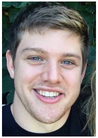
Enriching distributional linguistic representations with structured resources

Christopher Potts

Stanford Linguistics

Berkeley NLP Group, October 30, 2017

Ignacio Cases



Ben Lengerich

Andrew Maas

Central question for today

Central question for today

- Distributional representations are powerful and easy to obtain, but they tend to reflect only similarity (synonymy, connotation).

Central question for today

- Distributional representations are powerful and easy to obtain, but they tend to reflect only similarity (synonymy, connotation).
- Structured resources are sparse and hard to obtain, but they support learning rich, diverse semantic distinctions.

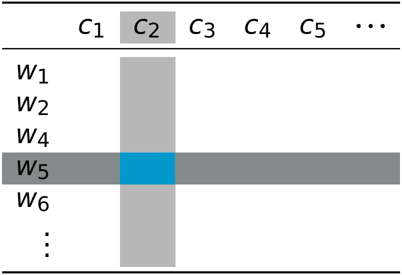
Central question for today

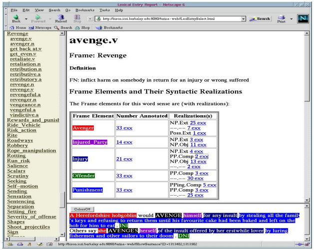
- Distributional representations are powerful and easy to obtain, but they tend to reflect only similarity (synonymy, connotation).
- Structured resources are sparse and hard to obtain, but they support learning rich, diverse semantic distinctions.
- Can we have the best aspects of both?

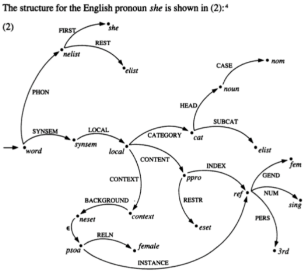
Distributed representations

	c_1	c_2	c_3	c_4	c_5	$\cdots$
w_1						
w_2						
w_4						
w_5						
w_6						
$\vdots$						

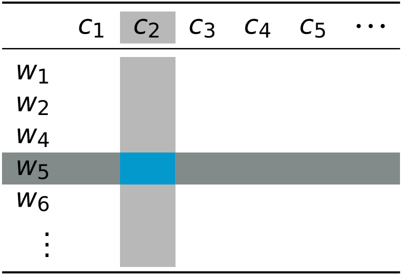
Distributed representations



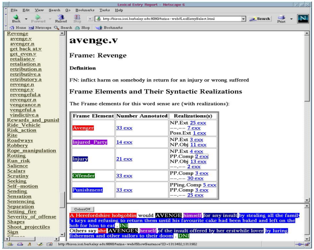


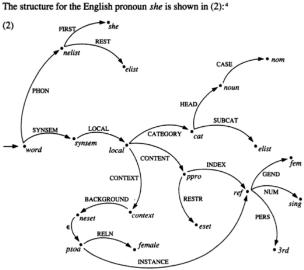


Distributed representations



The stock deteriorated.





Design choices

Matrix type	Weighting	Dim. reduction	Comparison
word × document	Probabilities	LSA	Euclidean
word × word	TF-IDF	PLSA	Cosine
adj. × modified noun	Observed/Expected	LDA	Dice
word × dependency	PMI	PCA	Jaccard
verb × arguments	Positive PMI	DCA	KL
⋮	⋮	⋮	⋮

Design choices

tokenization
annotation
tagging
parsing
feature selection
⋮ cluster texts by date/author/discourse context/. . .
↓ ↙

Matrix type	Weighting	Dim. reduction	Comparison
word × document	Probabilities	LSA	Euclidean
word × word	TF-IDF	PLSA	Cosine
adj. × modified noun	Observed/Expected	LDA	Dice
word × dependency	PMI	PCA	Jaccard
verb × arguments	Positive PMI	DCA	KL
⋮	⋮	⋮	⋮

Design choices

tokenization
annotation
tagging
parsing
feature selection
⋮ cluster texts by date/author/discourse context/...
↓ ↙

Matrix type	Weighting	Dim. reduction	Comparison
word × document	Probabilities	LSA	Euclidean
word × word	TF-IDF	PLSA	Cosine
adj. × modified noun	Observed/Expected	LDA	Dice
word × dependency	PMI	PCA	Jaccard
verb × arguments	Positive PMI	DCA	KL
⋮	⋮	⋮	⋮

LBL, word2vec, GloVE, etc.

David Lewis on truth and Markerese

David Lewis, ‘General semantics’ (1970)

David Lewis on truth and Markerese

Semantic markers are symbols: items in the vocabulary of an artificial language we may call Semantic Markerese. Semantic interpretation by means of them amounts merely to a translation algorithm from the object language to the auxiliary language Markerese.

David Lewis, 'General semantics' (1970)

David Lewis on truth and Markerese

Semantic markers are symbols: items in the vocabulary of an artificial language we may call Semantic Markerese. Semantic interpretation by means of them amounts merely to a translation algorithm from the object language to the auxiliary language Markerese. But we can know the Markerese translation of an English sentence without knowing the first thing about the meaning of the English sentence: namely, the conditions under which it would be true.

David Lewis, 'General semantics' (1970)

David Lewis on truth and Markerese

Semantic markers are symbols: items in the vocabulary of an artificial language we may call Semantic Markerese. Semantic interpretation by means of them amounts merely to a translation algorithm from the object language to the auxiliary language Markerese. But we can know the Markerese translation of an English sentence without knowing the first thing about the meaning of the English sentence: namely, the conditions under which it would be true. Semantics with no treatment of truth conditions is not semantics.

David Lewis, 'General semantics' (1970)

The meaning of life

Foreword

In the spring of 1976, Terry Parsons and Barbara Partee taught a course on Montague grammar, which I attended. On the second to the final day of class, Terry went around the room asking the students if there were any questions at all that remained unanswered, and promised to answer them on the last day of class. I asked if he really meant ANY question at all, which he emphatically said that he meant. As I had encountered a few questions in my lifetime that remained at least partially unresolved, I decided to ask one of them. What is life? What is the meaning of life? After all, Barbara and Terry had promised to provide answers to any question at all.

On the final day of class Barbara wore her Montague grammar T-shirt, and she and Terry busied themselves answering our questions. At long last, they came to my question. I anticipated a protracted and involved answer, but their reply was crisp and succinct. First Barbara, chalk in hand, showed me the meaning of life.

$$^\wedge\underline{\text{life}}'$$

Terry then stepped up and showed me what life really is.

$$^\vee{}^\wedge\underline{\text{life}}'$$

As we were asked to show on a homework assignment earlier in the year, this is equivalent to: $\underline{\text{life}}'$.

Leaving me astounded that I had been living in such darkness for all these years, the class then turned to the much stickier problem of pronouns.

Jerrold Katz on meaning

The arbitrariness of the distinction between form and matter reveals itself [. . .]

Jerrold J. Katz, *Semantic Theory* (1972)

Jerrold Katz on meaning

> *The arbitrariness of the distinction between form and matter reveals itself [. . .]*

The question “What is meaning?” broken down:

- What is synonymy?
- What is antonymy?
- What is superordination?
- What is semantic ambiguity?
- What is semantic truth (analyticity, metalinguistic, etc.)?
- What is a possible answer to a question?
- . . .

Jerrold J. Katz, *Semantic Theory* (1972)

Purely distributional representations

Purely distributional representations

- High-dimensional 😃

Purely distributional representations

- High-dimensional 😃
- Meaning from dense linguistic inter-relationships 😃

Purely distributional representations

- High-dimensional 😃
- Meaning from dense linguistic inter-relationships 😃
- Meaning *solely* from (nth-order) co-occurrence 😟

Purely distributional representations

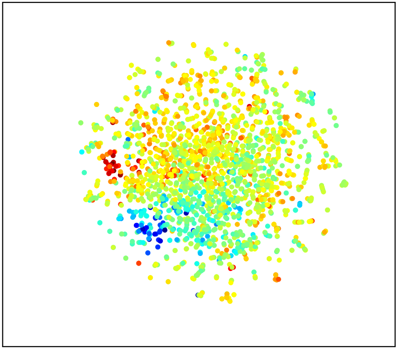
- High-dimensional 😃
- Meaning from dense linguistic inter-relationships 😃
- Meaning *solely* from (*n*th-order) co-occurrence 😟
- No grounding in physical or social contexts 😟

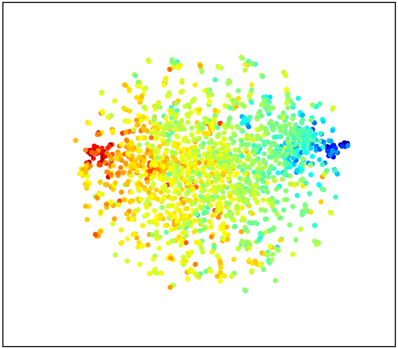
Purely distributional representations

- High-dimensional 😃
- Meaning from dense linguistic inter-relationships 😃
- Meaning *solely* from (*n*th-order) co-occurrence 😟
- No grounding in physical or social contexts 😟
- Not symbolic 😕

Grounding via supervision

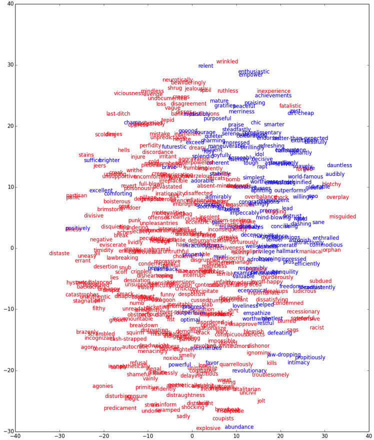
Word vectors to maximize unsupervised log-likelihood of words given documents and supervised prediction accuracy:

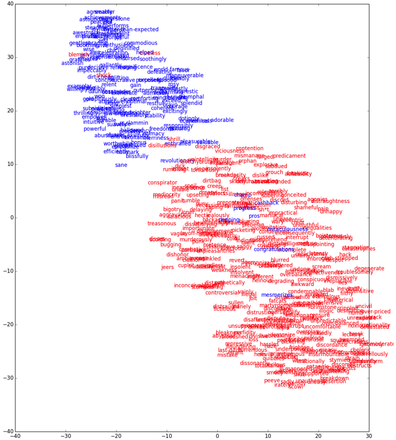




Maas et al., ‘Learning word vectors for sentiment analysis’ (2011)

Hidden representations from a deep classifier





Retrofitting

Faruqui et al., ‘Retrofitting word vectors to semantic lexicons’ (2015)

Faruqui et al.: Retrofitting with identity relations

$$\sum_{i \in \mathcal{V}} \alpha_i \left\| \boldsymbol{q}_i - \hat{\boldsymbol{q}}_i \right\|^2 + \sum_{(i,j,r) \in \mathcal{E}} \beta_{ij} \left\| \boldsymbol{q}_i - \boldsymbol{q}_j \right\|^2$$

- Balances fidelity to the original vector $\hat{\boldsymbol{q}}_i$
- against looking more like one's graph neighbors.
- Forces are balanced with $\alpha = 1$ and $\beta = \frac{1}{\textbf{Degree}(i)}$

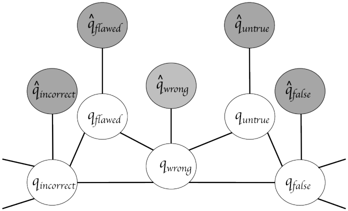


Figure 1: Word graph with edges between related words showing the observed (grey) and the inferred (white) word vector representations.

See also Hamilton et al., 'Inductive representation learning on large graphs' (2017)

What retrofitting to WordNet might do

What retrofitting to WordNet might do

- Cluster *mammal* with *dog* and *puppy* even though *mammal* has a different, unusual distributional profile.

What retrofitting to WordNet might do

- Cluster *mammal* with *dog* and *puppy* even though *mammal* has a different, unusual distributional profile.
- Avoid polarity mistakes like modeling *superb* and *awful* as similar (though beware those antonym edges!).

What retrofitting to WordNet might do

- Cluster *mammal* with *dog* and *puppy* even though *mammal* has a different, unusual distributional profile.
- Avoid polarity mistakes like modeling *superb* and *awful* as similar (though beware those antonym edges!).
- Holistic consistency:

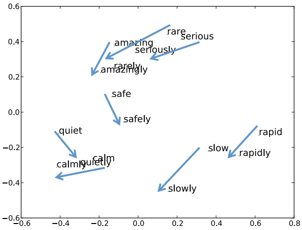
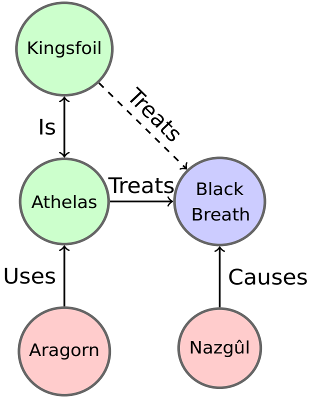


Figure 3: Two-dimensional PCA projections of 100-dimensional SG vector pairs holding the “adjective to adverb” relation, before (left) and after (right) retrofitting.

Concerns about identity retrofitting

- No attention to edge semantics; edges mean ‘similar to’.
- Presupposes a uniform initial embedding space
- No modeling of missing edges



Hand-build functions from Mrkšić et al.

- AntonymRepel:

$$\sum_{(i,j)\in A} \mathrm{ReLU}\left(1.0 - d(\boldsymbol{q}_i, \boldsymbol{q}_j)\right)$$

- SynonymAttract:

$$\sum_{(i,j)\in S} \mathrm{ReLU}\left(d(\boldsymbol{q}_i, \boldsymbol{q}_j) - 0\right)$$

- VectorSpacePreservation:

$$\sum_{i} \sum_{j\in N(i)} \mathrm{ReLU}\left(d(\boldsymbol{q}_i, \boldsymbol{q}_j) - d(\hat{\boldsymbol{q}}_i, \hat{\boldsymbol{q}}_j)\right)$$

Mrkšić et al., ‘Counter-fitting word vectors to linguistic constraints‘ (2017)

Retrofitting with functional relations

Lengerich et al. ‘Retrofitting distributional embeddings to knowledge graphs with functional relations’ (2017)

The framework

$$\sum_{i \in \mathcal{V}} \alpha_i \left\| \boldsymbol{q}_i - \hat{\boldsymbol{q}}_i \right\|^2 + \sum_{(i,j,r) \in \mathcal{E}} \beta_{ijr} f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) - \sum_{(i,j,r) \in \mathcal{E}^-} \beta_{ijr} f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) + \lambda \sum_r \rho(f_r)$$

Instantiations

Our framework

$$\sum_{i \in \mathcal{V}} \alpha_i \left\| \boldsymbol{q}_i - \hat{\boldsymbol{q}}_i \right\|^2 + \sum_{(i,j,r) \in \mathcal{E}} \beta_{ijr} f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) - \sum_{(i,j,r) \in \mathcal{E}^-} \beta_{ijr} f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) + \lambda \sum_r \rho(f_r)$$

Faruqui et al.

$$f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) = \left\| \boldsymbol{q}_i - \boldsymbol{q}_j \right\|^2$$

with $\beta_{ijr} = 0$

Instantiations

Our framework

$$\sum_{i \in \mathcal{V}} \alpha_i \left\| \boldsymbol{q}_i - \hat{\boldsymbol{q}}_i \right\|^2 + \sum_{(i,j,r) \in \mathcal{E}} \beta_{ijr} f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) - \sum_{(i,j,r) \in \mathcal{E}^-} \beta_{ijr} f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) + \lambda \sum_r \rho(f_r)$$

Linear

$$f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) = \left\| \boldsymbol{A_r} \boldsymbol{q}_j + \boldsymbol{b_r} - \boldsymbol{q}_i \right\|^2$$

- $\rho(f_r) = \|\boldsymbol{A_r}\|^2$
- We initialize $\boldsymbol{A_r} = \mathbf{1}$ and $\boldsymbol{b_r} = \mathbf{0}$
- Initialization can be different for different relations, e.g., $\boldsymbol{A}_{\text{antonym}} = \mathbf{-1}$

Instantiations

Our framework

$$\sum_{i \in \mathcal{V}} \alpha_i \left\| \boldsymbol{q}_i - \hat{\boldsymbol{q}}_i \right\|^2 + \sum_{(i,j,r) \in \mathcal{E}} \beta_{ijr} f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) - \sum_{(i,j,r) \in \mathcal{E}^-} \beta_{ijr} f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) + \lambda \sum_r \rho(f_r)$$

Simplest neural (akin to Latent Factor Models)

$$f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) = \tanh(\boldsymbol{q}_i^\top \boldsymbol{A_r} \boldsymbol{q}_j)$$

Instantiations

Our framework

$$\sum_{i \in \mathcal{V}} \alpha_i \left\| \boldsymbol{q}_i - \hat{\boldsymbol{q}}_i \right\|^2 + \sum_{(i,j,r) \in \mathcal{E}} \beta_{ijr} f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) - \sum_{(i,j,r) \in \mathcal{E}^-} \beta_{ijr} f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) + \lambda \sum_r \rho(f_r)$$

Neural Tensor Network

$$f_r(\boldsymbol{q}_i, \boldsymbol{q}_j) = \boldsymbol{u_r}^\top \tanh(\boldsymbol{q}_i^\top \mathcal{A}_r \boldsymbol{q}_j)$$

where $\mathcal{A}_r \in \mathbb{R}^{d \times d \times k}$ and $\rho(f_r) = \|\mathcal{A}_r\|^2 + \|\boldsymbol{u_r}\|^2$

Graph embedding penalty functions

TransE

$$f_r(\boldsymbol{q_i}, \boldsymbol{q_j}) = \left\|\boldsymbol{q_i} + \boldsymbol{a_r} - \boldsymbol{q_j}\right\|_2^2$$

Faruqui et al.'s model is the special case where $\boldsymbol{a_r} = \mathbf{0}$

Bordes et al. 'Translating embeddings for modeling multi-relational data' (2013)

Graph embedding penalty functions

TransH

$$f_r(\boldsymbol{q_i}, \boldsymbol{q_j}) = \left\| g_r(\boldsymbol{q_i}) + \boldsymbol{a_r} - g_r(\boldsymbol{q_j}) \right\|_2^2$$
$$g_r(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{w_r}^T \boldsymbol{x} \boldsymbol{w_r}$$

Wang et al. ‘Knowledge graph embedding by translating on hyperplanes‘ (2014)

Graph embedding penalty functions

TransR

$$f_r(\boldsymbol{q_i}, \boldsymbol{q_j}) = \left\| \boldsymbol{q_i M_r} + \boldsymbol{a} - \boldsymbol{q_j M_r} \right\|_2^2$$

Lin et al. ‘Learning entity and relation embeddings for knowledge graph completion‘ (2015)

Experimental paradigm: Edge prediction

When predicting edge type r:

1. Retrofit to a graph containing all edge-types except r.
2. Train a classifier to predict r from the concatenation of the two nodes' representations.
3. Training set uses 70% of r's edges; the rest are for testing.
4. Both train and test sets are balanced with an equivalent number of non-edges.

FrameNet evaluation

Model	'Inheritance' (2132/992)	'Using' (1552/668)	'Reframing' (544/312)	'Subframe' (356/168)	'Perspective On' (336/148)
None	87.58	88.59	85.60	91.24	89.59
Faruqui et al.	90.79	87.87	87.02	94.50	<u>94.24</u>
FR-Linear	**92.92**	<u>92.04</u>	<u>89.37</u>	<u>94.65</u>	**94.73**
FR-Neural	<u>92.46</u>	**92.54**	**89.57**	**95.65**	94.04

Model	'Precedes' (220/136)	'See Also' (268/76)	'Causative Of' (204/36)	'Inchoative Of' (60/16)
None	<u>87.30</u>	85.11	86.11	82.50
Faruqui et al.	85.26	83.81	84.49	78.33
FR-Linear	87.00	<u>91.93</u>	<u>92.09</u>	<u>82.50</u>
FR-Neural	**89.16**	**93.25**	**94.33**	**85.00**

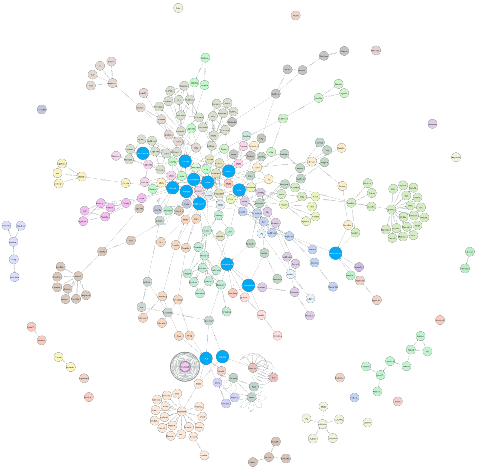
WordNet evaluations

Model	Word Similarity			Syntactic Relation
	WordSim-353	MTurk-771	MTurk-287	Google Analogy
None	0.512	0.538	0.671	0.772
Faruqui et al.	0.512	0.532	0.664	0.774
FR-Linear	**0.542**	**0.562**	**0.679**	**0.793**
FR-Neural	0.516	0.543	0.676	0.784

The Roam Core Public Health Knowledge Graph

- Diverse medical ontologies
- Provider profiles and networks
- Product approvals, recalls, adverse events
- County-level population and health stats
- Municipal and public-policy data
- Academic publications
- Clinical Trials summaries and stats
- Financial data

250 million nodes; 1 billion edges; 6 billion properties



Evaluation on the drug–disease subgraph

Entity Type	Count
Drug	223,019
Disease	95,559

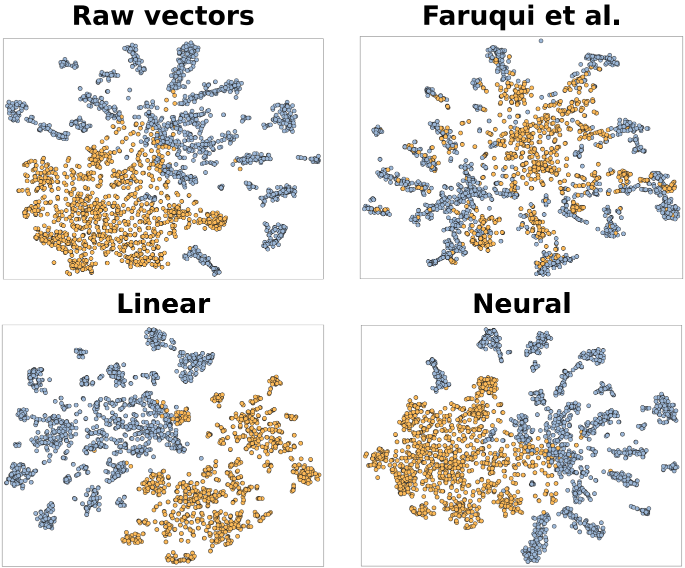
Edge Type	Connects	Count
Ingredient Of	Drug → Drug	49,218
Has Ingredient	Drug → Drug	49,208
Is A	Drug → Drug	28,297
Has Descendent	Disease → Disease	22,344
Treats	Drug → Disease	19,374
Has Active Ingredient	Drug → Drug	18,422
Has Child	Disease → Disease	18,066
Active Ingredient Of	Drug → Drug	17,175
Has TradeName	Drug → Drug	11,783
TradeName Of	Drug → Drug	11,783
Inverse Is A	Drug → Drug	10,369
Has Symptom	Disease → Disease	7,892
Part Of	Drug → Drug	6,882
Has Part	Drug → Drug	6,624
Same As	Drug → Drug	5,882
Precise Ingredient Of	Drug → Drug	3,562
Has Precise Ingredient	Drug → Drug	3,562
Possibly Equivalent To	Drug → Drug	1,233
Causative Agent of	Drug → Drug	1,070
Has Form	Drug → Drug	602
Form of	Drug → Drug	602
Component of	Drug → Drug	436
Includes	Disease → Disease	347
Has Dose Form	Drug → Drug	138

Disease representations from clinical text

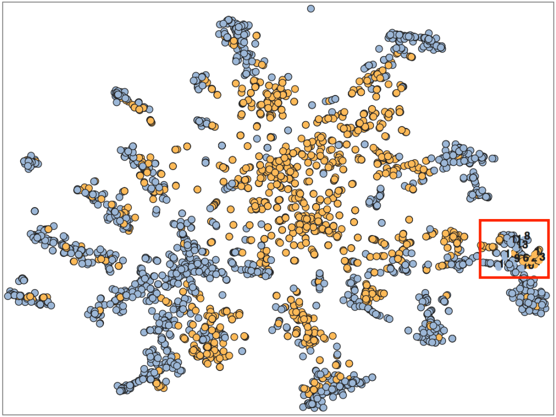
INDICATIONS FOR PROCEDURE: This is a 66-year-old female with past medical history of morbid obesity, obstructive sleep apnea, asthma, hypertension, and osteoarthritis who presents for revision of her previous bariatric surgery. The patient underwent vertical banded gastroplasty in 2000; however, had recurrent weight gain. The patients current BMI is 71. [. . .]

HISTORY OF PRESENT ILLNESS: The patient is a 51-year-old African American female postoperative day #1 status post sleeve gastrectomy. She has a history of hypertension, hyperlipidemia, chronic back pain, GERD, and previous laparoscopic band placement, which was later removed. [. . .]

A look at the embeddings with t-SNE

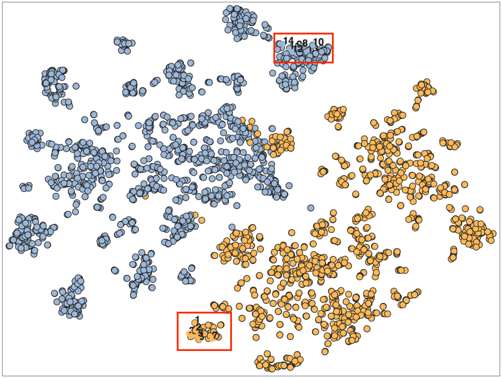


A closer look at the Faruqui et al. embeddings



1. Schizoaffective disorder, unspecified
2. Schizophrenia
3. Drug induced dystonia
4. Mild intellectual disabilities
5. Mixed anxiety and depressive disorder
6. Nonpsychotic mental disorder, unspecified
7. Panic disorder without agoraphobia
8. edluar
9. norpramin
10. imipramine pamoate
11. diethylpropion
12. buprenorphine/naloxone
13. lithium
14. pamelor

A closer look at the linear embeddings



1. Dissociative amnesia
2. Major depressv disord, single epsd, sev w/o psych features
3. Hypersomnia not due to a substance or known physiol cond
4. Insomnia
5. Schizoaffective disorder, unspecified
6. Schizophrenia
7. Conduct disorder, unspecified
8. naltrexone
9. amphetamine
10. geodon
11. chlorpromazine
12. haldol
13. vyvanse
14. lithobid

Drug–disease link prediction accuracies

Model	'Treats' (9152/2490)
None	72.02 ± 0.50
FR-Identity	72.93 ± 0.82
FR-Linear	$\mathbf{84.22} \pm \mathbf{0.82}$
FR-Neural	$\underline{73.52 \pm 0.89}$

Knowledge discovery

Model	Drug	Disease Target	Plausible
None	Naproxen	Ankylosing Spondylitis	Y
	Latanoprost	Superficial injury of ankle, foot and toes	N
	Pulmicort	Psoriasis, unspecified	Y
	Furosemide	Aneurysm of unspecified site	Y
	Desonide	Chlamydial lymphogranuloma (venereum)	N
FR-Identity	Latanoprost	Superficial injury of ankle, foot and toes	N
	Elixophyllin	Pneumonia in diseases classified elsewhere	Y
	Furosemide	Aneurysm of unspecified site	Y
	Oxistat	Mycosis fungoides	Y
	Trifluridine	Congenital Pneumonia	N
FR-Linear	Kenalog	Unspecified contact dermatitis	Y
	Kenalog	Pemphigus	Y
	Methyprednisolone Acetate	Nephrotic Syndrome	Y
	Furosemide	Aneurysm of unspecified site	Y
	Dexamethasone	Pemphigus	Y
FR-Neural	Onglyza	Type 2 diabetes mellitus	Y
	Pradaxa	Essential (primary) hypertension	Y
	Oxytocin	Pauciarticular juvenile rheumatoid arthritis	Y
	Terbutaline sulfate	HIV 2 as the cause of diseases classified elsewhere	N
	Lipitor	Cerebral infarction	Y

Knowledge discovery

Model	Drug	Disease Target	Plausible
None	Naproxen	Ankylosing Spondylitis	Y
	Latanoprost	Superficial injury of ankle, foot and toes	N
	Pulmicort	Psoriasis, unspecified	Y
	Furosemide	Aneurysm of unspecified site	Y
	Desonide	Chlamydial lymphogranuloma (venereum)	N
FR-Identity	Latanoprost	Superficial injury of ankle, foot and toes	N
	Elixophyllin	Pneumonia in diseases classified elsewhere	Y
	Furosemide	Aneurysm of unspecified site	Y
	Oxistat	Mycosis fungoides	Y
	Trifluridine	Congenital Pneumonia	N
FR-Linear	Kenalog	Unspecified contact dermatitis	Y
	Kenalog	Pemphigus	Y
	Methyprednisolone Acetate	Nephrotic Syndrome	Y
	Furosemide	Aneurysm of unspecified site	Y
	Dexamethasone	Pemphigus	Y
FR-Neural	Onglyza	Type 2 diabetes mellitus	Y
	Pradaxa	Essential (primary) hypertension	Y
	Oxytocin	Pauciarticular juvenile rheumatoid arthritis	Y
	Terbutaline sulfate	HIV 2 as the cause of diseases classified elsewhere	N
	Lipitor	Cerebral infarction	Y

Recent clinical trial!

Knowledge discovery

Model	Drug	Disease Target	Plausible
None	Naproxen	Ankylosing Spondylitis	Y
	Latanoprost	Superficial injury of ankle, foot and toes	N
	Pulmicort	Psoriasis, unspecified	Y
	Furosemide	Aneurysm of unspecified site	Y
	Desonide	Chlamydial lymphogranuloma (venereum)	N
FR-Identity	Latanoprost	Superficial injury of ankle, foot and toes	N
	Elixophyllin	Pneumonia in diseases classified elsewhere	Y
	Furosemide	Aneurysm of unspecified site	Y
	Oxistat	Mycosis fungoides	Y
	Trifluridine	Congenital Pneumonia	N
FR-Linear	Kenalog	Unspecified contact dermatitis	Y
	Kenalog	Pemphigus	Y
	Methyprednisolone Acetate	Nephrotic Syndrome	Y
	Furosemide	Aneurysm of unspecified site	Y
	Dexamethasone	Pemphigus	Y
FR-Neural	Onglyza	Type 2 diabetes mellitus	Y
	Pradaxa	Essential (primary) hypertension	Y
	Oxytocin	Pauciarticular juvenile rheumatoid arthritis	Y
	Terbutaline sulfate	HIV 2 as the cause of diseases classified elsewhere	N
	Lipitor	Cerebral infarction	Y

Existing label!

Knowledge discovery

Model	Drug	Disease Target	Plausible
None	Naproxen	Ankylosing Spondylitis	Y
	Latanoprost	Superficial injury of ankle, foot and toes	N
	Pulmicort	Psoriasis, unspecified	Y
	Furosemide	Aneurysm of unspecified site	Y
	Desonide	Chlamydial lymphogranuloma (venereum)	N
FR-Identity	Latanoprost	Superficial injury of ankle, foot and toes	N
	Elixophyllin	Pneumonia in diseases classified elsewhere	Y
	Furosemide	Aneurysm of unspecified site	Y
	Oxistat	Mycosis fungoides	Y
	Trifluridine	Congenital Pneumonia	N
FR-Linear	Kenalog	Unspecified contact dermatitis	Y
	Kenalog	Pemphigus	Y
	Methyprednisolone Acetate	Nephrotic Syndrome	Y
	Furosemide	Aneurysm of unspecified site	Y
	Dexamethasone	Pemphigus	Y
FR-Neural	Onglyza	Type 2 diabetes mellitus	Y
	Pradaxa	Essential (primary) hypertension	Y
	Oxytocin	Pauciarticular juvenile rheumatoid arthritis	Y
	Terbutaline sulfate	HIV 2 as the cause of diseases classified elsewhere	N
	Lipitor	Cerebral infarction	Y

Recent relabeling!

On the effective use of pretraining

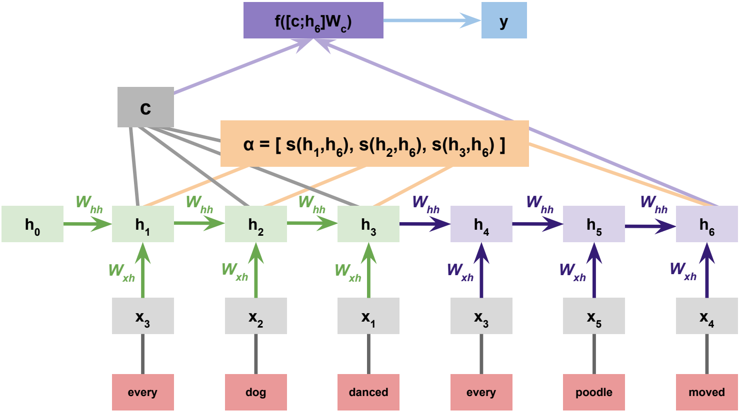
Cases et al., ‘On the effective use of pretraining for natural language inference‘ (2017)

Experimental setting: SNLI

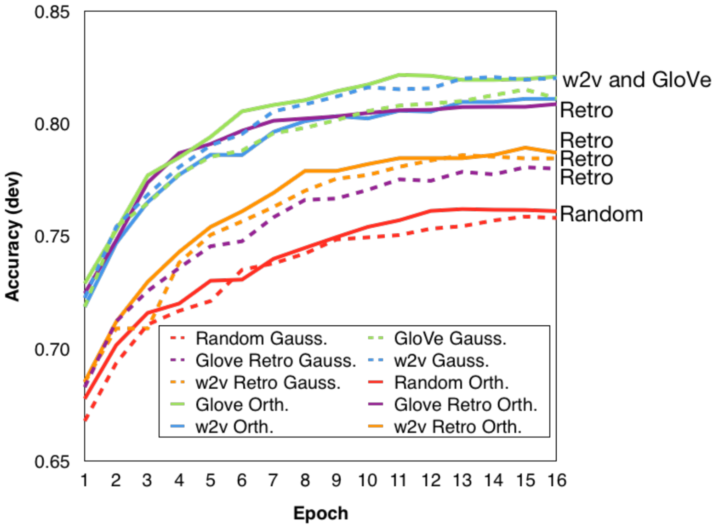
Premise	Labels	Conclusion
A man inspects the uniform of a figure in some East Asian country.	**contradiction** **c c c c c**	The man is sleeping
An older and younger man smiling.	**neutral** **n n e n n**	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	**entailment** **e e e e e**	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	**neutral** **n n e c n**	A happy woman in a fairy costume holds an umbrella.

From Bowman, *Modeling natural language semantics with learned representations* (2017)

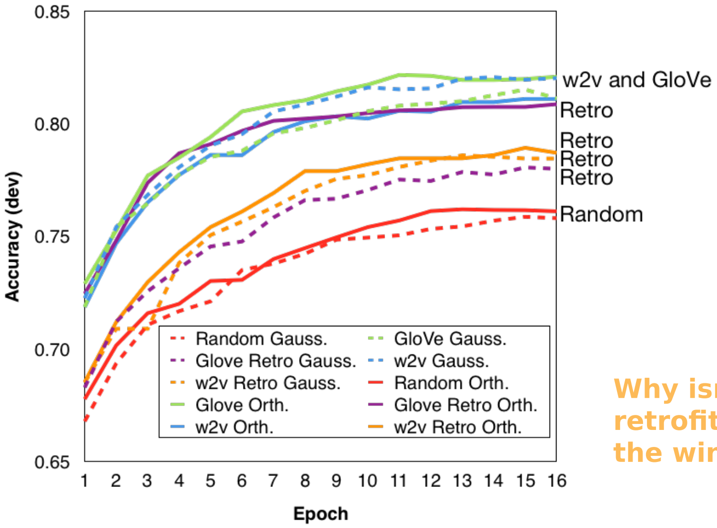
Bidirectional RNN with attention



Basic results

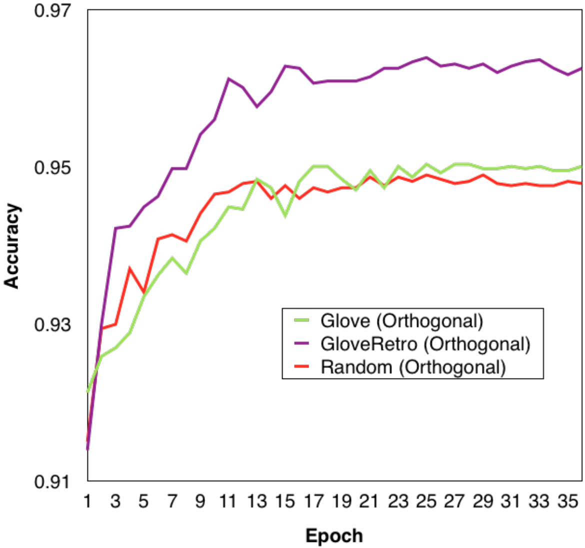


Basic results



Why isn't retrofitting the winner?

Lexical relations in WordNet



Trouble for compositional semantics?

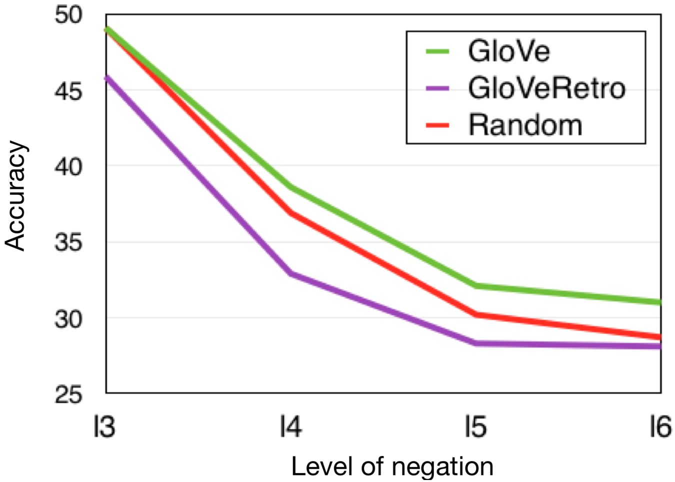
Negation

	not-p, not-q	p, not-q	not-p, q
p disjoint q	neutral	hyponym	hypernym
p equal q	equal	disjoint	disjoint
p neutral q	neutral	neutral	neutral
p hyponym q	hypernym	disjoint	neutral
p hypernym q	hyponym	neutral	disjoint

Examples

puppy	hyponym	mammal	⇒	not-puppy	hypernym	not-mammal
puppy	hyponym	mammal	⇒	puppy	disjoint	not-mammal
puppy	hyponym	mammal	⇒	not-puppy	neutral	mammal

Results for recursively applied negation



Conclusion

Conclusion

- Distributional representations are powerful and easy to obtain, but they tend to reflect only similarity (synonymy, connotation).

Conclusion

- Distributional representations are powerful and easy to obtain, but they tend to reflect only similarity (synonymy, connotation).
- Structured resources are sparse and hard to obtain, but they support learning rich, diverse semantic distinctions.

Conclusion

- Distributional representations are powerful and easy to obtain, but they tend to reflect only similarity (synonymy, connotation).
- Structured resources are sparse and hard to obtain, but they support learning rich, diverse semantic distinctions.
- Can we have the best aspects of both?

Conclusion

- Distributional representations are powerful and easy to obtain, but they tend to reflect only similarity (synonymy, connotation).
- Structured resources are sparse and hard to obtain, but they support learning rich, diverse semantic distinctions.
- Can we have the best aspects of both? **Yes!**

Conclusion

- Distributional representations are powerful and easy to obtain, but they tend to reflect only similarity (synonymy, connotation).
- Structured resources are sparse and hard to obtain, but they support learning rich, diverse semantic distinctions.
- Can we have the best aspects of both? **Yes!**
- And these methods can achieve the sort of grounding that linguists and psychologists endorse.

Conclusion

- Distributional representations are powerful and easy to obtain, but they tend to reflect only similarity (synonymy, connotation).
- Structured resources are sparse and hard to obtain, but they support learning rich, diverse semantic distinctions.
- Can we have the best aspects of both? **Yes!**
- And these methods can achieve the sort of grounding that linguists and psychologists endorse.
- But there remain open questions about how these enriched representations behave in complex systems.

Conclusion

- Distributional representations are powerful and easy to obtain, but they tend to reflect only similarity (synonymy, connotation).
- Structured resources are sparse and hard to obtain, but they support learning rich, diverse semantic distinctions.
- Can we have the best aspects of both? **Yes!**
- And these methods can achieve the sort of grounding that linguists and psychologists endorse.
- But there remain open questions about how these enriched representations behave in complex systems.

Thanks!