A Structured Vector Space Model for Word Meaning in Context

Katrin Erk & Sebastian Padó (2008)

0. Content

1. **Motivation**
    1. Drawbacks of Vector Composition
    2. Related Work: Mitchell and Lapata (2008)
2. **Structured Vector Space Model Defined**
    1. Lemma Meaning
    2. Meaning in Context
3. **Experimental Evaluation**
    1. Experiment 1
    2. Experiment 2
4. **Conclusion**

1. Motivation (Vector Composition)

Default **semantic space** models:

- Lemma meanings are encoded as high-dimensional vectors
- Each vector = **average over all possible usages** of a lemma

Problem: Word meaning can vary significantly depending on the **context** in which the word occurs

Solution: **Vector composition**

Ex: The meaning of lemma *a* in the context *b* is the vector *c,* which is a **function** of vectors *a* and *b:* $c = a \odot b.$

1. Motivation (Vector Composition)

Claim: Models of word meaning relying on vector composition **fail to sufficiently consider syntax** and are, as such, **limited in scope and scalability**

(1) a horse draws (horse is the subject of draws)

(2) draw a horse (horse is the object of draws)

→ (1) and (2) are assigned the **same representation** following vector composition!

1. Motivation (Vector Composition)

Claim: Models of word meaning relying on vector composition **fail to sufficiently consider syntax** and are, as such, **limited in scope and scalability**

Assumption: Vector *c* represents the meaning of the phrase *a*+*b*. **However**:

The **dimensionality** of *c* is **fixed** → can only encode a **limited amount** of structural information

Sentence length is **not fixed** → **no upper limit** on structural information that needs to be encoded

1. Motivation (Mitchell and Lapata (2008))

Evaluation of combinatorial models based on their representation of the combined meaning of **predicate** *p* and its **argument** *a*:

$$c = f(p, a, R, K)$$

c: Meaning of *p* + *a*

R: Syntactic relation between *p* and *a*

K: Additional, background knowledge

Conclusion: **Multiplicative models outperform additive ones**

1. Motivation (Mitchell and Lapata (2008))

Evaluation of combinatorial models based on their representation of the combined meaning of **predicate** *p* and its **argument** *a*:

$$c = f(p, a, R, K)$$

<u>However</u>: Concrete instantiations **ignore R and K**

→ Same limitations as default vector composition models w.r.t. scope and scalability!

2. SVS Model Defined

Intuition: The interpretation of a word in a context is guided by **expectations** about **typical events**:

(3) *catch a ball*

→ *catch*: an action that can be performed with a ball

→ *ball*: a thing that can be caught

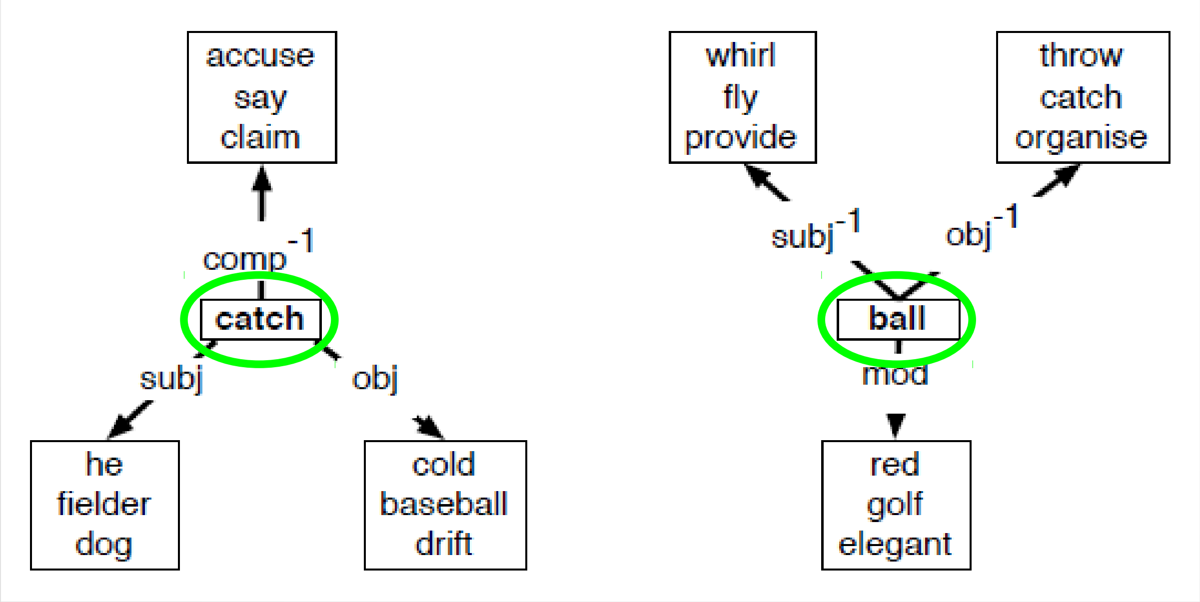
Cognitive evidence: Reading times, sentence processing etc.

Linguistic evidence: Selectional restrictions and preferences (ineducable from corpora)

2. SVS Model Defined (Lemma Meaning)

Proposal: Encode the meaning of each lemma as a **combination** of:

1) One vector modeling its **lexical meaning**
2) Set of vectors, each encoding the semantic expectations of exactly **one relation** the word supports



1) lexical vector

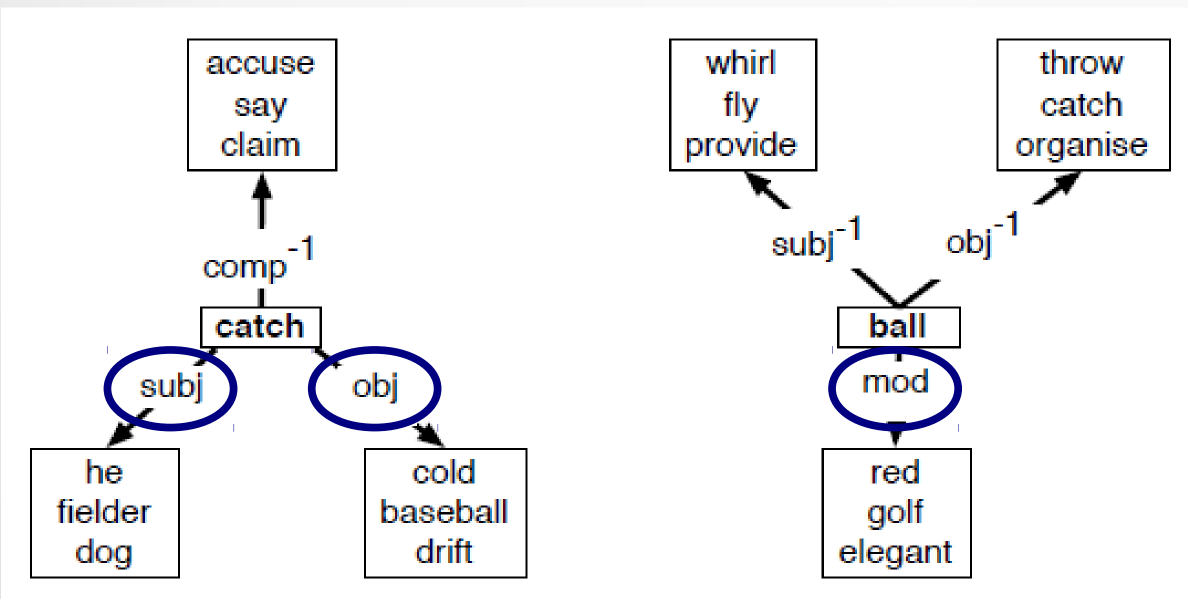
2a) **selectional preferences**

2b) **inverse selectional preferences**

2. SVS Model Defined (Lemma Meaning)

<u>Proposal</u>: Encode the meaning of each lemma as a **combination** of:

1) One vector modeling its **lexical meaning**
2) Set of vectors, each encoding the semantic expectations of exactly **one relation** the word supports



1) **lexical vector**

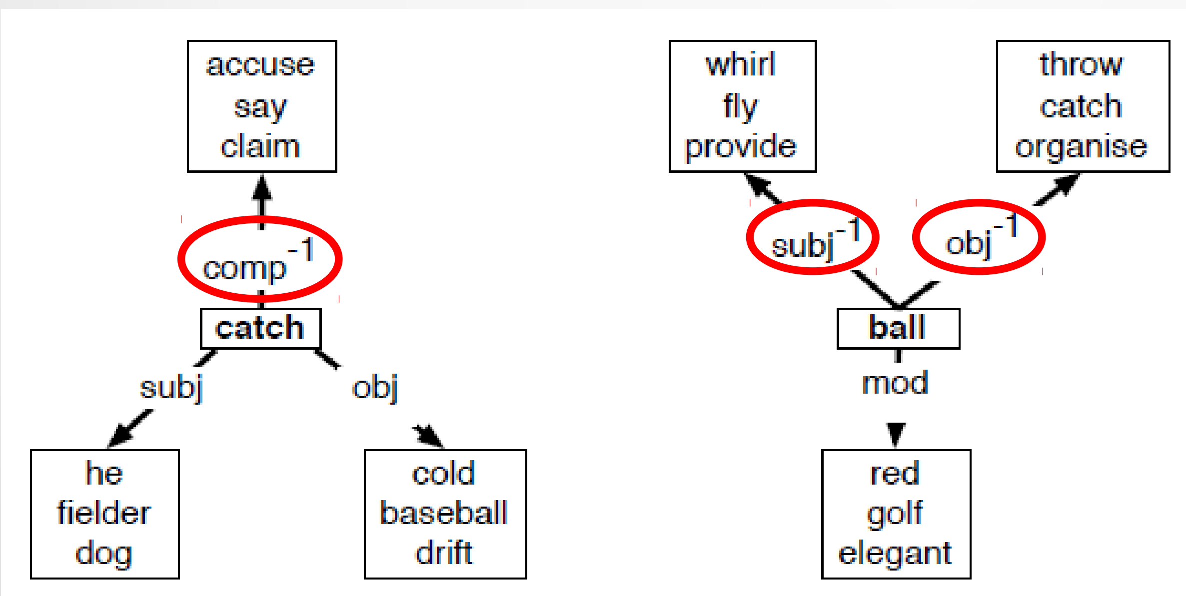
2a) **selectional preferences**

2b) **inverse selectional preferences**

2. SVS Model Defined (Lemma Meaning)

Proposal: Encode the meaning of each lemma as a **combination** of:

1) One vector modeling its **lexical meaning**
2) Set of vectors, each encoding the semantic expectations of exactly **one relation** the word supports



1) **lexical vector**

2a) **selectional preferences**

2b) **inverse selectional preferences**

2. SVS Model Defined (Lemma Meaning)

Proposal: Encode the meaning of each lemma as a **combination** of:

1) One vector modeling its **lexical meaning**
2) Set of vectors, each encoding the semantic expectations of exactly **one relation** the word supports

Formal Definition:

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

w: meaning of the lemma

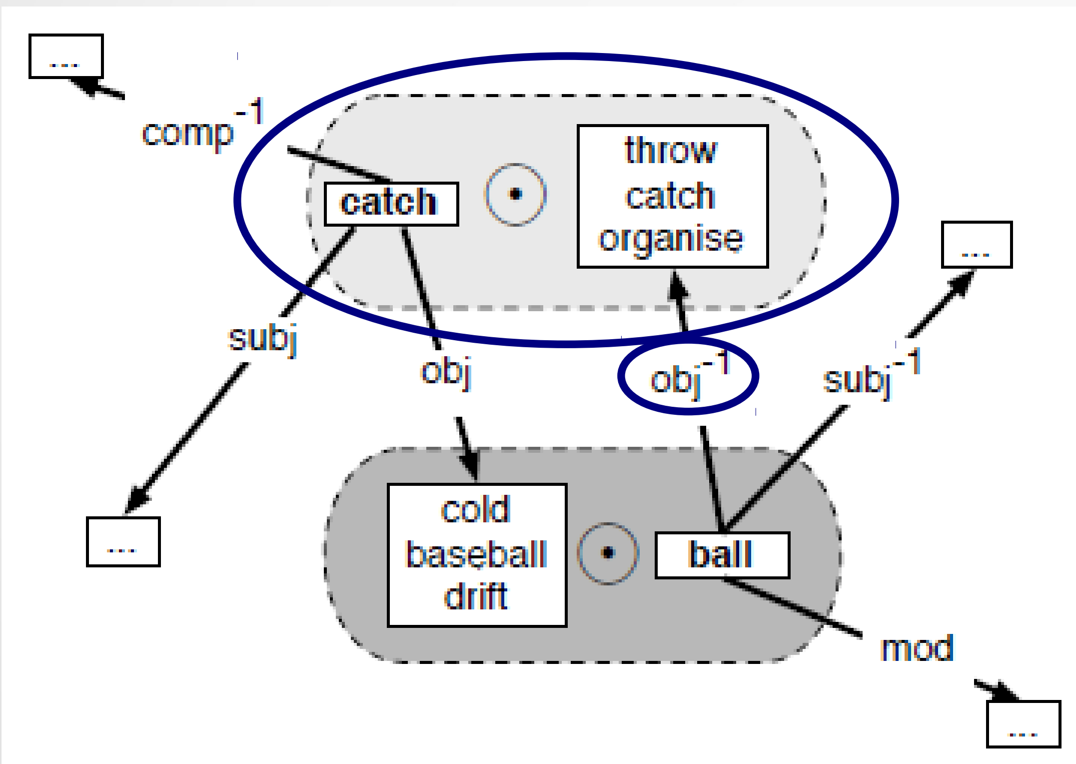
v: lexical vector describing word w

R: partial function mapping each relation label to a sel. pref. vector of w

R^{-1}: partial function mapping each role label to an inv. sel. pref. vector of w

2. SVS Model Defined (Meaning in Context)

Proposal: Compute the meaning of the word *a* in the context of the word *b* using their respective **selectional preferences**

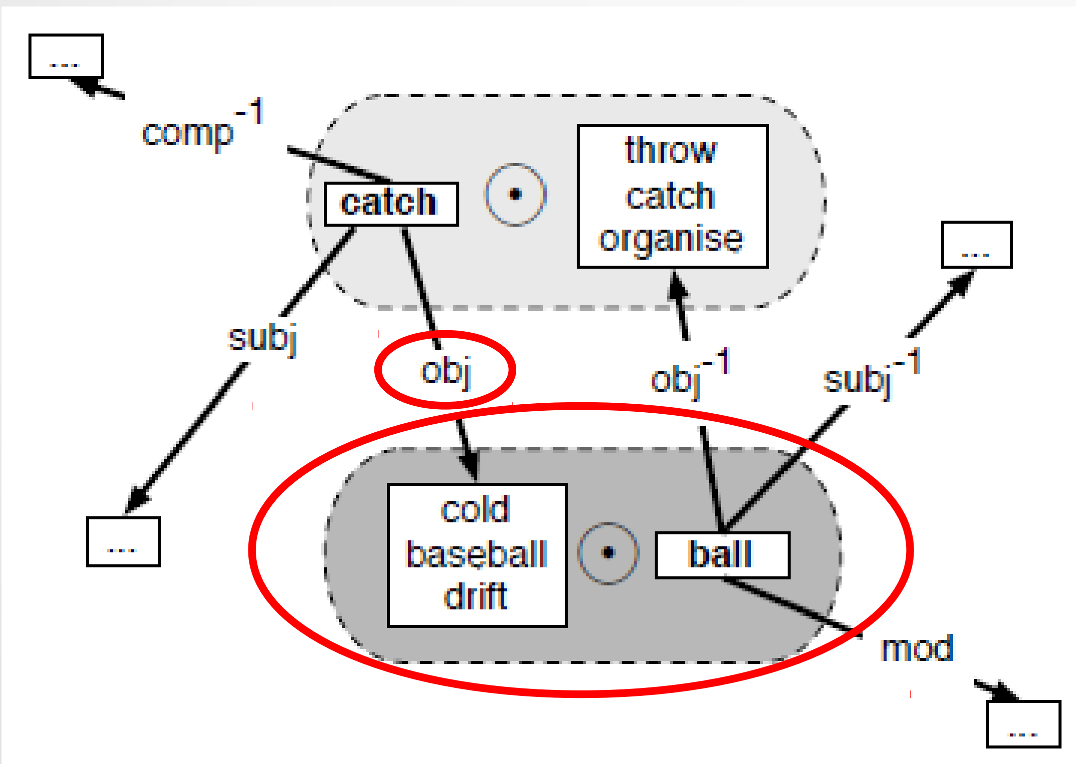


1) **Lexical vector** for *catch* is combined with **inverse object preference** of *ball*

2) **Lexical vector** of *ball* is combined with with **object preference** of *catch*

2. SVS Model Defined (Meaning in Context)

Proposal: Compute the meaning of the word *a* in the context of the word *b* using their respective **selectional preferences**



1) **Lexical vector** for *catch* is combined with **inverse object preference** of *ball*

2) **Lexical vector** of *ball* is combined with with **object preference** of *catch*

2. SVS Model Defined (Meaning in Context)

Proposal: Compute the meaning of the word *a* in the context of the word *b* using their respective **selectional preferences**

Formal Definition:

a = (v_a, R_a, R_a^{-1})

b = (v_b, R_b, R_b^{-1})

meaning of *a* + *b* = (*a'*, *b'*):

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

r = relation **linking *a* to *b***

2. SVS Model Defined (Meaning in Context)

Proposal: Compute the meaning of the word *a* in the context of the word *b* using their respective **selectional preferences**

Formal Definition:

meaning of *a* + *b* = (*a'*, *b'*):

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

$v_1 \odot v_2$: Direct vector combination function (e.g. addition)

- Combination **fails** if there is no relation *r* linking both words
- After *r* is filled, it is **deleted** from R_a and R_b^{-1}

2. SVS Model Defined (Meaning in Context)

Interim Conclusions:

- Combining **same words** through **different relations** generally results in **different adapted representations** → addresses issue of **scope**
- Proposed method produces **one context-adapted meaning per word** → addresses the issue of **scalability**
- Computing method is expressible within the framework proposed by Mitchell and Lapata (2008):

  K = semantic expectations of a and b

  $c = (a', b')$

3. Experimental Evaluation

Experiments 1 & 2: **Paraphrase tasks** determining how appropriate a predicate or argument paraphrase is in a context (i.e. judgments of the word *a* in the context of word *b*)

Two vector spaces:

Bag-of-words: Co-occurrence frequencies in a 10-word window

Dependency based: Co-occurring words have to be linked by a valid dependency path

Two baselines:

1. **Original vector of *a***
2. **Selectional preference of *b***

3. Experimental Evaluation

Three selectional preference models:

Selectional preference for word *b* and relation *r* is the w**eighted centroid of seen filler vectors** v_a:

1) **SELPREF**:

$$R_b(r)_{\text{SELPREF}} = \sum_{a:f(a,r,b)>0} f(a,r,b) \cdot \vec{v}_a$$

f(a,r,b) = **frequency** of a occurring in relation r to b in the training corpus (functions as a weight)

3. Experimental Evaluation

Three selectional preference models:

Selectional preference for word *b* and relation *r* is the w**eighted centroid of seen filler vectors** v_a:

2) **SELPREF-CUT**:

$$R_b(r)_{\text{SELPREF-CUT}} = \sum_{a: f(a,r,b) > \theta} f(a, r, b) \cdot \vec{v}_a$$

Aims to **reduce noise** introduced by **infrequent fillers**

Θ: Frequency **threshold**

3. Experimental Evaluation

<u>Three selectional preference models</u>:

Selectional preference for word *b* and relation *r* is the w**eighted centroid of seen filler vectors** v_a:

3) **SELPREF-POW**: $R_b(r)_{\text{SELPREF-POW}} = \langle v_1^n, \ldots, v_m^n \rangle$

Aims to **reduce noise** introduced by **low-valued dimensions**

High-count dimensions are inflated, low-count ones depressed

3. Experimental Evaluation (Experiment 1)

Replication of the Mitchell and Lapata (2008) experiment

Task: Paraphrase evaluation by comparison between a **predicate vector** and two predetermined, **highly dissimilar landmarks**

Method:

- **Cosine** as similarity measure
- **Nouns subj^{-1} preference** = second baseline (selfpref)
- For each of the thee models: **Verb's lexical vector** is combined with the **nouns subj^{-1} preference**
- **Evaluation scores**: Average landmark similarity, Spearman's ρ

3. Experimental Evaluation (Experiment 1)

Results:

- For the **bag-of-words** vector space, **SELPREF-POW outperformed its competitors**, including M&L's model
- For the **dependency-based** vector space, M&L's model performed better than SELPREF-POW numerically, with the difference **not being statistically significant**
- SVS makes **different predictions** then direct combination models and **produces different expectations for different relations**

3. Experimental Evaluation (Experiment 1)

Model	high	low	ρ
BOW space			
Target only	0.32	0.32	0.0
Selpref only	0.46	0.4	0.06**
M&L	0.25	0.15	0.20**
SELPREF	0.32	0.26	0.12**
SELPREF-CUT, θ=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	0.2	0.2	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, θ=10	0.2	0.13	0.13**
SELPREF-POW, n=30	0.08	0.04	0.22**
Upper bound	–	–	0.4

3. Experimental Evaluation (Experiment 2)

Identify paraphrase fit for a **broader range of constructions**

Task: Ranking problem – appropriate paraphrases should be ranked higher than ones not provided by human annotators

Method:

- Three types of sentences: V-SUBJ, V-OBJ, N-OBJ
- Target words annotated with **contextually fitting paraphrases**
- Paraphrase similarity measured as follows:

  V-SUBJ: verb + nouns' $subj^{-1}$ **V-OBJ**: verb + noun's obj^{-1}

  N-OBJ: noun + verb's obj

3. Experimental Evaluation (Experiment 2)

Results:

- Best results obtained by **SELPREF-POW**, which performs significantly better than M&L's model

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	52.0	53.4
SELPREF-POW, n=30	**63.1**	**55.8**	**56.9**

Measure used: Mean 'out of ten' precision

4. Conclusion

- SVS (SELPREF-POW) outperformed M&L's model in both experiments

- SVS achieved a better performance in a task utilizing more 'realistic' paraphrases (Experiment 2)

- Only single words considered as context; integration of information from multiple relations as the next step

4. Questions

- Does the SVS approach sufficiently address the problems of scope and scalability? Are there any immediately apparent improvements one might suggest?

- M&L's comparatively simple direct combination model performs relatively well despite failing to consider syntactic relations. Why is that the case? What implications, if any, does that have for the role syntactic relations may play in language processing?