Vector Representations of Word Meaning in Context

Lea Frermann

Universität des Saarlandes

May 23, 2011

Outline

Outline

Motivation

- **Context and syntactic structure are essential for modelling semantic similarity.**

- Example 1
  (a) It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.
  (b) That day the office manager, who was drinking, hit the problem sales worker with a bottle, but it was not serious.
- Example 2
  (a) catch a ball
  (b) catch a disease
  (c) attend a ball

Logical vs. Distributional Representation of Semantics

Modelling word semantics in a distributional way:

+ Rich and easily available resources
+ High coverage and robust
+ Little hand-crafting necessary
- Vectors represent the semantics of one word in isolation
- Compositionality is hard to achieve

→ **Augment vector representations in a way that allows incorporation of context/syntactic information**

Outline

Vector Representation of Word-level Semantics

	animal	stable	village	gallop	jokey	
horse	0	6	2	10	4	$= \mathbf{u}$
run	1	8	4	4	0	$= \mathbf{v}$

- Vector Dimensions: Co-occurring words
- Values: Co-occurrence frequencies

Vector Composition

Define a set of possible models:

$$\mathbf{p} = f(\mathbf{u},\mathbf{v},R,K)$$

- **p** = resulting vector
- f = function which combines the two vectors (addition, multiplication, combination of both)
- **u,v** = vectors representing individual words
- R = syntactic relation between words represented by **u,v**
- K = additional knowledge

Vector Representation of Word-level Semantics

- Fix the relation R
- Ignore additional knowledge K
- Independence Assumption: Only the i^{th} component of u/v influences the i^{th} component of p.

$$p_i = u_i + v_i$$
$$p_i = u_i * v_i$$

	animal	stable	village	gallop	jokey	
horse	0	6	2	10	4	$= \mathbf{u}$
run	1	8	4	4	0	$= \mathbf{v}$

Additive Model: $\mathbf{p} = [1\ 14\ 6\ 14\ 4]$
Multiplicative Model: $\mathbf{p} = [0\ 48\ 8\ 40\ 0]$

Vector Representation of Word-level Semantics

- Loosen symmetry assumption in
- Introduce weights

$\rightarrow$ Semantically important words can have higher influence

$$p_i = \alpha n_i + \beta v_i$$

- Optimized weights: $\alpha = 20$ and $\beta = 80$
  - $\mathbf{n}$ = noun vector and $\mathbf{v}$ = verb vector

Vector Representation of Word-level Semantics

- Corresponds to the model introduced in Kintsch(2001)
- Re-introduce additional knowledge K
- $(\mathbf{d})$ = vectors of n distributional neighbors of the predicate
- Makes the additional model sensitive to syntactic structure

$$\mathbf{p} = \mathbf{u} + \mathbf{v} + \sum \mathbf{d}$$

- Kintsch's optimal parameters:
  - m most similar neighbors to the predicate = 20
  - from m, select k most similar neighbors to its argument = 1

Vector Representation of Word-level Semantics

- Combine additional and multiplicative models
- Avoids the multiplication-by-zero problem

$$p_i = \alpha n_i + \beta v_i + \gamma n_i v_i$$

- Optimized weights: $\alpha = 0$ and $\beta = 95$ and $\gamma = 5$
  - $\mathbf{n}$ = noun vector and $\mathbf{v}$ = verb vector

Evaluation

- How is the verb's meaning influenced in the context of its subject?
- Measure similarity of reference verb relative to **landmarks**
  - **Landmark** = Synonym of the reference verb in context of the given subject
  - Chosen to be as dissimilar as possible according to WordNet similarity

Noun	Reference	High	Low
The fire	glowed	burned	beamed
The face	glowed	beamed	burned
The child	strayed	roamed	digressed
The discussion	strayed	digressed	roamed
The sales	slumped	declined	slouched
The shoulders	slumped	slouched	declined

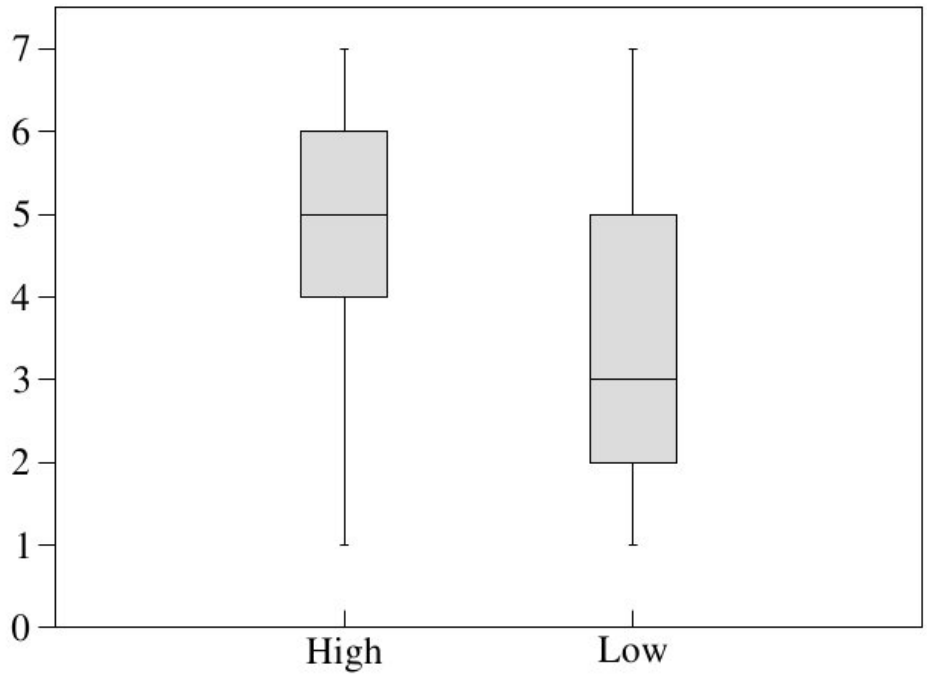
Figure: Example Stimuli with High and Low similarity landmarks.

Evaluation –Pretests

- Compile a list of intransitive verbs from CELEX
- Extract all verb-subject pairs that occur $>$ 50 times in the British National Corpus
- Pair these verbs with two landmarks
- Pick the subset of verbs with least variation in human similarity ratings
- Result: 15 verbs x 4 nouns x 2 landmarks = 120 sentences

Evaluation –Experiments

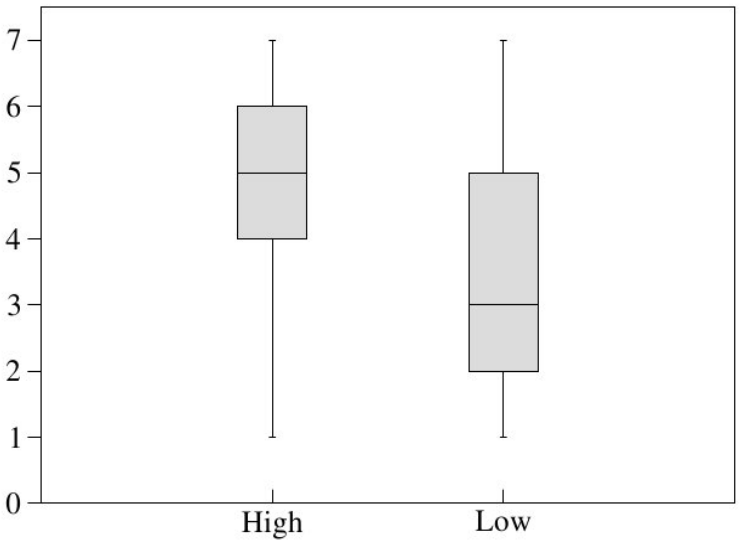
- Humans are shown reference sentence and landmark
- Rate similarity on a scale from 1-7
- Significant correlation
- Inter-human agreement $\rho = 0.4$



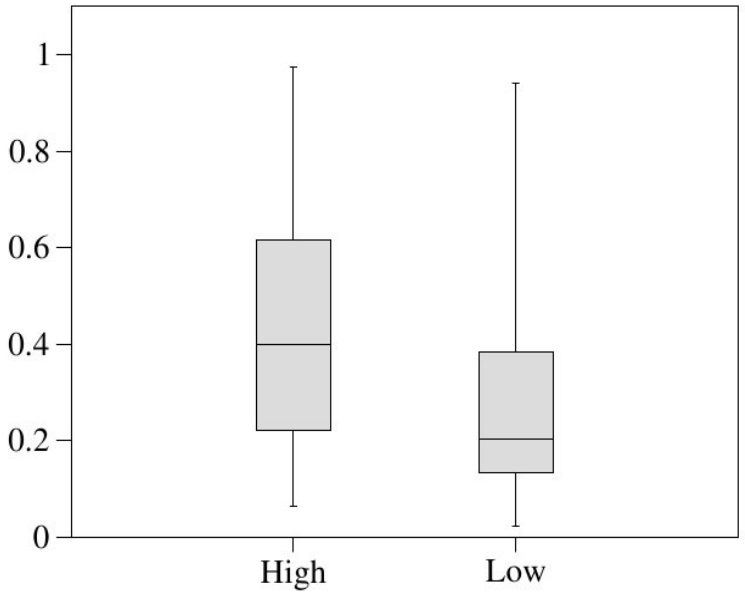
Evaluation –Model Parameters

- 5 context words on either side of the reference verb
- 2000 most frequent context words as vector components
- Vector values: $\frac{p(ContextWord|TargetWord)}{p(ContextWord)}$
- Cosine similarity for vector comparison

Evaluation –Results I



(a) Human ratings for High and Low similarity items



(b) Multiplication Model ratings for High and Low similarity items

Evaluation –Results II

Model	High	Low	ρ
Noncomp	0.27	0.26	0.08**
Add	0.59	0.59	0.04*
WeightAdd	0.35	0.34	0.09**
Kintsch	0.47	0.45	0.09**
Multiply	**0.42**	**0.28**	0.17**
Combined	0.38	0.28	0.19**
UpperBound	4.94	3.25	0.40**

Figure: Model means for High and Low similarity items and correlation coefficients with human judgments (*: $p<0.05$, **: $p< 0.01$)

Conclusion

- Component-wise vector multiplication outperforms vector addition
- Basic representation of word meaning as syntax-free bag-of-words-based vectors
- Their actual instantiations of models are insensitive to syntactic relations and word order
- Future Work:
  - Include more linguistic information
  - Evaluation on larger and more realistic data sets

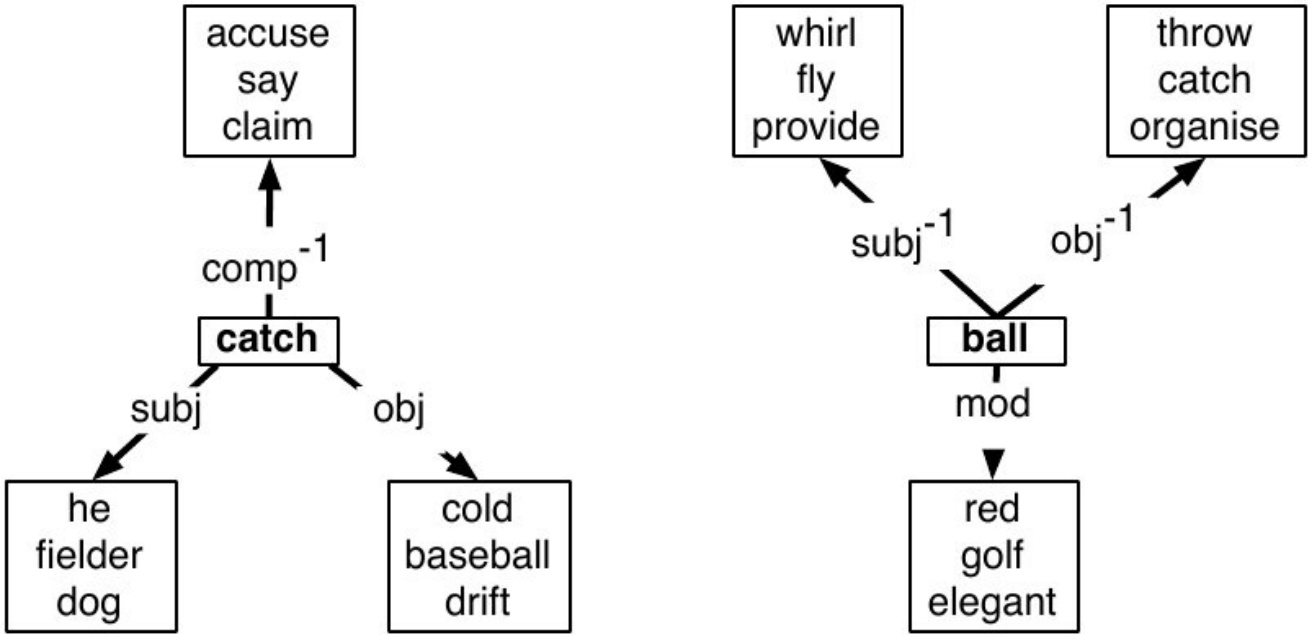
Outline

General Idea

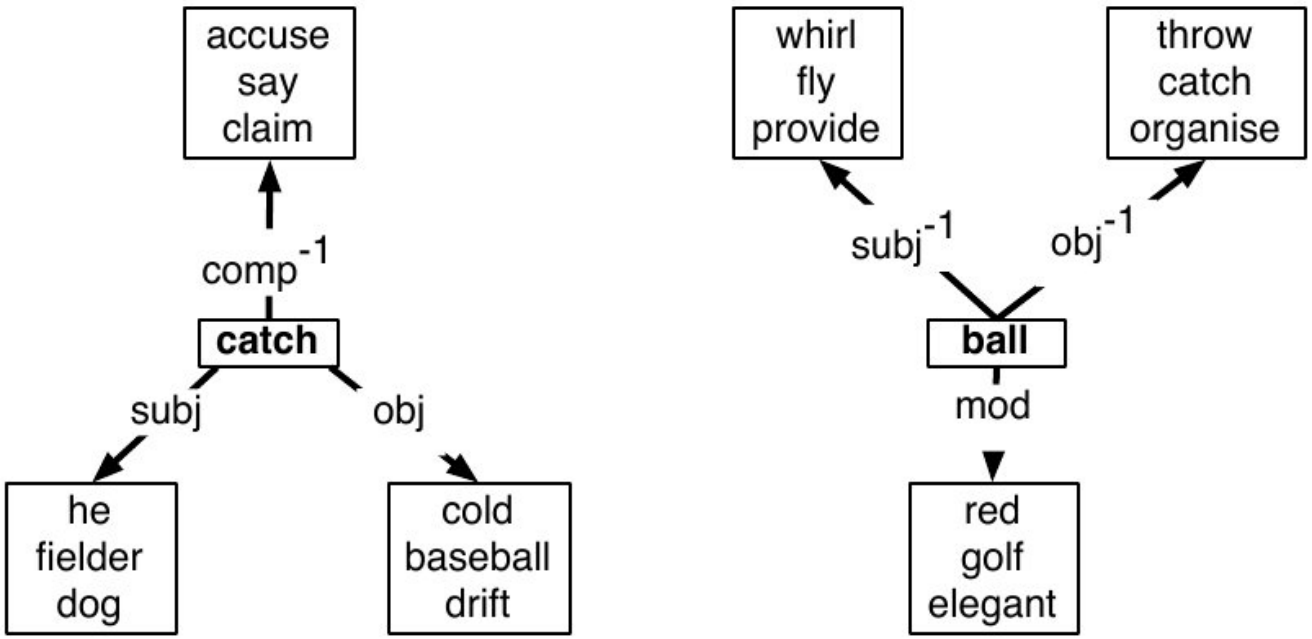
- Problem 1: Lack of syntactic information
- Problem 2: Scaling up
  - A vector with fixed dimensionality can encode a fixed amount of information
  - There is no limit on sentence length
- Construct a structured vector space, containing a word's meaning as well as its selectional preferences
- Meaning of word a in context of word b = combination of a with b's selectional preferences

- Re-introduce additional knowledge K into the models!

Representing Lemma Meaning



Representing Lemma Meaning



Represent each word w as a combination of vectors in vector space D:

a) One vector modeling the lexical meaning (v)
b) A set of vectors modeling w's selectional preferences
   - $R : R \rightarrow D$
   - $R^{-1} : R \rightarrow D$

$$w = (v, R, R^{-1})$$

Selectional Preferences

- Selectional Preference of word b and relation r = centroid of seen filler vectors $\overrightarrow{v_a}$

$$R_b(r)_{\textbf{SELPREF}} = \sum_{a:f(a,r,b)>0} f(a,r,b) * \overrightarrow{v_a}$$

- *f(a,r,b)* = frequency of a occurring in relation r to b in the British National Corpus

Two Variations

- Alleviate noise caused by infrequent filler vectors

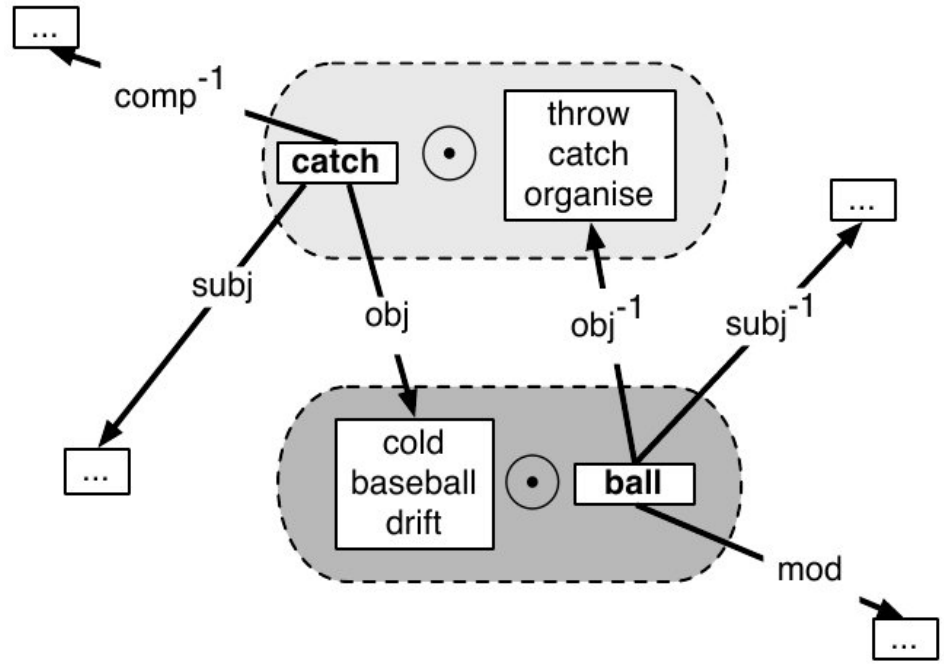
$$R_b(r)_{\text{SELPREF-CUT}} = \sum_{a:f(a,r,b)>\theta} f(a,r,b) * \overrightarrow{v_a}$$

- Alleviate noise caused by low-valued vector dimensions

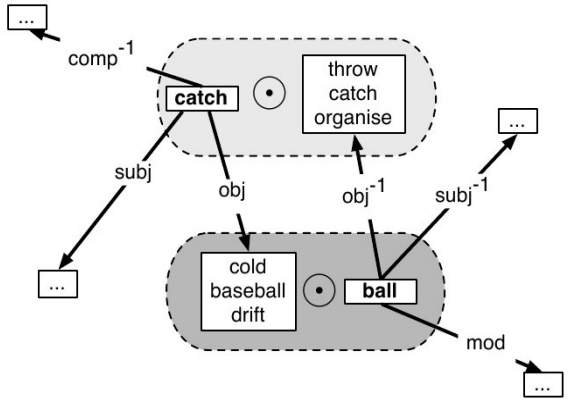
$$R_b(r)_{\text{SELPREF-POW}} =< v_1^n, \ldots, v_m^n >$$

Computing Meaning in Context

Verb meaning combined with the centroid of the vectors of the verbs to which the noun can stand in an object relation



Computing Meaning in Context



$$
\begin{aligned}
a' &= (v_a \odot R_b^{-1}(r) \quad , R_a - r \quad , R_a^{-1}) \\
b' &= (v_b \odot R_a(r) \quad , R_b \quad , R_b^{-1} - r)
\end{aligned}
$$

$(a', b') =$ vector representing meaning of word $a = (v_a, R_a, R_a^{-1})$ in the context of word $b = (v_b, R_b, R_b^{-1})$

$r \in R =$ relation which links a to b

Vector Spaces

1. Bag-of-Words space (BOW)
   - Co-occurrence frequencies of target and context within a context window of 10 (Mitchell and Lapata)

2. Dependency-based space (SYN)
   - Target and context word must be linked by a valid dependency path

Evaluation I –Results: Part 1

Model	high	low	ρ
	BOW space		
Target only	0.32	0.32	0.0
Selpref only	0.46	0.40	0.06**
M&L	0.25	0.15	0.20**
SELPREF	0.32	0.26	0.12**
SELPREF-CUT, $\theta = 10$	0.31	0.24	0.11**
SELPREF-POW, $n = 20$	0.11	0.03	**0.27****
Upper bound	—	—	0.4
	SYN space		
Target only	0.20	0.20	0.08**
Selpref only	0.27	0.21	0.16**
M&L	0.13	0.06	**0.24****
SELPREF	0.22	0.16	0.13**
SELPREF-CUT, $\theta = 10$	0.20	0.13	0.13**
SELPREF-POW, n=30	0.08	0.04	0.22**
Upper bound	—	—	0.4

Figure: Mean cosine similarity for High and Low similarity items and correlation coefficients with human judgments (**: p$<$ 0.01)

Evaluation I –Results: Part 2

Model	lex.vector	$subj^{-1}$ vs. obj^{-1}
SELPREF	0.23 (0.09)	0.88 (0.07)
SELPREF-CUT (10)	0.20 (0.10)	0.72 (0.18)
SELPREF-POW (30)	0.03 (0.08)	0.52 (0.48)

Figure: Average similarity (and standard deviation); cosine similarity in SYN space

- Column 1:
  - To what extent does the difference in method (combination with words' lexical vectors vs. selpref vectors) translate to a difference in predictions?
- Column 2:
  - Does syntax-aware vector combination make a difference?

Evaluation II –Settings

Paraphrase ranking for a broader range of constructions

- Data: SemEval 1 *lexical substitution* data set
  - 10 instances of each of 200 target words in sentential contexts
  - Contextually appropriate paraphrases for each instance; rated by humans

- Subset of constructions used for evaluation:
  (a) target intransitive verbs with noun subjects
  (b) target transitive verbs with noun objects
  (c) target nouns occurring as objects of verbs

Evaluation II –Settings

Rank paraphrases on the basis of their cosine-similarity to:

- SELPREF-POW (30)
  - V-SUBJ: verb & noun's subj^{-}1 preferences
  - V-OBJ: verb & noun's obj^{-}1 preferences
  - N-OBJ: noun & verb's obj preferences
- Mitchell and Lapata
  - Direct noun-verb combination

Evaluation II –Settings

"Out of ten" evaluation metric:

$$P_{10} = 1/|I| \sum_i \frac{\sum_{s \in M_i \cap G_i} f(s,i)}{\sum_{s \in G_i} f(s,i)}$$

- G_i = Gold Parse for item i
- M_i = model's top ten paraphrases for i
- f(s,i) = frequency of s as paraphrase for i

Evaluation II –Results

Model	V-SUBJ	V-OBJ	N-OBJ
Target only	47.9	47.4	49.6
Selpref only	54.8	51.4	55.0
M&L	50.3	52.2	53.4
SELPREF-POW, n=30	**63.1**	**55.8**	**56.9**

$\rightarrow$ Knowledge about a single context word (although not necessarily informative) can already lead to significant improvement

Conclusion: Word Meaning in Context

- A model of word meaning and selectional preferences in a structured vector space
- Outperforms the bag-of-words model of Mitchell and Lapata
- Evaluation on a broader range of relations and realistic paraphrase candidates
- Future work:
  - Integrating information from multiple relations (eg. both Subject and Object)
  - Application of models to more complex NLP problems

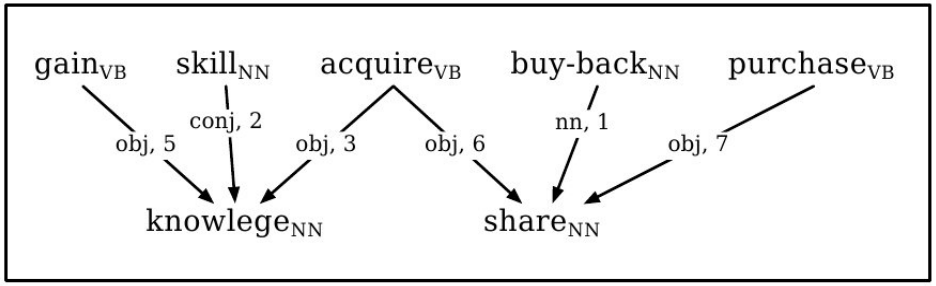
Outline

Basic idea

Assumes richer internal structure of vector representations

- Model relation-specific co-occurrence frequencies
- Use syntactic second-order vector representations
  - Reduces data sparseness caused by use of syntax
  - Makes vector transformations possible, which avoids complementary information in vectors for different parts of speech

1st-Order Context Vectors

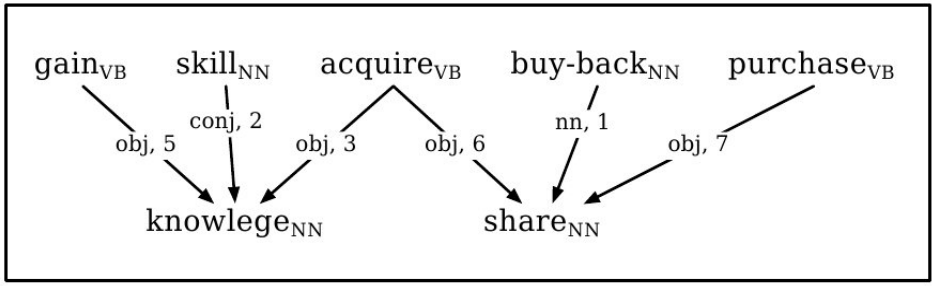


$$[w] = \sum_{r \in R, w' \in W} \omega(w, r, w') * \overrightarrow{e}_{r,w'}$$

In vector space V_1 $\{\overrightarrow{e}_{r,w'} | r \in R, w' \in W\}$

$$[knowledge] = < 5_{(OBJ^{-1}, gain)}, 2_{(CONJ^{-1}, skill)}, 3_{(OBJ^{-1}, acquire)}, \cdots >$$

2nd-Order Context Vectors I

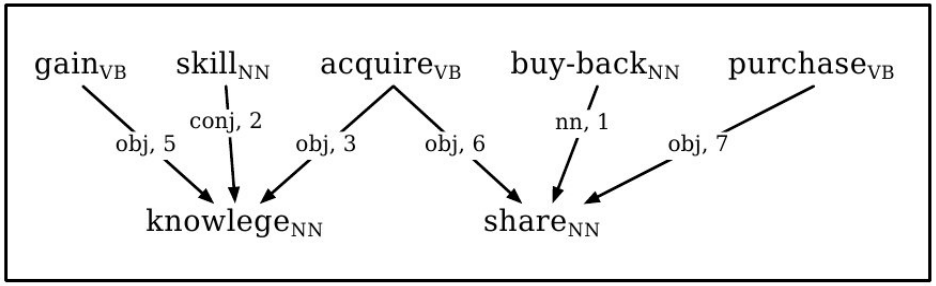


- All words that can be reached in the co-occurrence graph with 2 steps
- Dimensions = (r,w',r',w''), generalized to (r,r',w'')
- Vectors contain paths of the form (r, r^{-1}, w'')

$\rightarrow$ relate a word to other words that are possible substitution candidates

If r = OBJ and r' = OBJ^{-1} then the coefficients of $\overrightarrow{e}_{r,r',w''}$ in $[[w]]$ characterize the distribution of verbs w'' sharing objects with w.

2nd-Order Context Vectors II



$$[[w]] = \sum_{r \in R, w'' \in W} \left(\sum_{w' \in W} \omega(w, r, w') * \omega(w', r', w'') \right) \overrightarrow{e}_{r,r',w''}$$

In Vector space V_2 $\{\overrightarrow{e}_{r,r',w'} | r, r' \in R, w' \in W\}$

$[[Acquire]] =$
$< 15_{(OBJ, OBJ^{-1}, gain)}, 6_{(OBJ, CONJ^{-1}, skill)}, 42_{(OBJ, OBJ^{-1}, purchase)}, \cdots >$

Combining Context Vectors

$$[[w_{r:w'}]] = [[w]] x L_r([w'])$$

$[[acquire]]$
$< 15_{(OBJ, OBJ^{-1}, gain)}, 6_{(OBJ, CONJ^{-1}, skill)}, 42_{(OBJ, OBJ^{-1}, purchase)}, \cdots >$

$L_r([knowledge])$
$< 5_{(OBJ^{-1}, gain)}, 2_{(CONJ^{-1}, skill)}, 3_{(CONJ^{-1}, skill)}, \cdots >$

$[[acquire_{OBJ:knowledge}]]$
$< 75_{(OBJ, OBJ^{-1}, gain)}, 12_{(OBJ, CONJ^{-1}, skill)}, 0_{(OBJ, OBJ^{-1}, purchase)}, \cdots >$

Contextualization of multiple vectors

To contextualize multiple words, take the sum of pairwise contextualizations

$$[[w_{r_1:w_1,\ldots,r_n:w_n}]] = \sum_{k=1}^{n} [[w_{r_k:w_k}]]$$

Vector Space

- Obtain dependency trees from the parsed English Gigaword corpus (Stanford parser)
- Obtain 3.9 mio dependency triples
- Compute the vector space from a subset, exceeding a threshold in pmi and frequency of occurrence

Evaluation I –Procedure

Sentence	Paraphrases
Teacher education students will **acquire** the knowledge and skills required to [...]	gain 4; amass 1; receive 1

Compare contextually constrained 2^{nd} order vector of the target verb to unconstrained 2^{nd} order vectors of the paraphrase candidates:

$[[acquire_{SUBJ:student,OBJ:knowledge}]]$ vs. $[[gain]], [[amass]], [[receive]], \ldots$

Evaluation I –Metrics

1. "Out of ten" (P_{10})
2. Generalized Average Precision

$$GAP = \frac{\sum_{i=1}^{n} I(x_i)p_i}{\sum_{i=1}^{R} I(y_i)\overline{y}_i}$$

   - x_i = the weight of i^{th} item in the gold standard, or 0 if it does not appear
   - $I(x_i) = 1$ if $x_i > 0$, 0 otherwise
   - $\overline{y}_i$ = average weight of the ranked gold standard list $y_1, ..., y_i$
   - $p_i = \frac{\sum_{k=1}^{i} x_k}{i}$

$\rightarrow$ Rewards the correct order of a ranked list

Evaluation I –Results

Model	GAP	P_{10}
Random baseline	26.03	54.25
E&P (add, object)	29.93	66.20
E&P (min, subject & object)	32.22	64.86
1^{st} order contextualized	36.09	59.35
2^{nd} order uncontextualized	37.65	66.32
Full model	**45.94**	**73.11**

Evaluation II

Rank WordNet senses of a word w in context

- Word sense =
  centroid of the second-order vectors of the synset members +
  centroid of the sense's hypernyms scaled down by factor 10

- Compare contextually constrained 2^{nd} order vector of the target verb to unconstrained 2^{nd} order vectors of the paraphrase

Evaluation II –Results

Word	Present paper	WN-Freq	Combined
ask	0.344	0.369	0.431
add	0.256	0.164	0.270
win	0.236	0.343	0.381
average	0.279	0.291	**0.361**

Figure: Correlation of model predictions and human ratings (Spearman's ρ)

; Upper Bound: 0.544

Conclusion

- A model for adapting vector representations of words according to their context
- Detailed syntactic information through combinations of 1*st* and 2*nd* order vectors
- Outperforms state of the art systems and improves weakly supervised word sense assignment
- Future work:
  - Generalization to larger syntactic contexts by recursive integration of information

Outline

Conclusion

- Syntactic and contextual information is essential for vector representations of word meaning
- Multiplicative vector combination results in the most accurate models
- Context as vector representations of a word's selectional preferences for each relation
- Context as interfering 1*st* and 2*nd* order context vectors of words
- Evaluation on word sense similarity, paraphrase ranking and word sense ranking
- Future work:
  - Scale up models to allow for more contextual information
  - Scale up models to adapt them to more complex NLP applications

Thank you for your attention!

Bibliography

Katrin Erk and Sebastian Padó.
A structured vector space model for word meaning in context.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 897–906, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata.
Vector-based models of semantic composition.
In *In Proceedings of ACL-08: HLT*, pages 236–244, 2008.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal.
Contextualizing semantic representations using syntactically enriched vector models.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 948–957, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.