Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing

Angeliki Lazaridou    Eva Maria Vecchi    Marco Baroni

University Of Trento

21 October
EMNLP 2013
Structural Ambiguity

- Different parses of the same sentence are tied to distinct meanings.
- Alternative meanings can lead to rather less semantically plausible interpretations...

Example

*Live fish transporters and fishermen always eat pasta with tuna ...*
Structural Ambiguity

- Different parses of the same sentence are tied to distinct meanings.
- Alternative meanings can lead to rather less semantically plausible interpretations...

Example

**Live fish transporters and fishermen always eat pasta with tuna ...**

- NP bracketing Are we talking about fish transporters that are not dead??
Structural Ambiguity

- Different parses of the same sentence are tied to distinct meanings.
- Alternative meanings can lead to rather less semantically plausible interpretations...

Example

*Live fish transporters and fishermen always eat pasta with tuna...*

- NP bracketing  Are we talking about fish transporters that are not dead??
- PP attachment  Can we use tuna instead of cutlery for eating pasta?
Introduction

Structural Ambiguity

- Different parses of the same sentence are tied to distinct meanings.
- Alternative meanings can lead to rather less semantically plausible interpretations...

Example

**Live fish transporters and fishermen** always eat pasta with tuna ...

- NP bracketing  Are we talking about fish transporters that are not dead??
- PP attachment  Can we use tuna instead of cutlery for eating pasta?
- Coordination  Are both fishermen and fish transporters live???
Correct **syntactic parsing** is steered by **semantic information**. [Fillmore, 1968]
Semantics for parse disambiguation

**Lexical co-occurrence statistics (e.g. PMI)**

Co-occurrence statistics can tell apart syntactically plausible from less plausible constructions.

- NP bracketing [Lauer, 1995, Nakov and Hearst, 2005, Pitler et al., 2010, Vadas and Curran, 2011],
- PP attachment [Lapata and Keller, 2004]
- Full parsing [Bansal and Klein, 2011]

**Compositional Semantic Models**

Syntactically plausible constructions have “better” vectorial representations.

- Full parsing [Le et al., 2013, Socher et al., 2013]
NP Bracketing based on Compositional Semantics Models

(miracle home) run vs miracle (home run)

How semantically plausible is it? vs How semantically plausible is it?

VS lexical co-occurrence statistics vs lexical co-occurrence statistics

[Le et al., 2013, Socher et al., 2013]
[Vecchi et al., 2011]
[Lauer, 1995, Nakov and Hearst, 2005, Pittler et al., 2010, Vadas and Curran, 2011]
Recap

Distributional Semantic Models (DSMs)

- A representation of meaning based on the *Distributional Hypothesis* ...
Recap

Compositional Distributional Semantic Models (cDSMs)

- Represent meaning **beyond words** useful for paraphrase extraction etc.
- Solution à la Frege...

...operationalized in DSM with different **composition functions** of word vectors. [Baroni and Zamparelli, 2010, Coecke et al., 2010, Mitchell and Lapata, 2010, Socher et al., 2012]
Measuring Semantic Plausibility in cDSMs

Plausibility measures inspired by Vecchi et al, 2011

**cosine**: Cosine similarity between composed phrase and head N

**density**: Average similarity between composed phrase and its top 10 neighbors

**entropy**: Entropy calculated from the resulting composed vector

| Low cosine values, less plausible | Low density values, less plausible | High entropy values, less plausible |
|----------------------------------|-----------------------------------|------------------------------------|
| residential steak                | residential steak                 | c1 c2 c3...c101 c102...c1023 c1024 |
| red steak                        | red steak                         |                                   |
| steak                            |                                   |                                   |

Lazaridou, Vecchi, Baroni (University Of Trento)
Noun Phrase Dataset\(^1\)

- **Source:** Penn TreeBank
  - flat structure in NPs
    - always right bracketed
    - e.g. *local (phone company)* but also *blood (pressure medicine)*
  - Incorporate annotations by [Vadas and Curran, 2007a]
- **Extract** **Adj**ecive-**Noun-Noun** and **Noun-Noun-Noun**

| Type of NP | #   | Example                  |
|------------|-----|--------------------------|
| A (N N)    | 1296| *local phone company*    |
| (A N) N    | 343 | *crude oil sector*       |
| N (N N)    | 164 | *miracle home oil*       |
| (N N) N    | 424 | *blood pressure medicine*|
| **Total**  | **2227** | -                        |

\(^1\)\url{http://clic.cimec.unitn.it/~angeliki.lazaridou/datasets/NP_dataset.tar.gz}
Semantic Composition

Basic Composition

| Model            | Composition function                                                                 | Source                                                                 |
|------------------|-------------------------------------------------------------------------------------|----------------------------------------------------------------------|
| weighted additive | $w_1 \vec{crude} + w_2 \vec{oil}$                                                   | [Mitchell and Lapata, 2010]                                          |
| dilation          | $||\vec{crude}||^2 \vec{oil} + (\lambda - 1) \langle \vec{crude}, \vec{oil} \rangle \vec{crude}$ | [Mitchell and Lapata, 2010]                                          |
| full additive     | $W_1 \vec{crude} + W_2 \vec{oil}$                                                   | [Guevara, 2010]                                                      |
| lexical function  | $A_{\vec{crude}} \vec{oil}$                                                        | [Baroni and Zamparelli, 2010]                                       |

- Training phase with DISSECT$^2$ for learning the parameters

$^2$http://clic.cimec.unitn.it/composes/toolkit/
Semantic Composition

Recursive Composition

\[ f(\text{crude oil}, \text{sector}) \]

| Model                  | Composition function                                      | Refs                                  |
|------------------------|----------------------------------------------------------|---------------------------------------|
| weighted additive      | \( w_1 \text{crude oil} + w_2 \text{sector} \)           | [Mitchell and Lapata, 2010]           |
| dilation                | \( ||\text{crude oil}||^2_2 \text{sector} + (\lambda - 1)\langle \text{crude oil}, \text{sector} \rangle \text{crude oil} \) | [Mitchell and Lapata, 2010]           |
| full additive           | \( W_1 \text{crude oil} + W_2 \text{sector} \)          | [Guevara, 2010]                       |
| lexical function       | \( \text{crude oil} + \text{sector} \)                   | [Baroni and Zamparelli, 2010]         |
The task
NP bracketing as binary classification

Goal: (blood pressure) medicine or blood (pressure medicine)?

Alternative bracketings → different composed vectors → different plausibility scores

Feature vector: features extracted from its left and right bracketing.

SVM with Radial Basis Function

Split dataset in 10 folds, 1 for tuning and 9 for cross validation

\[^3\]http://scikit-learn.org/stable/
The baselines

- **Goal**: \((\text{blood pressure}) \text{ medicine}\) or \(\text{blood} (\text{pressure medicine})\)?
- **right**: always right bracketed \(\rightarrow\) \(\text{blood} (\text{pressure medicine})\)
- **pos**: NNN as left and ANN as right bracketed \(\rightarrow\) \((\text{blood pressure}) \text{ medicine}\)
Experimental Setup

The features

Features: $f_{\text{basic}}$

blood pressure medicine

$f(\text{blood})$ \quad $f(\text{pressure})$ \quad $f(\text{medicine})$

density entropy cosine

density entropy cosine
The features

**blood pressure medicine**

Features: $f_{rec}$

![Diagram showing the composition of features for blood pressure medicine]
The features

**blood pressure medicine**

Features: $f_{\text{basic+rec}}$

(blood pressure) medicine blood (pressure medicine)

$\text{density}$ $\text{entropy}$ $\text{cosine}$

$\text{density}$ $\text{entropy}$ $\text{cosine}$
The features

**blood pressure medicine**

Features: **pmi**

\[
\log \frac{P(\text{blood,pressure})}{P(\text{blood})P(\text{pressure})}
\]

\[
\log \frac{P(\text{pressure,medicine})}{P(\text{pressure})P(\text{medicine})}
\]
Results: Compositional semantics vs PMI

| Features          | Accuracy |
|-------------------|----------|
| right             | 65.6     |
| pos               | 77.3     |
| lexfunc\textsubscript{basic} | 74.6     |
| lexfunc\textsubscript{rec}  | 74.0     |
| lexfunc\textsubscript{basic+rec} | 76.2     |
| wadd\textsubscript{basic} | 75.9     |
| wadd\textsubscript{rec}  | 78.2     |
| wadd\textsubscript{basic+rec} | 78.7     |
| pmi               | 81.2     |

- **dil** and **fulladd** outperformed by **right** baseline
- **pos** strong competitor
- **wadd** and **lexfunc** better than current behavior of parsers and comparable to **pos**
- Recursive composition more informative than basic
  - **oil sector** still makes sense, it is crude (oil sector) that refers to a weird concept!
- Semantic plausibility measures not better than **pmi**; 😞
### Results: Compositional semantics vs PMI

| Features          | Accuracy |
|-------------------|----------|
| right             | 65.6     |
| pos               | 77.3     |
| lexfunc<sub>basic</sub> | 74.6     |
| lexfunc<sub>rec</sub>   | 74.0     |
| lexfunc<sub>basic+rec</sub> | 76.2     |
| wadd<sub>basic</sub>  | 75.9     |
| wadd<sub>rec</sub>    | 78.2     |
| wadd<sub>basic+rec</sub> | 78.7     |
| pmi               | 81.2     |

- **dil** and **fulladd** outperformed by **right** baseline
- **pos** strong competitor
- **wadd** and **lexfunc** better than current behavior of parsers and comparable to **pos**
- Recursive composition more informative than basic
  - **oil sector** still makes sense, it is crude (oil sector) that refers to a weird concept!
- Semantic plausibility measures not better than **pmi**; 😞
### Results: Compositional semantics vs PMI

| Features          | Accuracy |
|-------------------|----------|
| right             | 65.6     |
| pos               | 77.3     |
| lexfunc_basic     | 74.6     |
| lexfunc_rec       | 74.0     |
| lexfunc_basic+rec | 76.2     |
| wadd_basic        | 75.9     |
| wadd_rec          | 78.2     |
| wadd_basic+rec    | 78.7     |
| pmi               | 81.2     |

- **dil** and **fulladd** outperformed by **right** baseline
- **pos** strong competitor
- **wadd** and **lexfunc better** than current behavior of parsers and **comparable** to **pos**
- recursive composition more informative than basic
  - **oil sector** still makes sense, it is crude (oil sector) that refers to a weird concept!
- semantic plausibility measures not better than **pmi**; 😞
### Results: Compositional semantics vs PMI

| Features      | Accuracy |
|---------------|----------|
| right         | 65.6     |
| pos           | 77.3     |
| lexfunct_{basic} | 74.6    |
| lexfunct_{rec}  | 74.0     |
| lexfunct_{basic+rec}  | 76.2 |
| wadd_{basic}  | 75.9     |
| wadd_{rec}    | 78.2     |
| wadd_{basic+rec} | 78.7    |
| pmi           | 81.2     |

- **dil** and **fulladd** outperformed by **right** baseline
- **pos** strong competitor
- **wadd** and **lexfunc** better than current behavior of parsers and **comparable** to **pos**
- recursive composition more informative than basic
  - **oil sector** still makes sense, it is **crude (oil sector)** that refers to a weird concept!
- semantic plausibility measures not better than **pmi**; 😞
Results: Compositional semantics vs PMI

| Features           | Accuracy |
|--------------------|----------|
| right              | 65.6     |
| pos                | 77.3     |
| lexfunc<sub>basic</sub> | 74.6     |
| lexfunc<sub>rec</sub>  | 74.0     |
| lexfunc<sub>basic+rec</sub> | 76.2     |
| wadd<sub>basic</sub>  | 75.9     |
| wadd<sub>rec</sub>    | 78.2     |
| wadd<sub>basic+rec</sub> | 78.7     |
| pmi                | 81.2     |

- **dil** and **fulladd** outperformed by **right** baseline
- **pos** strong competitor
- **wadd** and **lexfunc** better than current behavior of parsers and comparable to **pos**
- recursive composition more informative than basic
  - **oil sector** still makes sense, it is crude (oil sector) that refers to a weird concept!
- **semantic plausibility measures not better than pmi 😐**
Results: Compositional semantics combined with PMI

| Features                  | Accuracy |
|---------------------------|----------|
| pmi                       | 81.2     |
| pmi + lexfunc\textsubscript{basic+rec} | 82.9     |
| pmi + wadd\textsubscript{basic+rec}   | 85.6     |

- Error analysis: only 30% of the mistakes between wadd\textsubscript{basic+rec} and pmi are common.
- Combining compositional semantics with pmi significantly ($p < 0.001$) outperforms pmi alone. 😊
- What makes PMI different from compositional semantics?
Hypothesis 1:
- Compositional models are more robust for low frequency NPs, for which PMI estimates will be less accurate.
- $\text{wadd}_{\text{basic+rec}}$ performed 8% better than $\text{pmi}$ on low frequency phrases only.

Hypothesis 2:
- Compositional models can be more useful in cases of weak lexicalization (=low PMI scores)
Conclusions

- Semantic plausibility can improve NP parsing.
- Our approach and current state-of-the-art PMI features are complementary; the combination results in increased performance.

- Extend to full parsing
  - Can we use the same plausibility measures for other kind of headed phrases (e.g. PP-attachment)?
- Need of more plausibility measures.
  - Conduct qualitative evaluation of nearest neighbors of valid and invalid parses of NPs.
Thank you for your attention!

https://sites.google.com/site/lazaridouangeliki/
Bansal, M. and Klein, D. (2011).
Web-scale features for full-scale parsing.
In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 693–702, Portland, Oregon, USA.

Baroni, M. and Zamparelli, R. (2010).
Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space.
In Proceedings of EMNLP, pages 1183–1193, Boston, MA.

Coecke, B., Sadrzadeh, M., and Clark, S. (2010).
Mathematical foundations for a compositional distributional model of meaning.
Linguistic Analysis, 36:345–384.

Fillmore, C. (1968).
The case for case.
In Bach, E. and Harms, R., editors, Universals in Linguistic Theory, pages 1–89. Holt, Rinehart and Winston, New York.

Guevara, E. (2010).
A regression model of adjective-noun compositionality in distributional semantics.
In Proceedings of GEMS, pages 33–37, Uppsala, Sweden.

Lapata, M. and Keller, F. (2004).
The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of nlp tasks.
In HLT-NAACL 2004: Main Proceedings, pages 121–128, Boston, Massachusetts, USA.
Lauer, M. (1995).
Corpus statistics meet the noun compound: some empirical results.
In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 47–54.

Le, P., Zuidema, W., and Scha, R. (2013).
Learning from errors: Using vector-based compositional semantics for parse reranking.
In *Proceedings of the ACL 2013 Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria.

Marcus, M. P. (1980).
*Theory of syntactic recognition for natural languages*.
MIT press.

Mitchell, J. and Lapata, M. (2010).
Composition in distributional models of semantics.
*Cognitive Science*, 34(8):1388–1429.

Nakov, P. and Hearst, M. (2005).
Search engine statistics beyond the n-gram: Application to noun compound bracketing.
In *Proceedings of CoNLL*, pages 17–24, Stroudsburg, PA, USA.

Pitler, E., Bergsma, S., Lin, D., and Church, K. (2010).
Using web-scale n-grams to improve base NP parsing performance.
In *Proceedings of the COLING*, pages 886–894, Beijing, China.
Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. (2013).
Parsing with compositional vector grammars.
In Proceedings of ACL, Sofia, Bulgaria.

Socher, R., Huval, B., Manning, C., and Ng, A. (2012).
Semantic compositionality through recursive matrix-vector spaces.
In Proceedings of EMNLP, pages 1201–1211, Jeju Island, Korea.

Vadas, D. and Curran, J. (2007a).
Adding noun phrase structure to the Penn Treebank.
In Proceedings of ACL, pages 240–247, Prague, Czech Republic.

Vadas, D. and Curran, J. R. (2007b).
Large-scale supervised models for noun phrase bracketing.
In Proceedings of the PACLING, pages 104–112.

Vadas, D. and Curran, J. R. (2011).
Parsing noun phrases in the penn treebank.
Comput. Linguist., 37(4):753–809.
Results

Dependency vs Adjacency PMI

blood pressure medicine

\[
\begin{align*}
\log \frac{P(\text{blood,pressure})}{P(\text{blood})P(\text{pressure})} && \log \frac{P(\text{blood,pressure})}{P(\text{blood})P(\text{pressure})} \\
\log \frac{P(\text{pressure,medicine})}{P(\text{pressure})P(\text{medicine})} && \log \frac{P(\text{blood,medicine})}{P(\text{blood})P(\text{medicine})}
\end{align*}
\]

Figure: Adjacency PMI

Figure: Dependency PMI

- 2 alternative methods in the literature for the calculation of PMI for NP bracketing disambiguation.
  - Adjacency PMI \[\text{[Marcus, 1980]}\]
  - Dependency PMI \[\text{[Lauer, 1995]}\]

- On NPs extracted from Penn TreeBank, the Adjacency model has shown to outperform the Dependency. \[\text{[Vadas and Curran, 2007b]}\]