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?

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???

Structural Ambiguity

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

vs

(miracle home) run

miracle (home run)

[Le et al., 2013,
Socher et al., 2013]

How semantically plausible is it?

How semantically plausible is it?

[Vecchi et al, 2011]

vs

vs

lexical co-occurence statistics

lexical co-occurence statistics

[Lauer, 1995,
Nakov and Hearst, 2005,
Pitler et al., 2010,
Vadas and Curran, 2011]

Recap

Distributional Semantic Models (DSMs)

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

```
he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
 surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
 in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
 But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
 they love the sun , the moon and the stars . None of
```

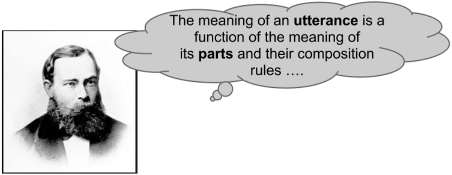
→

	planet	night	full	shadow	shine	crescent
moon	10	22	43	16	29	12
sun	14	10	4	15	45	0
dog	0	4	2	10	0	0

Recap

Compositional Distributional Semantic Models (cDSms)

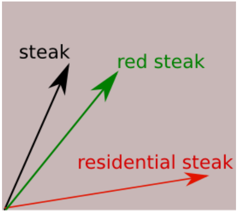
- Represent meaning **beyond words** useful for paraphrase extraction etc.
- Solution à la Frege...



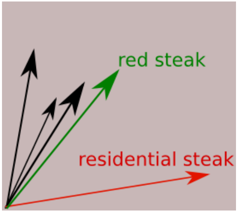
- ...operationalized in DSM with different **composition functions** of word vectors. [Baroni and Zamparelli, 2010, Coecke et al., 2010, Mitchell and Lapata, 2010, Socher et al., 2012]

Measuring Semantic Plausibility in cDSMs

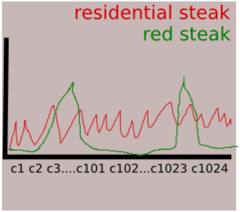
Plausibility measures inspired by Vecchi et al, 2011



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

Noun Phrase Dataset[1]

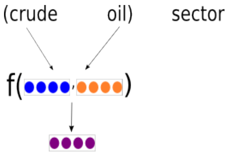
- Source: Penn TreeBank
  - flat structure in NPs
    - always right bracketed
    - e.g. *local (phone company)* but also *blood (pressure medicine)*
  - Incorporate annotations by [Vadas and Curran, 2007a]
- Extract **A**djective-**N**oun-**N**oun and **N**oun-**N**oun-**N**oun

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] http://clic.cimec.unitn.it/~angeliki.lazaridou/datasets/NP_dataset.tar.gz

Semantic Composition

Basic Composition



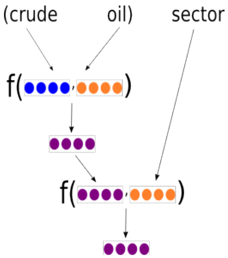
Model	Composition function	
weighted **add**itive	$w_1 \vec{crude} + w_2 \vec{oil}$	[Mitchell and Lapata, 2010]
dilation	$\|\|\vec{crude}\|\|_2^2 \vec{oil} + (\lambda - 1)\langle \vec{crude}, \vec{oil}\rangle \vec{crude}$	[Mitchell and Lapata, 2010]
full additive	$W_1 \vec{crude} + W_2 \vec{oil}$	[Guevara, 2010]
lexical **func**ion	$A_{crude} \vec{oil}$	[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



Model	Composition function	
weighted **add**itive	$w_1 \vec{crude\ oil} + w_2 \vec{sector}$	[Mitchell and Lapata, 2010]
dilation	$\|\|\vec{crude\ oil}\|\|_2^2 \vec{sector} + (\lambda - 1)\langle \vec{crude\ oil}, \vec{sector} \rangle \vec{crude\ oil}$	[Mitchell and Lapata, 2010]
full additive	$W_1 \vec{crude\ oil} + W_2 \vec{sector}$	[Guevara, 2010]
lexical **func**ion	$\vec{crude\ oil} + \vec{sector}$	[Baroni and Zamparelli, 2010]

The task

NP bracketing as binary classification

blood pressure medicine

- **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[3]
- Split dataset in 10 folds, 1 for tuning and 9 for cross validation

[3] `http://scikit-learn.org/stable/`

The baselines

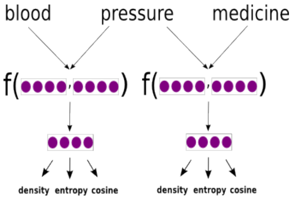
blood pressure medicine

- **Goal:** *(blood pressure) medicine* or *blood (pressure medicine)*?
- **right**: always right bracketed → *blood (pressure medicine)*
- **pos**: NNN as left and ANN as right bracketed → *(blood pressure) medicine*

The features

blood pressure medicine

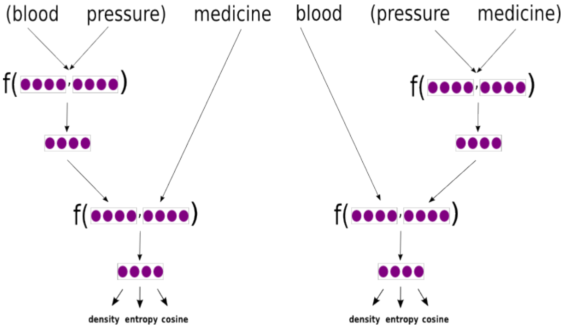
Features: $\mathbf{f_{basic}}$



The features

blood pressure medicine

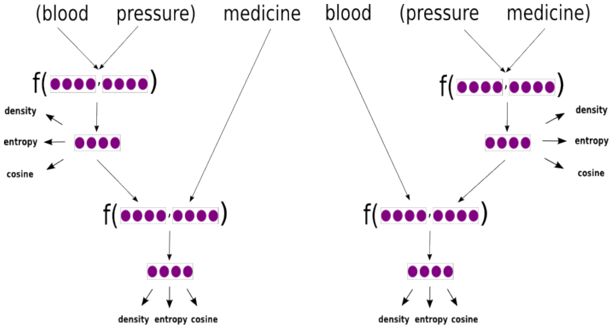
Features: $\mathbf{f}_{\mathbf{rec}}$



The features

blood pressure medicine

Features: $\mathbf{f}_{\mathbf{basic+rec}}$



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$_{\textbf{basic}}$	74.6
lexfunc$_{\textbf{rec}}$	74.0
lexfunc$_{\textbf{basic+rec}}$	76.2
wadd$_{\textbf{basic}}$	75.9
wadd$_{\textbf{rec}}$	78.2
wadd$_{\textbf{basic+rec}}$	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$_{basic}$	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
lexfunc$_{\text{basic}}$	74.6
lexfunc$_{\text{rec}}$	74.0
lexfunc$_{\text{basic+rec}}$	76.2
wadd$_{\text{basic}}$	75.9
wadd$_{\text{rec}}$	78.2
wadd$_{\text{basic+rec}}$	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
$\text{lexfunc}_{\text{basic}}$	74.6
$\text{lexfunc}_{\text{rec}}$	74.0
$\text{lexfunc}_{\text{basic+rec}}$	76.2
$\text{wadd}_{\text{basic}}$	75.9
wadd_{rec}	78.2
$\text{wadd}_{\text{basic+rec}}$	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$_{\textbf{basic}}$	74.6
lexfunc$_{\textbf{rec}}$	74.0
lexfunc$_{\textbf{basic+rec}}$	76.2
wadd$_{\textbf{basic}}$	75.9
wadd$_{\textbf{rec}}$	78.2
wadd$_{\textbf{basic+rec}}$	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 combined with PMI

Features	Accuracy
pmi	81.2
pmi+lexfunc$_{\mathbf{basic+rec}}$	82.9
pmi+wadd$_{\mathbf{basic+rec}}$	**85.6**

- Error analysis: only 30% of the mistakes between **wadd$_{\mathbf{basic+rec}}$** and **pmi** are common.
- Combining compositional semantics with **pmi** significantly ($p < 0.001$) outperforms **pmi** alone. ☺
- What makes PMI different from compositional semantics?

Results: Compositional semantics combined with PMI

- Hypothesis 1:
  - Compositional models are more robust for **low frequency NPs**, for which PMI estimates will be less accurate.
  - $\mathbf{wadd}_{\mathbf{basic+rec}}$ performed 8% better than **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/

References I

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.

References II

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.

References III

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.

Dependency vs Adjacency PMI

blood pressure medicine

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

$$\log \frac{P(\text{pressure,medicine})}{P(\text{pressure})P(\text{medicine})}$$

Figure: Adjacency PMI

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

$$\log \frac{P(\text{blood,medicine})}{P(\text{blood})P(\text{medicine})}$$

Figure: Dependency PMI

- 2 alternative methods in the literature for the calculation of PMI for NP bracketing disambiguation.
  - Adjacency PMI [Marcus, 1980]
  - Dependency PMI [Lauer, 1995]
- On NPs extracted from Penn TreeBank, the Adjacency model has shown to outperform the Dependency. [Vadas and Curran, 2007b]