Learning Semantic Composition to Detect Non-compositionality of Multiword Expressions

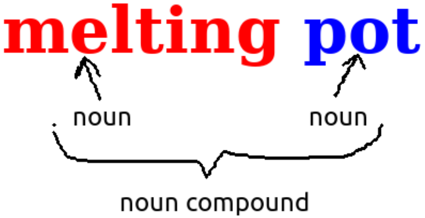
Majid Yazdani, Meghdad Farahmand, James Henderson

January 19, 2019

The concept of noun compound compositionality

melting pot

The concept of noun compound compositionality



The concept of noun compound compositionality

melting pot

1. melt: to become or make something become liquid as a result of heating + pot: a deep round container used for cooking things in
2. melting pot: a place or situation in which large numbers of people, ideas, etc. are mixed together

The concept of noun compound compositionality

melting pot

Should we count the phrase as one semantic unit or separate?

Compositional and non-compositional compounds

A compound is called **compositional** if it can be semantically derived from its components (e.g. taxi driver)
Otherwise it is called **non-compositional** (e.g. silk road)

Related work

- ADD: $a.s1 + b.s2 = s3$
- MULT: $a.s1.s2 = s3$
- COMB: $a.s1 + b.s2 + c.s1.s2 = s3$
- WORD1: $a.s1 = s3$
- WORD2: $a.s2 = s3$

$$(a\mathbf{v1} + b\mathbf{v2})_i = a.\mathrm{v1}_i + b.\mathrm{v2}_i$$

$$(\mathbf{v1v2})_i = \mathrm{v1}_i.\mathrm{v2}_i$$

Figure: Score-based and vector-based baseline models from Reddy et al.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In IJCNLP, pages 210–218

Related work

Model	ρ	R^2
ADD	**0.686**	0.613
MULT	0.670	0.428
COMB	0.682	**0.615**
WORD1	0.669	0.548
WORD2	0.515	0.410
a**v1**+b**v2**	**0.714**	**0.620**
v1v2	0.650	0.501

Figure: Spearman's ρ and determination coefficient of presented models

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In IJCNLP, pages 210–218

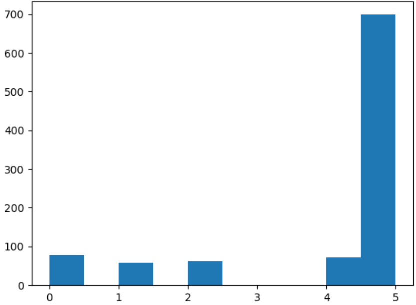
Briefly on dataset

Unannotated part:
70k noun-noun compounds from english Wikipedia dump

Annotated part:
1042 selected compounds with human evaluations
0 - "idiomatic" (**non-compositional**)
5 - "literal" (**compositional**)

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In Proceedings of the 11th Workshop on Multiword Expressions (MWE- NAACL 2015). Association for Computational Lin- guistics

Briefly on dataset



non-compositional $\longrightarrow$ compositional

Briefly on dataset

Around **80%** of compounds (vast majority) are **compositional**

Research concepts

1. Most compounds can be counted as compositional
2. Compositional compounds are decomposable to components; they should be *independently* predictable

Research concepts

A projection f for compound components w_i, w_j can be learned such that error

$$e_{ij} = ||f(\phi(w_i), \phi(w_j)) - \phi(w_i, w_j)||$$

is minimized for given examples

Research concepts

(non-)compositionality detection:

- Low error — projection fits compound well, **compositional**
- High error — projection does not fit compound well, **non-compositional**

Linear projection

$$f(\phi(w_i), \phi(w_j)) = [\phi(w_i), \phi(w_j)]\theta_{2d \times d}$$

Minimizing functional

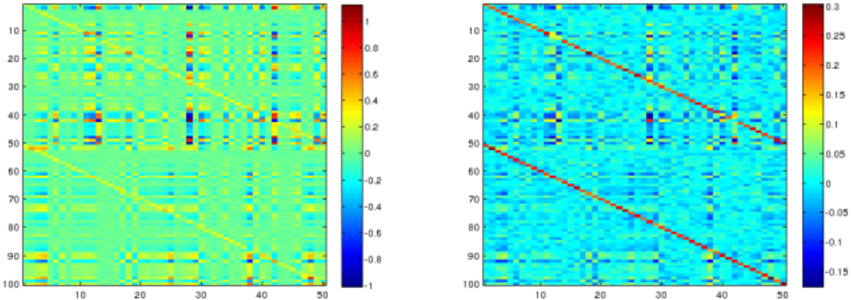
$$\min_{\theta} ||[\phi(w_i), \phi(w_j)]\theta_{2d \times d} - \phi(w_i, w_j)||$$

Via mutli-variant linear regression

Matrix sparsity via L1-regularization

$$\min_{\theta} ||[\phi(w_i), \phi(w_j)]\theta - \phi(w_i, w_j)|| + \lambda|\theta|$$

Matrix sparsity via L1-regularization



(a) Linear Projection

(b) Sparse Linear Projection

Figure: Linear transformation matrix of compositionality for embeddings of size 50

Polynomial (quadratic) projection

$$f(\phi(w_i), \phi(w_j)) = \psi([\phi(w_i), \phi(w_j)])\theta_{2d \times d}$$

In quadratic case

$$\psi(x) = x_1^2, \cdots x_n^2, x_1 x_2, \cdots x_{n-1} x_n, x_1, \cdots x_n$$

$x_1^2, \cdots x_n^2$ — **pure quadratic** terms
$x_1 x_2, \cdots x_{n-1} x_n, x_1$ — **interaction** terms
$x_1, \cdots x_n$ — **linear** terms

Feed-forward neural networks

$$f(\phi(w_i), \phi(w_j)) = \sigma([\phi(w_i), \phi(w_j)]W_{ih})W_{ho}$$

Experimental results

Model	Spearman ρ	NDCG	F_1
ADD (Reddy et al.)	0.2083	0.8139	0.3695
MULT (Reddy et al.)	0.0918	0.7600	0.3561
Sparse Linear	0.3758	0.8425	0.4640
Linear	0.3809	0.8425	0.4641
Sparse Pure Quad.	0.3785	0.8411	0.4705
Pure Quad.	0.3857	0.8468	0.4701
Sparse Interaction	**0.4103**	**0.8582**	**0.4871**
Interaction	0.4025	0.8569	0.4864
Quadratic	0.4025	0.8559	0.4834
Sparse NN	0.3708	0.8504	0.4635
NN	0.3751	0.8497	0.4547

Figure: Results for each model's ability to predict non-compositionality

Auto-reconstructive models

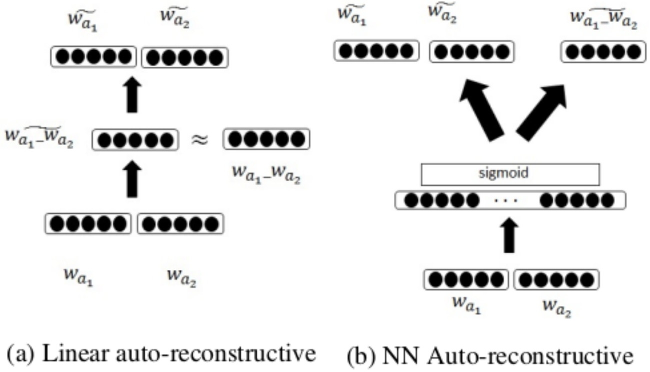


Figure: Auto-reconstructive linear and neural network models

Auto-reconstructive linear models

Minimizing functional

$$\min_{\theta,\theta'} ||X\theta - Y|| + \lambda||A\theta\theta' - A||$$

Y — matrix of precomputed compound embeddings
X — concatenations of components' embeddings for these compounds
A — concatenation of components' embeddings of all compounds

Auto-reconstructive neural networks

Minimizing functional

$$\min_{W_{ih}, W_{hi}, W_{oh}} ||\sigma(XW_{ih})W_{ho} - Y|| + \lambda||\sigma(AW_{ih})W_{hi} - A||$$

Experimental results

Model	Spearman ρ	NDCG	F_1
Linear	0.3809	0.8425	0.4641
Linear + auto	0.3752	0.8455	0.4655
Interaction	0.4025	0.8569	0.4864
Interaction + auto	0.3929	0.8571	0.4895
NN	0.3740	0.8450	0.4634
NN + auto	0.3998	0.8517	0.4912

Figure: Results comparing the auto- reconstructive models' ability to predict non-compositionality.

Non-compositionality detection using latent annotations

Solving optimization problem

$$\min_{\lambda_{ij},\theta} \sum_{ij} \lambda_{ij} e_{ij}^2$$

With following constraints:

$$\sum_{ij} \lambda_{ij} = N - B$$

$$\lambda_{ij} \in \{0, 1\}$$

N — number of all compounds
B — desired number of **non-compositional** compounds
"EM-like" algorithm

Experimental results

Model	Spearman ρ	NDCG	F_1
Linear	0.3809	0.8425	0.4641
Linear + la	0.3780	0.8460	0.4629
Interaction	0.4025	0.8569	0.4864
Interaction + la	0.4056	0.8630	0.4834
NN	0.3740	0.8450	0.4634
NN + la	0.3923	0.8536	0.4815

Figure: Results comparing the latent annotation models' ability to predict non-compositionality.

Conclusions

- For semantic composition, polynomial projections tend to do better than linear ones
- Sparsity via L1-regularization generally improves detection quality
- Auto-reconstructive models — improvement debatable
- Latent annotations — some further improvement

Discussion time