Hierarchical Embeddings for Hypernymy Detection and Directionality

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, Ngoc Thang Vu

Institute for Natural Language Processing - University of Stuttgart, Germany

{nguyenkh, koepermn, schulte, thangvu}@ims.uni-stuttgart.de

Motivation

- **Hypothesis**: each common context of a hyponym–hypernym relation is an indicator to determine which of two words is semantically more general.
- **Goal**: learn the hierarchical embeddings for hypernymy **detection** and **directionality**.
- **Procedure**: **strengthen** the distributional **similarity** of **hypernym pairs** and generate a distributional **hierarchy** between hyponyms and hypernyms.

Contributions

1. Propose a novel neural model *HyperVec* to learn hierarchical embeddings for hypernymy addressing detection and directionality tasks.
2. Present an unsupervised measure to score hypernym relations based on *HyperVec*.
3. The *HyperVec* is able to generalize over unseen hypernymy pairs.
4. The *HyperVec* outperforms both state-of-the-art unsupervised measures and embedding models.

Models

1. Hierarchical Hypernymy:

- Learn hierarchical embeddings in a specific order. The similarity score for hypernymy is higher than the similarity score for other relations:

$$\mathrm{L}_{(w,c)} = \frac{1}{\#(w,u)} \sum_{u \in \mathbb{H}^{+}(w,c)} \partial(\vec{w}, \vec{u})$$

- Learn the distributional hierarchy between hypernyms and hyponyms, as an indicator to differentiate between hypernym and hyponym:

$$\mathrm{L}_{(v,w,c)} = \sum_{v \in \mathbb{H}^{-}(w,c)} \partial(\vec{v}, \vec{w})$$

- Incorporate the Skip-gram with negative sampling model:

$$\mathrm{J}_{(w,c)} = \#(w,c) \log \sigma(\vec{w}, \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w}, \vec{c}_N)]$$

- The final objective functionis defined as follows

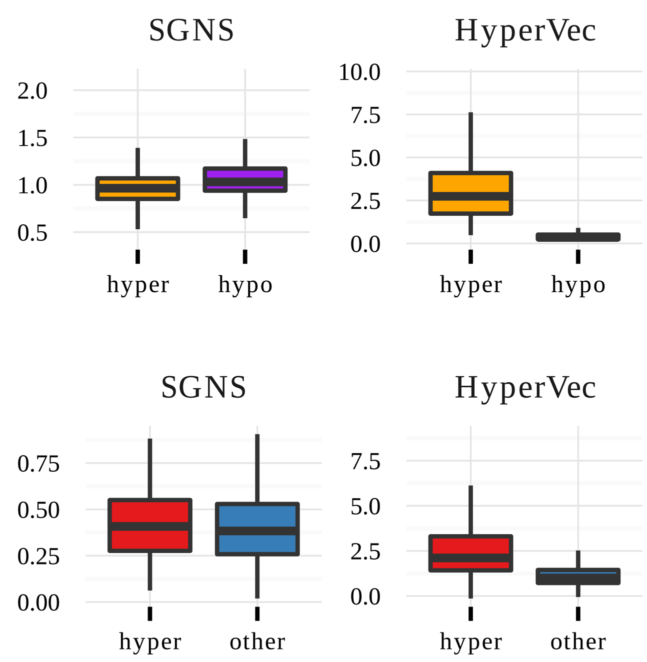
$$\mathrm{J} = \sum_{w \in V_W} \sum_{c \in V_C} \mathrm{J}_{(w,c)} + \mathrm{L}_{(w,c)} + \mathrm{L}_{(v,w,c)}$$

2. Unsupervised Hypernymy Measure:

- *HyperVec* shows the two following properties:

1. high similarity between hypernyms and hyponyms.
2. hierarchy between hypernyms and their hyponyms.

- The measure is defined as follows:
$HyperScore(u,v) = cos(\vec{u}, \vec{v}) * \frac{\|\vec{v}\|}{\|\vec{u}\|}$



Unsupervised Hypernymy Detection and Directionality

Dataset	Baseline	HyperScore
EVALution	0.353	**0.538**
BLESS	0.051	**0.454**
Lenci&Benotto	0.382	**0.574**
Weeds	0.441	**0.850**

	BLESS	WBLESS	BIBLESS
Kiela et al.	0.88	0.75	0.57
Santus et al.	0.87	–	–
Weeds et al.	–	0.75	–
SGNS	0.44	0.48	0.34
HyperVec	**0.92**	**0.87**	**0.81**

Experimental Settings

- ENCOW14A corpus ≈14.5 billion tokens.
- Baseline: default SGNS (word2vec).
- 100 dimensions, window 5, negative samples:15, learning rate 0.025.
- Learn HyperVec for nouns and verbs.

Supervised Classification

- **SVM classifier** based on four components: conc. + diff. + cos + magnitude(hyper)

Models	BLESS	ENTAILMENT
Yu et al. (2015)	0.90	0.87
Tuan et al. (2016)	0.93	0.91
HyperVec	**0.94**	**0.91**

Graded Entailment

- *HyperLex*: dataset of graded lexical entailment.
- Provides soft lexical entailment on a continous scale e.g *duck-animal* is 5.6 out of 6.0 but reversed *animal-duck* is only 1.0.
- 2 616 word pairs, seven semantic relations, and two word classes (nouns and verbs).
- We compared HyperScore against the most prominent state-of-the-art models.

Measures		Embeddings	
Model	ρ	Model	ρ
FR	0.279	SGNS	0.205
DEM	0.180	PARAGRAM	0.320
SLQS	0.228	OrderEmb	0.191
WN	0.234	Word2Gauss	0.206
VIS	0.209	HyperScore	**0.540**

Generalizing Hypernymy I

Motivation: explore HyperVecs potential for generalization

- Rely on a small seed set only, rather than using a large set of training data
- Learn only based on the 200 concepts (and their hyponyms) from the **BLESS** dataset
- Performance measured using Average Precision (AP) ranking measure

Dataset	Baseline	HyperScore
EVALution	0.353	**0.390**
Lenci/Benotto	0.382	**0.448**
Weeds	0.441	**0.585**

Generalizing Hypernymy II

- project (default) representations from any arbitrary language into our modified English HyperVec space
- mapping function between source and target space using least-squares error method
- $DE \rightarrow EN$ and $IT \rightarrow EN$ word translations based on Europarl
- compare the original vs. mapped representation on hypernymy ranking retrieval task

German	Hyp/All	Hyp/Syn	Hyp/Ant
DE → SGNS	0.28	0.48	0.40
DE → $\mathrm{EN}_{HyperScore}$	**0.37**	**0.65**	**0.47**
Italian	**Hyp/All**	**Hyp/Syn**	**Hyp/Ant**
IT → SGNS	0.38	0.50	0.60
IT → $\mathrm{EN}_{HyperScore}$	**0.44**	**0.57**	**0.65**