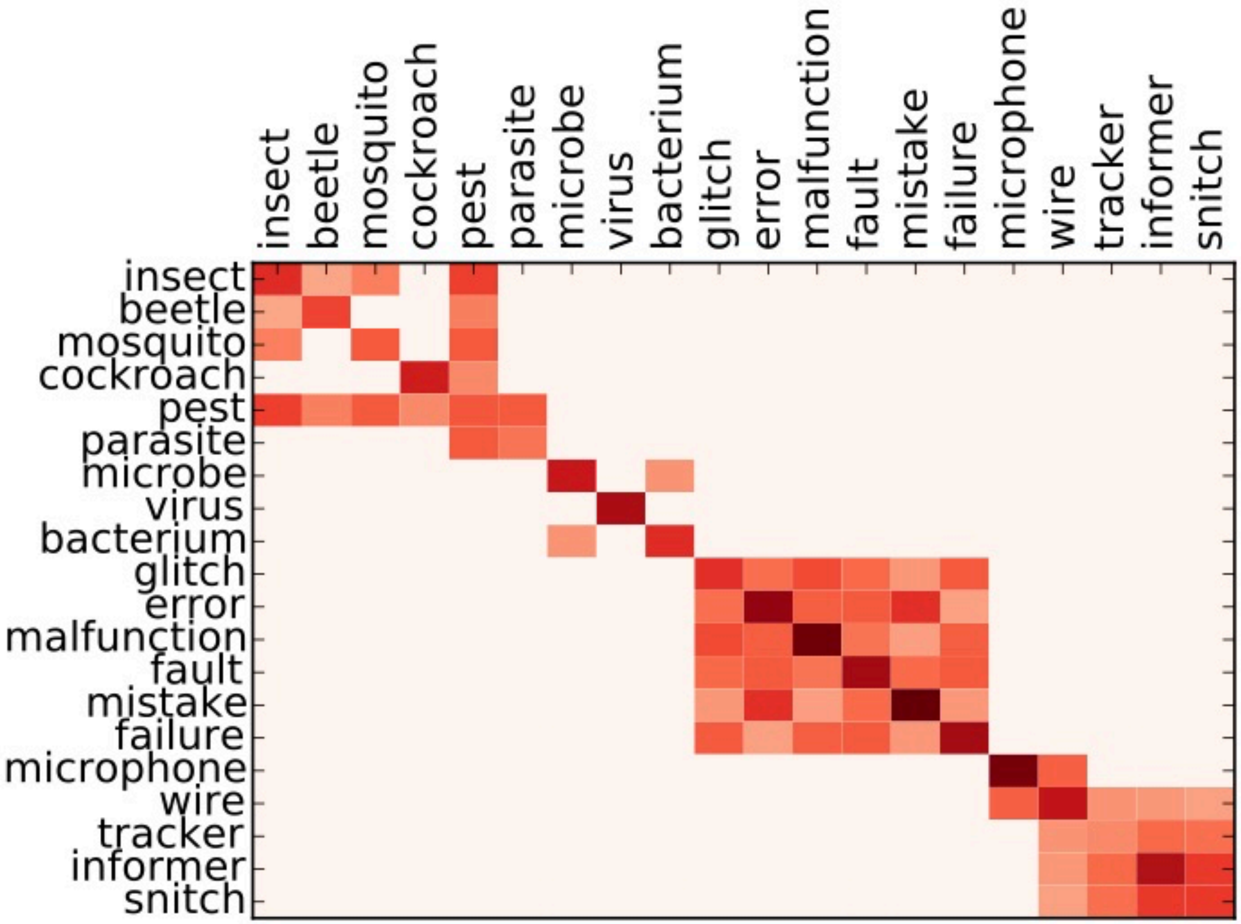
Do the vector based representations of words that we’ve looked at so far handle word sense well? No! All senses of a word are collapsed into the same word vector.

One solution would be to learn a separate representation for each sense. However, it is hard to enumerate a discrete set of senses for a word.

A good semantic model should be able to automatically capture variation in meaning without a manually specified sense inventory.

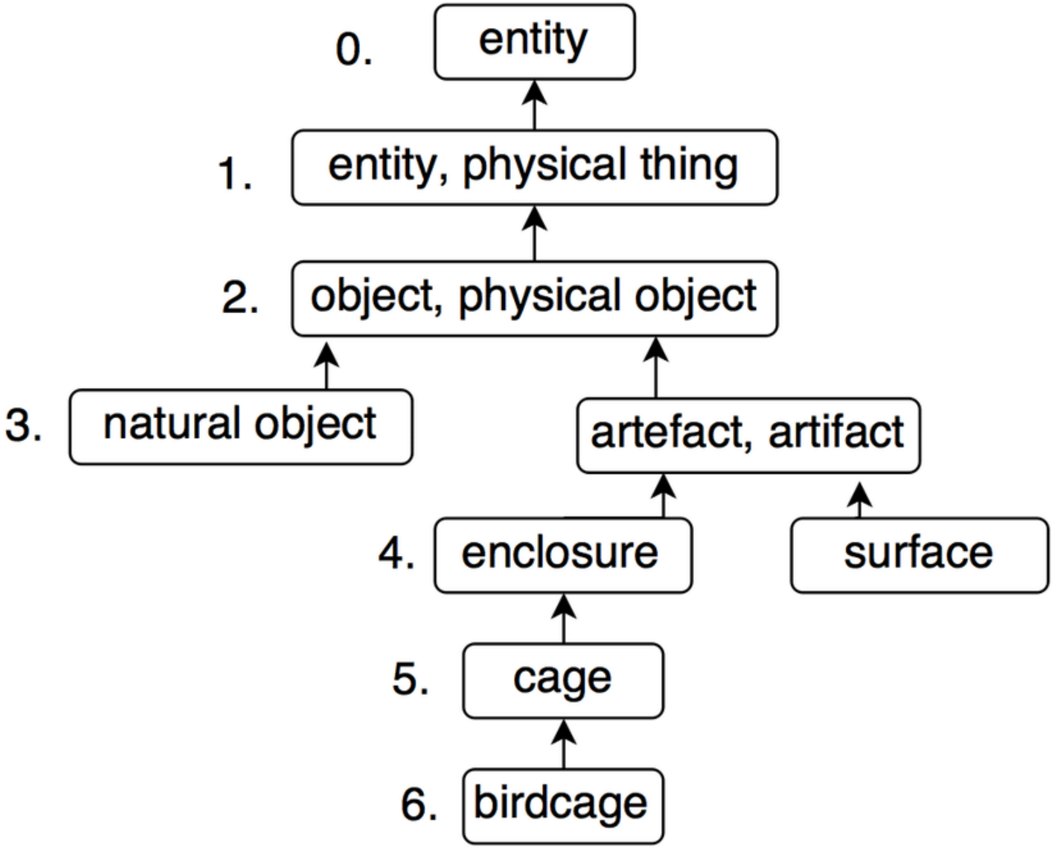
Clustering Paraphrases by Word Sense. Anne Cocos and Chris Callison-Burch. NAACL 2016.

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*).



The Distributional Inclusion Hypotheses and Lexical Entailment. Maayan Geffet and Ido Dagan. ACL 2005.

Goal: Hypernomy

Distributional inclusion hypotheses, which correspond to the two directions of inference relating distributional feature inclusion and lexical entailment. Let *vi* and *wj* be two word senses of words *w* and *v*, and let *vi => wj* denote the (directional) entailment relation between these senses. Assume further that we have a measure that determines the set of *characteristic* features for the meaning of each word sense. Then we would hypothesize:

Hypothesis I:

If *vi => wj* then all the characteristic features of *vi* are expected to appear with *wj*.

Hypothesis II:

If all the characteristic features of *vi* appear with *wj* then we expect that *vi => wj*.

Distributional Lexical Entailment by Topic Coherence. Laura Rimell. EACL 2014.

Goal: Hypernomy

Distributional Inclusion Hypothesis (DIH) states that a hyperonym occurs in all the contexts of its hyponyms.

For example, **lion** is a hyponym of **animal**, but **mane** is a likely context of **lion** and unlikely for **animal**, contradicting the DIH.

Rimell proposes measuring hyponymy using coherence: the contexts of a general term minus those of a hyponym are coherent, but the reverse is not true.

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.**

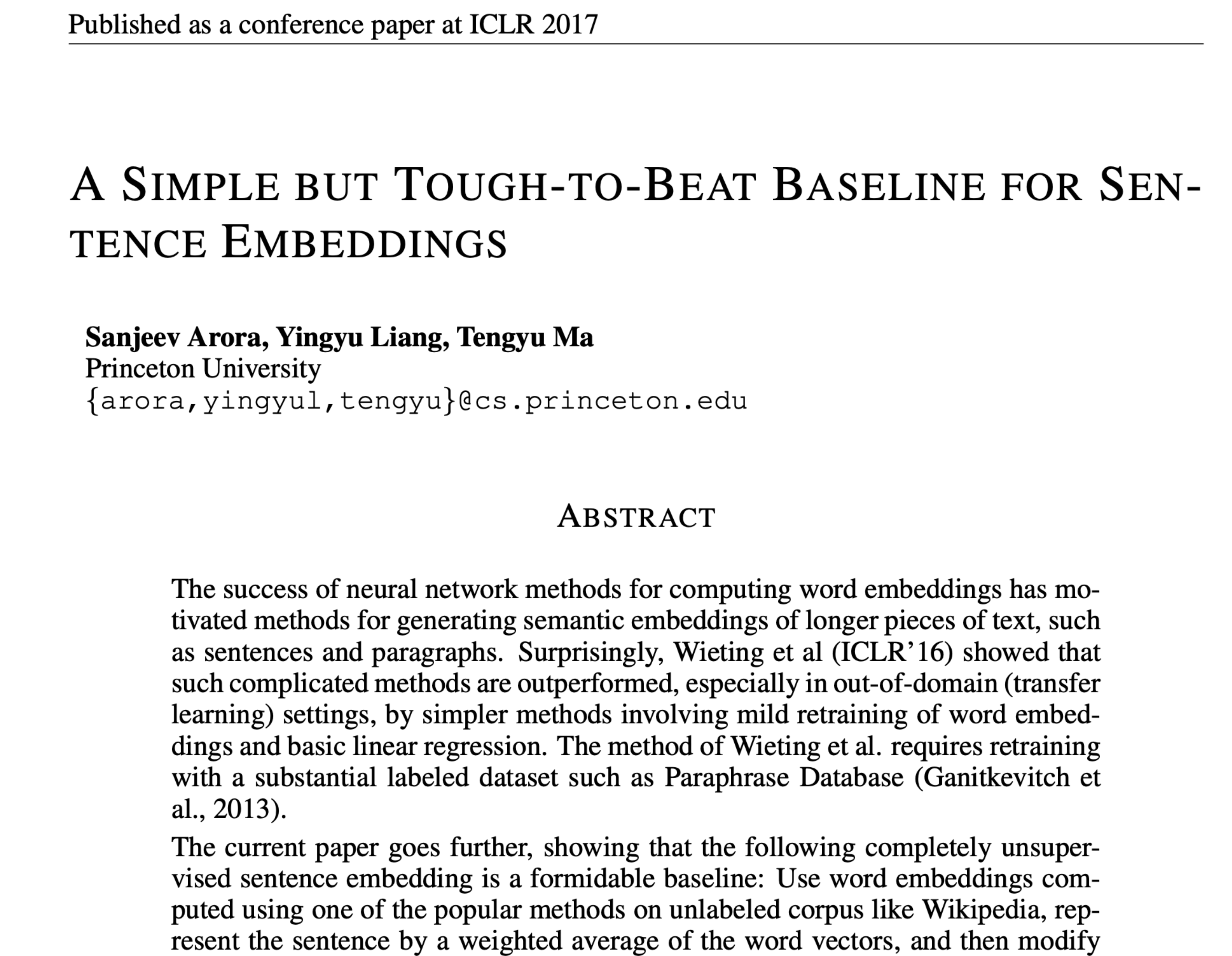
Vector-based models of semantic composition. Jeff Mitchell and Mirella Lapata. ACL 2010.

Goal: Compositionality

For vector space models, we have the challenge of how to **compose** word vectors to construct **phrase representation**. One option is to represent phrases as vectors too.

If we use the same vector space as for words, the challenge is then to find a composition function that maps a pair of vectors onto a new vectors.

Mitchell and Lapata experimented with a variety of functions and found that component-wise multiplication was as good or better than other functions that they tried.



A Simple but Tough-to-Beat Baseline for Sentence Embeddings

Sanjeev Arora, Yingyu Liang, Tengyu Ma
Princeton University
{arora,yingyul,tengyu}@cs.princeton.edu

Abstract

The success of neural network methods for computing word embeddings has motivated methods for generating semantic embeddings of longer pieces of text, such as sentences and paragraphs. Surprisingly, Wieting et al (ICLR'16) showed that such complicated methods are outperformed, especially in out-of-domain (transfer learning) settings, by simpler methods involving mild retraining of word embeddings and basic linear regression. The method of Wieting et al. requires retraining with a substantial labeled dataset such as Paraphrase Database (Ganitkevitch et al., 2013).

The current paper goes further, showing that the following completely unsupervised sentence embedding is a formidable baseline: Use word embeddings computed using one of the popular methods on unlabeled corpus like Wikipedia, represent the sentence by a weighted average of the word vectors, and then modify

Vector-based models of semantic composition. Jeff Mitchell and Mirella Lapata. ACL 2010.

Goal: Compositionality

The problem with componentwise multiplication is that it is commutative and therefore insensitive to word order.

These two sentences contain exactly the same words, but they do not have the same meaning:

1. It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.
2. hat day the office manager, who was drinking, hit the problem sales worker with a bottle, but it was not serious.

Goal: Grounding

A semantic model should capture how language relates to the world via sensory perception and motor control.

The process of connecting language to the world is called **grounding**.

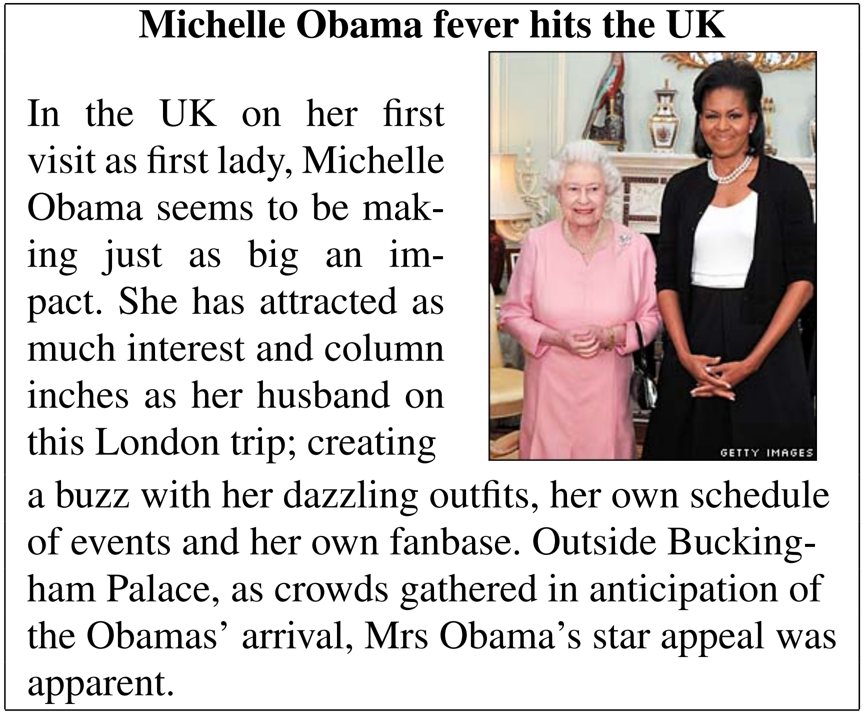
Vector space models that rely entirely on how words co-occur with other words is not grounded, since they are constructed solely from text.

Yansong Feng and Mirella Lapata (2010). Visual Information in Semantic Representation. Proceedings of NAACL.

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



Michelle Obama fever hits the UK

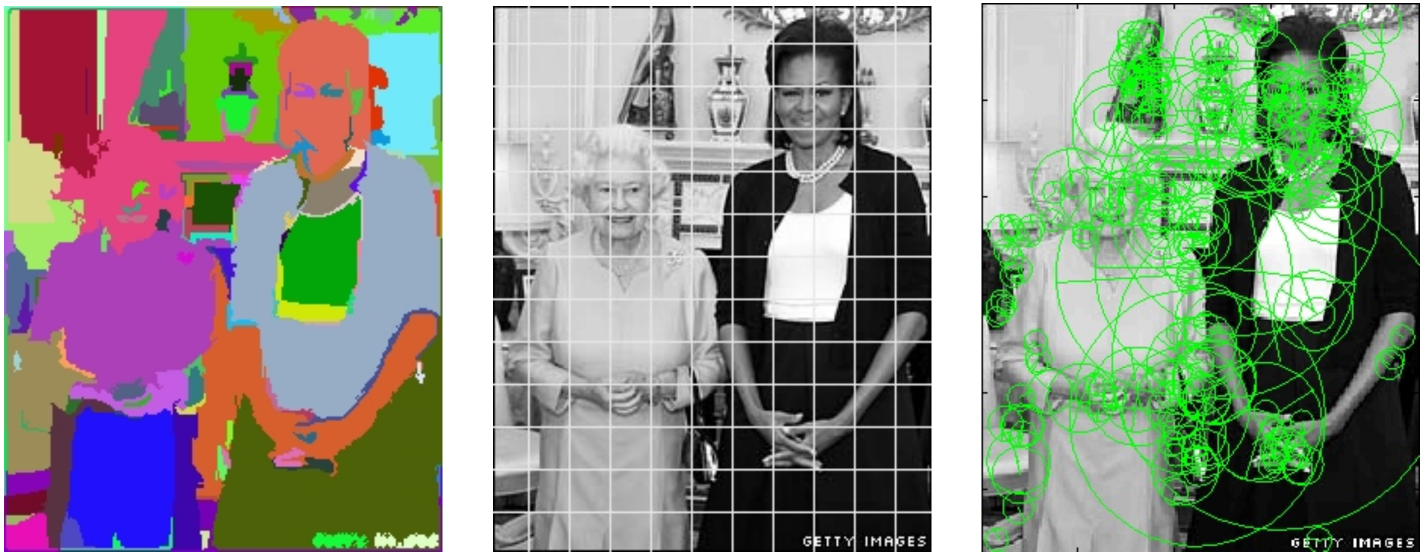
In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact. She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase. Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

Yansong Feng and Mirella Lapata (2010). Visual Information in Semantic Representation. Proceedings of NAACL.

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.

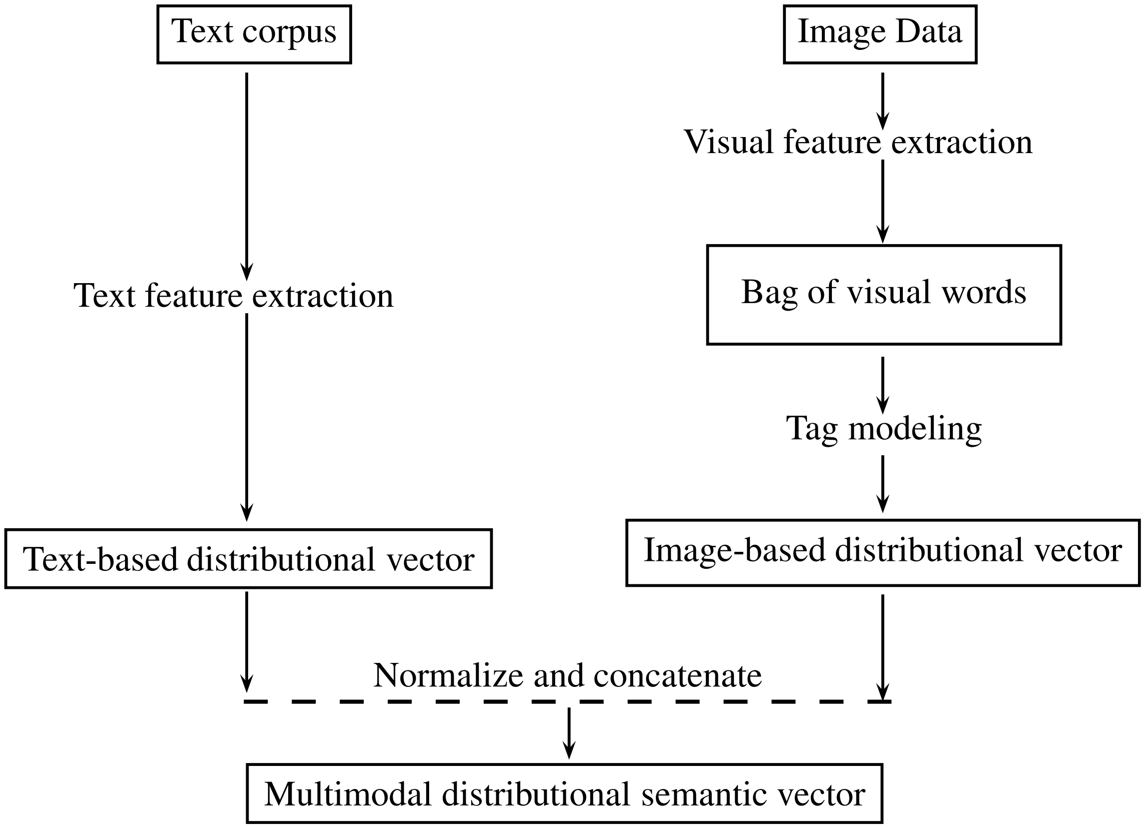
Use collections of documents that contain pictures



Elia Bruni, Giang Binh Tran, Marco Baroni (2011). Distributional semantics from text and images. Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics

Goal: Grounding

How can we ground a distributional semantic model? Simplest way train word vectors, and then concatenate them with image vectors.



Text corpus

Text feature extraction

Text-based distributional vector

Image Data

Visual feature extraction

Bag of visual words

Tag modeling

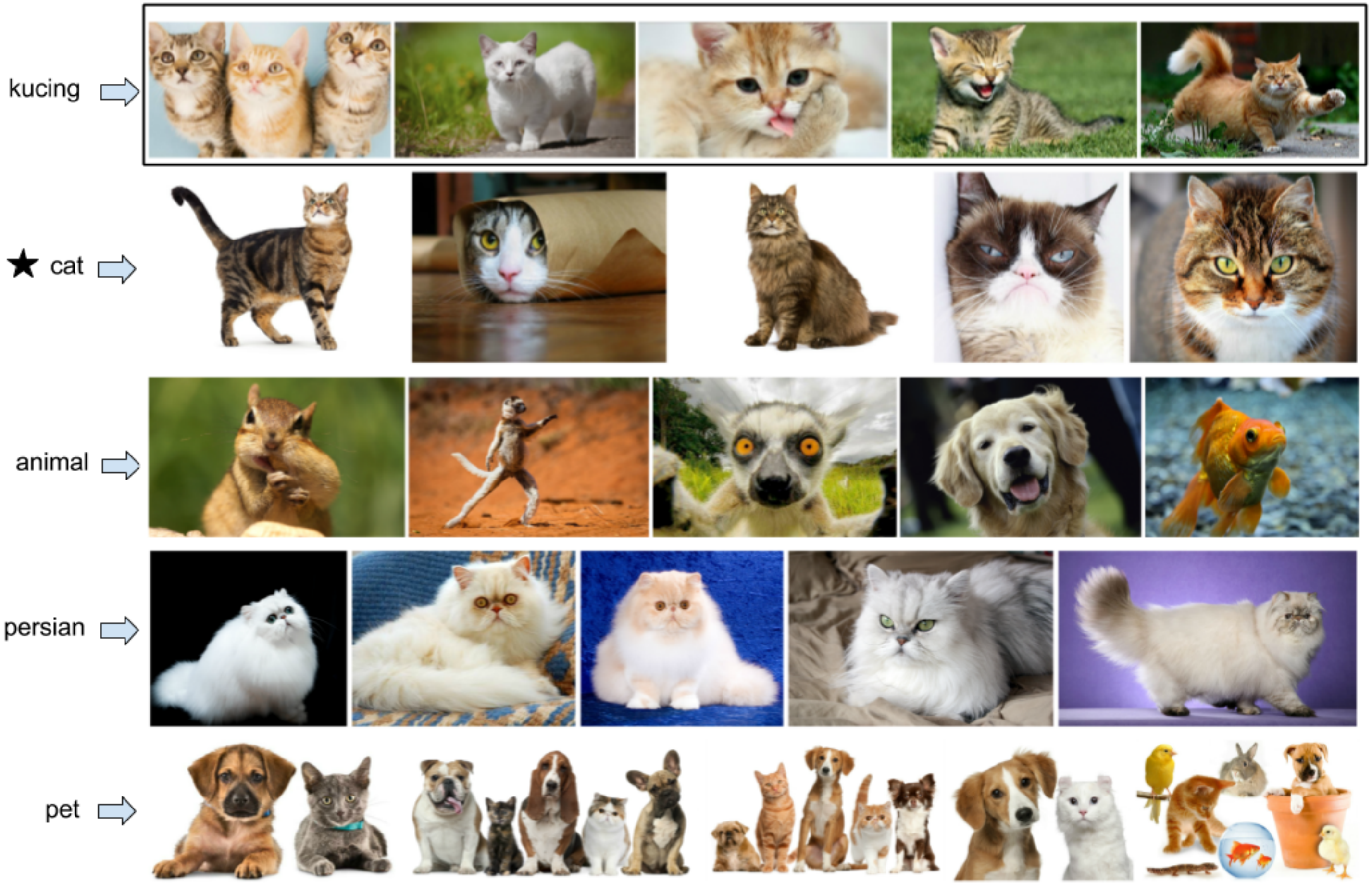
Image-based distributional vector

Normalize and concatenate

Multimodal distributional semantic vector

Learning Translations via Images with a Massively Multilingual Image Dataset. John Hewitt*, Daphne Ippolito*, Brendan Callahan, Reno Kriz, Derry Wijaya and Chris Callison-Burch. ACL 2018

Goal: Grounding

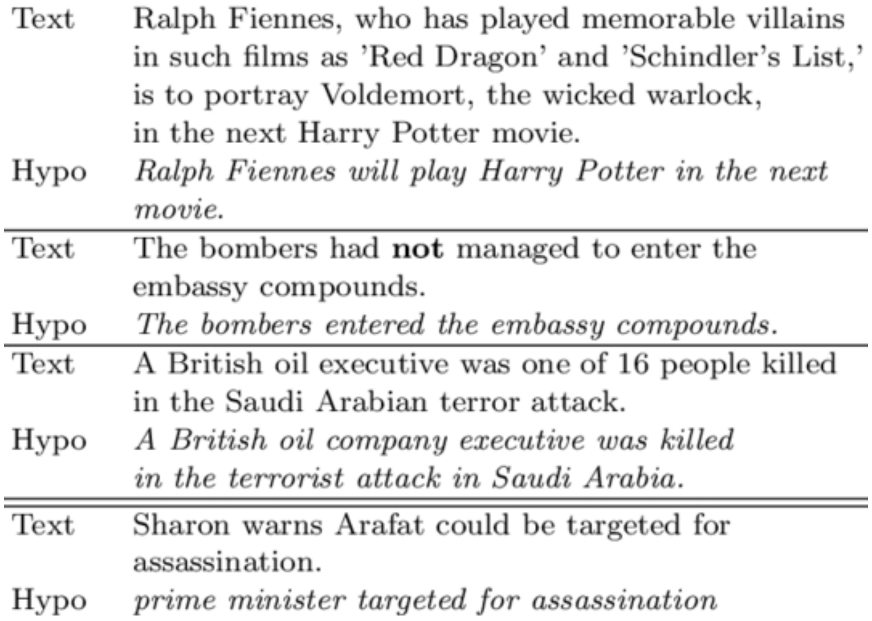


Recognizing Textual Entailment: Models and Applications. Ido Dagan, Dan Roth, Mark Sammons, Fabio Massimo Zanzotto. Synthesis Lectures on Human Language Technologies, 2013.

Goal: Logical inference

Sentences can express complex thoughts and build change of reasoning. Logic formalize this. One goal of semantic models is to support the logical notions of *truth* and of *entailment*.

Vectors do not have logical structure, but they can be used in a system that computes entailment. One challenge problem that is proposed for NLU is the task of *recognizing textual entailment.*



Text	Ralph Fiennes, who has played memorable villains in such films as 'Red Dragon' and 'Schindler's List,' is to portray Voldemort, the wicked warlock, in the next Harry Potter movie.
Hypo	*Ralph Fiennes will play Harry Potter in the next movie.*
Text	The bombers had **not** managed to enter the embassy compounds.
Hypo	*The bombers entered the embassy compounds.*
Text	A British oil executive was one of 16 people killed in the Saudi Arabian terror attack.
Hypo	*A British oil company executive was killed in the terrorist attack in Saudi Arabia.*
Text	Sharon warns Arafat could be targeted for assassination.
Hypo	*prime minister targeted for assassination*

Goal: Context Dependence

One goal of a semantic model is to capture how meaning depends on context. For example, a *small elephant* is not a *small animal*, but a *large ant* is. The meanings of *small* and *large* depend on the nouns that they modify.

Similarly performing word sense disambiguation requires understanding how a word is used in context.

The KGB planted a ***bug*** *in the Oval Office.*

I found a ***bug*** *swimming in my soup.*

Recent large language models like ELMo and BERT create different vectors for words depending on the sentences that they appear in.

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)