TTIC 31190: Natural Language Processing

Kevin Gimpel

Winter 2016

Lecture 14: Introduction to Computational Semantics

Announcements

- if you haven’t emailed me to set up a 15-minute meeting to discuss your project proposal, please do so
  - times posted on course webpage
  - let me know if none of those work for you
- Assignment 3 due Feb 29
- email me to sign up for your (10-minute) class presentation on 3/3 or 3/8

Roadmap

- classification
- words
- lexical semantics
- language modeling
- sequence labeling
- neural network methods in NLP
- syntax and syntactic parsing
- semantic compositionality
- semantic parsing
- unsupervised learning
- machine translation and other applications

Roadmap

- classification
- words
- lexical semantics
- language modeling
- sequence labeling
- neural network methods in NLP
- syntax and syntactic parsing
- computational semantics (today)
  - compositionality
  - semantic parsing
- machine translation (Thursday)
- other NLP applications (next Tuesday)

Compositional Semantics

- "how should the meanings of words combine to create the meaning of something larger?"
- there's currently a lot of work in producing vector representations of sentences and documents
- simplest case: how should two word vectors be combined to create a vector for a bigram?
- explosion of work in this area in the neural network era, but earlier work began ~2007

Evaluating Compositional Semantics

- compute similarity of two bigrams under your model, then compute correlation with human judgments:

		BigramSim
television programme	tv set	5.8
training programme	education course	5.7
bedroom window	education officer	1.3

(Mitchell and Lapata, 2010)

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 34 (2010) 1388–1429

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2010.01106.x

Composition in Distributional Models of Semantics

Jeff Mitchell, Mirella Lapata

School of Informatics, University of Edinburgh

Received 6 May 2009; received in revised form 22 March 2010; accepted 25 March 2010

Abstract

Vector-based models of word meaning have become increasingly popular in cognitive science. The appeal of these models lies in their ability to represent meaning simply by using distributional information under the assumption that words occurring within similar contexts are semantically similar. Despite their widespread use, vector-based models are typically directed at representing words in isolation, and methods for constructing representations for phrases or sentences have received little attention in the literature. This is in marked contrast to experimental evidence (e.g., in sentential priming) suggesting that semantic similarity is more complex than simply a relation between isolated words. This article proposes a framework for representing the meaning of word combinations in vector space. Central to our approach is vector composition, which we operationalize in terms of additive and multiplicative functions. Under this framework, we introduce a wide range of composition models that we evaluate empirically on a phrase similarity task.

Bigram Composition Functions

J. Mitchell, M. Lapata/Cognitive Science 34 (2010)

Table 5
Composition functions considered in our experiments

Model	Function
Additive	$p_i = u_i + v_i$
Kintsch	$p_i = u_i + v_i + n_i$
Multiplicative	$p_i = u_i \cdot v_i$
Tensor product	$p_{i,j} = u_i \cdot v_j$
Circular convolution	$p_i = \sum_j u_j \cdot v_{i-j}$
Weighted additive	$p_i = \alpha v_i + \beta u_i$
Dilation	$p_i = v_i \sum_j u_j u_j + (\lambda - 1) u_i \sum_j u_j v_j$
Head only	$p_i = v_i$
Target unit	$p_i = v_i(t_1 t_2)$

Bigram Similarity Results

J. Mitchell, M. Lapata/Cognitive Science 34 (2010)

Table 6
Correlation coefficients of model predictions with subject similarity ratings (Spearman's ρ) using a simple semantic space

Model	Adjective–Noun	Noun–Noun	Verb–Object
Additive	.36	.39	.30
Kintsch	.32	.22	.29
Multiplicative	.46	.49	.37
Tensor product	.41	.36	.33
Convolution	.09	.05	.10
Weighted additive	.44	.41	.34
Dilation	.44	.41	.38
Target unit	.43	.34	.29
Head only	.43	.17	.24
Humans	.52	.49	.55

Why does multiplication work?

- these vectors are built from co-occurrence counts (like in the first part of Assignment 2)
- so element-wise multiplication is like performing an AND operation on context counts

- when using skip-gram word vectors (or other neural network-derived vectors), addition often works better

A Comparison of Vector-based Representations for Semantic Composition

William Blacoe and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
w.b.blacoe@sms.ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

In this paper we address the problem of modeling compositional meaning for phrases and sentences using distributional methods. We experiment with several possible combinations of representation and composition, exhibiting varying degrees of sophistication. Some are shallow while others operate over syntactic structure, rely on parameter learning, or require access to very large corpora. We find that shallow approaches are as good as more computationally intensive alternatives with regards to two particular tests: (1) phrase similarity and (2) paraphrase detection. The sizes of the involved training corpora and the

word sense discrimination (Schütze, 1998), language modeling (Bellegarda, 2000), and the identification of analogical relations (Turney, 2006).

While much research has been directed at the most effective ways of constructing representations for individual words, there has been far less consensus regarding the representation of larger constructions such as phrases and sentences. The problem has received some attention in the connectionist literature, particularly in response to criticisms of the ability of connectionist representations to handle complex structures (Smolensky, 1990; Plate, 1995). More recently, several proposals have been put forward for computing the meaning of word combina-

Results

	dim.	c.m.	Adj-N	N-N	V-Obj
SDS (BNC)	2000	+	0.37	0.38	0.28
	2000	⊙	**0.48**	**0.50**	**0.35**
	100	RAE	0.31	0.30	0.28
NLM (BNC)	50	+	0.28	0.26	0.24
	50	⊙	0.26	0.22	0.18
	100	RAE	0.19	0.24	0.28

Table 3: Correlation coefficients of model predictions with subject similarity ratings (Spearman's ρ); columns show dimensionality: fixed or varying (see Section 2.1), composition method: $+$ is additive vector composition, $\odot$ is component-wise multiplicative vector composition, RAE is Socher et al. (2011a)'s recursive auto-encoder.

SDS = simple distributional semantic

NLM = neural language model

	Topical	Paraphrastic
Bigrams	BigramSim (Mitchell and Lapata, 2010)	

Topical

Paraphrastic

Bigrams

BigramSim

(Mitchell and Lapata, 2010)

		BigramSim
television programme	tv set	5.8
training programme	education course	5.7
bedroom window	education officer	1.3

Topical

Paraphrastic

Bigrams

BigramSim (Mitchell and Lapata, 2010)

BigramPara (Wieting et al., 2015)

		BigramSim	BigramPara
television programme	tv set	5.8	1.0
training programme	education course	5.7	5.0
bedroom window	education officer	1.3	1.0

Topical

Paraphrastic

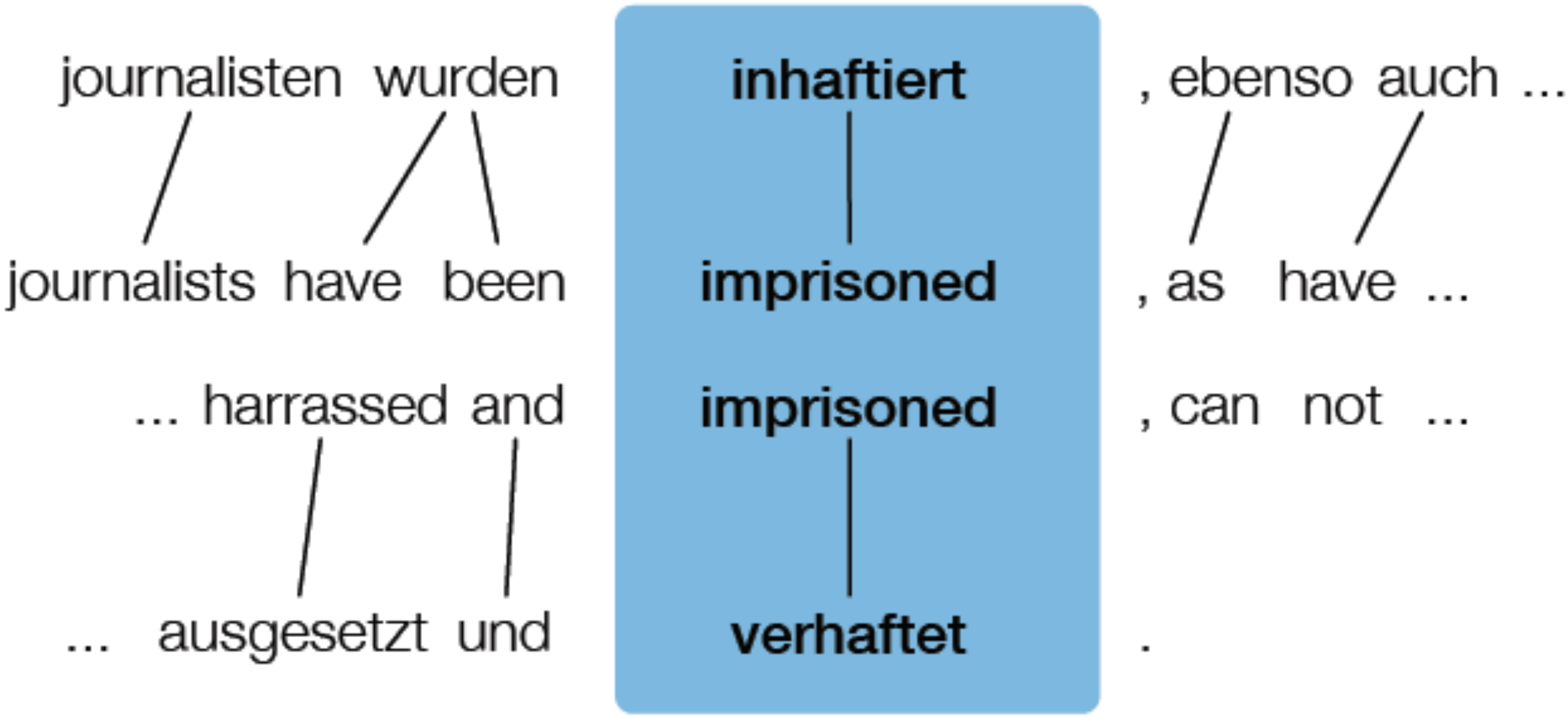
		PhrasePara
can not be separated from	is inseparable from	5.0
hoped to be able to	looked forward to	3.4
come on , think about it	people , please	2.2
how do you mean that	what worst feelings	1.6

Phrases

PhrasePara
(Wieting et al., 2015)

Training Data: Paraphrase Database

(Ganitkevitch, Van Durme, and Callison-Burch, 2013)



from Ganitkevitch and Callison-Burch (2014)

- currently there is a lot of work on designing functional architectures for bigram, phrase, and sentence similarity
  - e.g., word averaging, recurrent neural networks, LSTMs, recursive neural networks, etc.
- our recent results find that, for sentence similarity, word averaging is a surprisingly strong baseline

Towards Universal Paraphrastic Sentence Embeddings

John Wieting **Mohit Bansal** **Kevin Gimpel** **Karen Livescu**
Toyota Technological Institute at Chicago, Chicago, IL, 60637, USA
{jwieting,mbansal,kgimpel,klivescu}@ttic.edu

Model	Pavlick et al. (2015) (test)
PARAGRAM-PHRASE	60.0
iRNN	60.0
projection	58.4
DAN	60.1
RNN	60.3
LSTM (o.g.)	60.9
LSTM (no o.g.)	**61.3**
skip-thought	39.3
GloVe	44.8
PARAGRAM-SL999	55.3

on similar data to training data, LSTM does best

word averaging

adding layers to word averaging

Dataset	50%	75%	Max	PP	iRNN	proj.	DAN	RNN	LSTM (o.g.)	LSTM (no o.g.)	ST	GloVe	PSL
STS 2012 Average	54.5	59.5	70.3	58.7	58.4	**60.0**	56.0	48.1	46.4	51.0	30.8	52.5	52.8
STS 2013 Average	45.3	51.4	65.3	55.8	56.7	**56.8**	54.2	44.7	41.5	45.2	24.8	42.3	46.4
STS 2014 Average	64.7	71.4	76.7	70.9	70.9	**71.3**	69.5	57.7	51.5	59.8	31.4	54.2	59.5
STS 2015 Average	70.2	75.8	80.2	**75.8**	75.6	74.8	72.7	57.2	56.0	63.9	31.0	52.7	60.0
2014 SICK	71.4	79.9	82.8	71.6	71.2	**71.6**	70.7	61.2	59.0	63.9	49.8	65.9	66.4
2015 Twitter	49.9	52.5	61.9	52.9	52.9	52.8	**53.7**	45.1	36.1	47.6	24.7	30.3	36.3

but when evaluating on other datasets, word averaging models do best!

“You can’t cram the meaning of a whole %&!$# sentence into a single $&!#* vector!”

--Ray Mooney

“You can’t map all sentences into a cold, sterile space of meaningless, uninterpretable dimensions. Symbolic representations can encode meaning much more efficiently.”

--my interpretation

"You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!"

--Ray Mooney

Why must we choose?

Neural architectures for text understanding can combine discrete (symbolic) and continuous representations

Syntax and Semantics

- **syntax**: rules, principles, processes that govern sentence structure of a language
- **semantics**: what the sentence means

- we saw syntactic parsing, which produces a syntactic structure of a sentence
  - helps to disambiguate attachments, coordinations, sometimes word sense
- now we’ll look at semantic parsing, which roughly means “produce a semantic structure of a sentence”

Several Kinds of Semantic Parsing

- semantic role labeling (SRL)
- frame-semantic parsing
- “semantic parsing” (first-order logic)
- abstract meaning representation (AMR)
- dependency-based compositional semantics

Semantic Role Labeling

Who did *what* to *whom* at *where?*

The police officer detained the suspect at the scene of the crime

Agent Predicate Theme Location

J&M/SLP3

Can we figure out that these have the same meaning?

XYZ corporation **bought** the stock.

They **sold** the stock to XYZ corporation.

The stock was **bought** by XYZ corporation.

The **purchase** of the stock by XYZ corporation...

The stock **purchase** by XYZ corporation...

A Shallow Semantic Representation: Semantic Roles

Predicates (bought, sold, purchase) represent an **event**

semantic roles express the abstract role that arguments of a predicate can take in the event

More specific ←→ More general

buyer **agent** **proto-agent**

J&M/SLP3

Getting to semantic roles

Neo-Davidsonian event representation:

Sasha broke the window

Pat opened the door

$$\exists e,x,y\ Breaking(e) \wedge Breaker(e, Sasha) \wedge BrokenThing(e,y) \wedge Window(y)$$

$$\exists e,x,y\ Opening(e) \wedge Opener(e, Pat) \wedge OpenedThing(e,y) \wedge Door(y)$$

Subjects of break and open: **Breaker** and **Opener**

Deep roles specific to each event (breaking, opening)

Hard to reason about them for applications like QA

Thematic roles

- **Breaker** and **Opener** have something in common!
  - Volitional actors
  - Often animate
  - Direct causal responsibility for their events
- Thematic roles are a way to capture this semantic commonality between *Breakers* and *Eaters*
  - they are both AGENTS
- The *BrokenThing* and *OpenedThing* are THEMES.
  - prototypically inanimate objects affected in some way by the action

A Typical Set of Thematic Roles

Thematic Role	Definition	Example
AGENT	The volitional causer of an event	*The waiter* spilled the soup.
EXPERIENCER	The experiencer of an event	*John* has a headache.
FORCE	The non-volitional causer of the event	*The wind* blows debris from the mall into our yards.
THEME	The participant most directly affected by an event	Only after Benjamin Franklin broke *the ice*...
RESULT	The end product of an event	The city built a *regulation-size baseball diamond*...
CONTENT	The proposition or content of a propositional event	Mona asked *"You met Mary Ann at a supermarket?"*
INSTRUMENT	An instrument used in an event	He poached catfish, stunning them *with a shocking device*...
BENEFICIARY	The beneficiary of an event	Whenever Ann Callahan makes hotel reservations *for her boss*
SOURCE	The origin of the object of a transfer event	I flew in *from Boston*.
GOAL	The destination of an object of a transfer event	I drove *to Portland*.

J&M/SLP3

Problems with Thematic Roles

Hard to create standard set of roles or formally define them

Often roles need to be fragmented to be defined.

- Levin and Rappaport Hovav (2015): two kinds of INSTRUMENTS
  - **intermediary instruments** that can appear as subjects
    - The cook opened the jar with the new gadget.
    - The new gadget opened the jar.
  - **enabling instruments** that cannot
    - Shelly ate the sliced banana with a fork.
    - *The fork ate the sliced banana.

Alternatives to thematic roles

1. **Fewer roles**: generalized semantic roles, defined as prototypes (Dowty 1991)

   PROTO-AGENT

   PROTO-PATIENT

   PropBank

2. **More roles**: Define roles specific to a group of predicates

   FrameNet

Semantic role labeling (SRL)

- The task of finding the semantic roles of each argument of each predicate in a sentence.
- FrameNet versus PropBank:

[You]	can't	[blame]	[the program]	[for being unable to identify it]
COGNIZER		TARGET	EVALUEE	REASON

[The San Francisco Examiner]	issued	[a special edition]	[yesterday]
ARG0	TARGET	ARG1	ARGM-TMP

History

- semantic roles as a intermediate semantics, used early in
  - machine translation (Wilks, 1973)
  - question-answering (Hendrix et al., 1973)
  - spoken-language understanding (Nash-Webber, 1975)
  - dialogue systems (Bobrow et al., 1977)
- early SRL systems

  Simmons 1973, Marcus 1980:

    - parser followed by hand-written rules for each verb
    - dictionaries with verb-specific case frames (Levin 1977)

Why Semantic Role Labeling?

- A useful shallow semantic representation
- Improves NLP tasks like:
  - question answering

    Shen and Lapata 2007, Surdeanu et al. 2011
  - machine translation

    Liu and Gildea 2010, Lo et al. 2013

PropBank

- Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106

PropBank Roles

Following Dowty 1991

Proto-Agent

- Volitional involvement in event or state
- Sentience (and/or perception)
- Causes an event or change of state in another participant
- Movement (relative to position of another participant)

Proto-Patient

- Undergoes change of state
- Causally affected by another participant
- Stationary relative to movement of another participant

PropBank Roles

- Following Dowty 1991
  - Role definitions determined verb by verb, with respect to the other roles
  - Semantic roles in PropBank are thus verb-sense specific.
- Each verb sense has numbered argument: Arg0, Arg1, Arg2,…

  Arg0: PROTO-AGENT

  Arg1: PROTO-PATIENT

  Arg2: usually: benefactive, instrument, attribute, or end state

  Arg3: usually: start point, benefactive, instrument, or attribute

  Arg4 the end point

  (Arg2-Arg5 are not really that consistent, causes a problem for labeling)

PropBank Frame Files

agree.01
Arg0: Agreer
Arg1: Proposition
Arg2: Other entity agreeing

Ex1: $[_{\text{Arg0}}$ The group] *agreed* $[_{\text{Arg1}}$ it wouldn't make an offer].
Ex2: $[_{\text{ArgM-TMP}}$ Usually] $[_{\text{Arg0}}$ John] *agrees* $[_{\text{Arg2}}$ with Mary] $[_{\text{Arg1}}$ on everything].

Advantage of a ProbBank Labeling

increase.01 "go up incrementally"

Arg0: causer of increase
Arg1: thing increasing
Arg2: amount increased by, EXT, or MNR
Arg3: start point
Arg4: end point

This would allow us to see the commonalities in these 3 sentences:

[$_{\text{Arg0}}$ Big Fruit Co.] increased [$_{\text{Arg1}}$ the price of bananas].
[$_{\text{Arg1}}$ The price of bananas] was increased again [$_{\text{Arg0}}$ by Big Fruit Co.]
[$_{\text{Arg1}}$ The price of bananas] increased [$_{\text{Arg2}}$ 5%].

Modifiers or adjuncts of the predicate: Arg-M

ArgM-TMP	when?	yesterday evening, now
LOC	where?	at the museum, in San Francisco
DIR	where to/from?	down, to Bangkok
MNR	how?	clearly, with much enthusiasm
PRP/CAU	why?	because ... , in response to the ruling
REC		themselves, each other
ADV	miscellaneous	
PRD	secondary predication	...ate the meat raw

J&M/SLP3

Capturing descriptions of the same event by different nouns/verbs

[$_{\text{Arg1}}$ The price of bananas] increased [$_{\text{Arg2}}$ 5%].
[$_{\text{Arg1}}$ The price of bananas] rose [$_{\text{Arg2}}$ 5%].
There has been a [$_{\text{Arg2}}$ 5%] rise [$_{\text{Arg1}}$ in the price of bananas].

FrameNet

- Baker et al. 1998, Fillmore et al. 2003, Fillmore and Baker 2009, Ruppenhofer et al. 2006
- Roles in PropBank are specific to a verb
- Role in FrameNet are specific to a **frame: a** background knowledge structure that defines a set of frame-specific semantic roles, called **frame elements**,
  - includes a set of predicates that use these roles
  - each word evokes a frame and profiles some aspect of the frame

“Change position on a scale” Frame

frame consists of words that indicate change of ITEM’s position on a scale (the ATTRIBUTE) from starting point (INITIAL VALUE) to end point (FINAL VALUE)

[$_{\text{ITEM}}$ Oil] *rose* [$_{\text{ATTRIBUTE}}$ in price] [$_{\text{DIFFERENCE}}$ by 2%].

[$_{\text{ITEM}}$ It] has *increased* [$_{\text{FINAL\_STATE}}$ to having them 1 day a month].

[$_{\text{ITEM}}$ Microsoft shares] *fell* [$_{\text{FINAL\_VALUE}}$ to 7 5/8].

[$_{\text{ITEM}}$ Colon cancer incidence] *fell* [$_{\text{DIFFERENCE}}$ by 50%] [$_{\text{GROUP}}$ among men].

steady *increase* [$_{\text{INITIAL\_VALUE}}$ from 9.5] [$_{\text{FINAL\_VALUE}}$ to 14.3] [$_{\text{ITEM}}$ in dividends]

[$_{\text{DIFFERENCE}}$ 5%] [$_{\text{ITEM}}$ dividend] *increase*...

J&M/SLP3

“Change position on a scale” Frame

VERBS:	dwindle	move	soar	escalation	shift
advance	edge	mushroom	swell	explosion	tumble
climb	explode	plummet	swing	fall	
decline	fall	reach	triple	fluctuation	**ADVERBS:**
decrease	fluctuate	rise	tumble	gain	increasingly
diminish	gain	rocket		growth	
dip	grow	shift	**NOUNS:**	hike	
double	increase	skyrocket	decline	increase	
drop	jump	slide	decrease	rise	

“Change position on a scale” Frame

Core Roles	
ATTRIBUTE	The ATTRIBUTE is a scalar property that the ITEM possesses.
DIFFERENCE	The distance by which an ITEM changes its position on the scale.
FINAL_STATE	A description that presents the ITEM’s state after the change in the ATTRIBUTE’s value as an independent predication.
FINAL_VALUE	The position on the scale where the ITEM ends up.
INITIAL_STATE	A description that presents the ITEM’s state before the change in the ATTRIBUTE’s value as an independent predication.
INITIAL_VALUE	The initial position on the scale from which the ITEM moves away.
ITEM	The entity that has a position on the scale.
VALUE_RANGE	A portion of the scale, typically identified by its end points, along which the values of the ATTRIBUTE fluctuate.
Some Non-Core Roles	
DURATION	The length of time over which the change takes place.
SPEED	The rate of change of the VALUE.
GROUP	The GROUP in which an ITEM changes the value of an ATTRIBUTE in a specified way.