Domain-Specific Word Embeddings with Structure Prediction

David Lassner[*,1,2], Stephanie Brandl[*,1,2,3], Anne Baillot[4], Shinichi Nakajima[1,2,5]

[*]Authors contributed equally.

[1]TU Berlin, [2]BIFOLD, [3]University of Copenhagen, [4]Le Mans Université, [5]RIKEN Center for AIP

eMail: davidlassner@gmail.com, brandl@di.ku.dk

Introduction

- 2 new methods to calculate dynamic word embeddings, e.g., across time or domain:
  - *Word2Vec with Structure Constraint (W2VConstr)*:
    Domain-specific embeddings are learned under regularization of a given structure.
  - *Word2Vec with Structure Prediction (W2VPred)*:
    Domain-specific embeddings and sub-corpora structure are learned simultaneously.

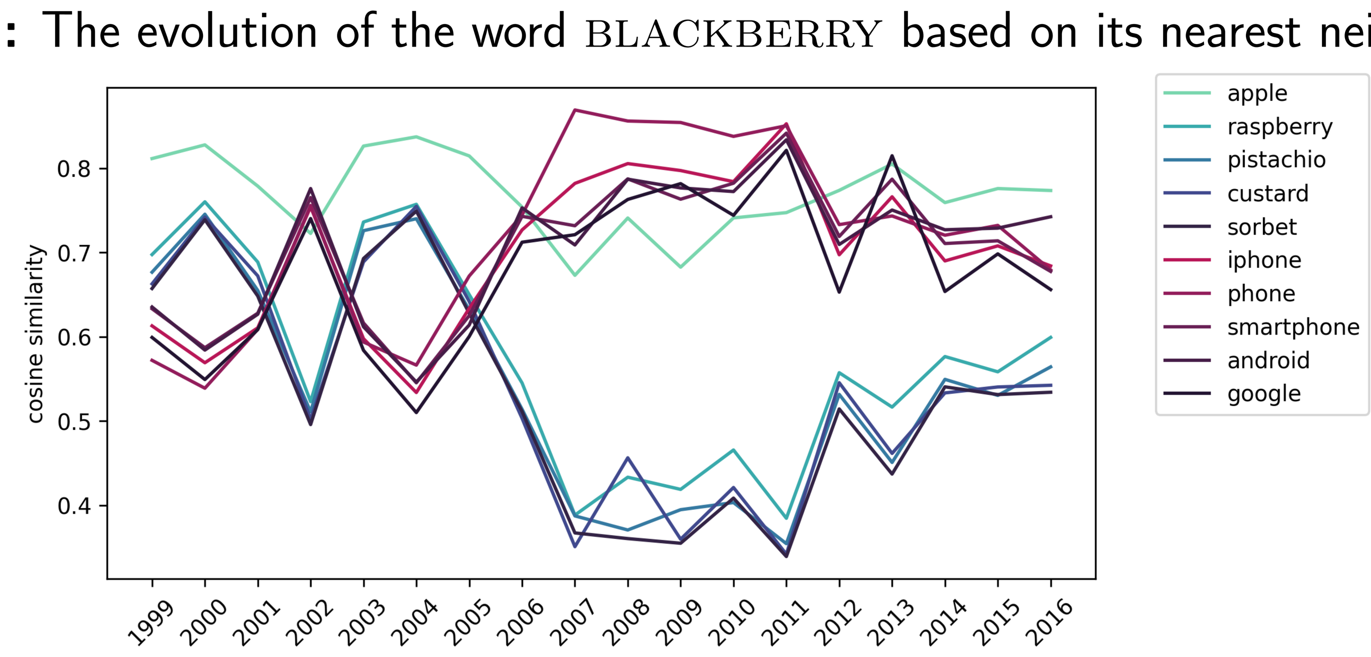
$$\min_{U_t} L_{\mathrm{F}} + \tau L_{\mathrm{RD}} + \lambda L_{\mathrm{S}}, \quad \text{where}$$

$$L_{\mathrm{F}} = \left\| Y_t - U_t U_t^\top \right\|_F^2, L_{\mathrm{RD}} = \left\| D \right\|_F, L_{\mathrm{S}} = \sum_{t'=1}^{T} W_{t,t'} D_{t,t'}.$$

- We test our methods on 4 different datasets with different structures (sequences, trees and general graphs), domains (news, wikipedia, high literature) and languages (en and de).
- We show how W2VPred can be used in an explorative setting to raise novel research questions in the field of Digital Humanities.

Example

Fig.1: The evolution of the word BLACKBERRY based on its nearest neighbors



Data

New York Times

- English news articles
- headlines, lead texts and paragraphs
- published online and offline
- Jan 1990-June 2016
- 100k articles

Wiki Field of Science

- English Wikipedia
- categories selected by fields of science and technology (OECD)
- 4 clusters: 1 main category + 3 subcategories
- 226k articles

Wikipedia Philosophy

- English Wikipedia
- categories in *philosophy*
- 5 main categories + 2 largest subcategories for each
- 41k articles

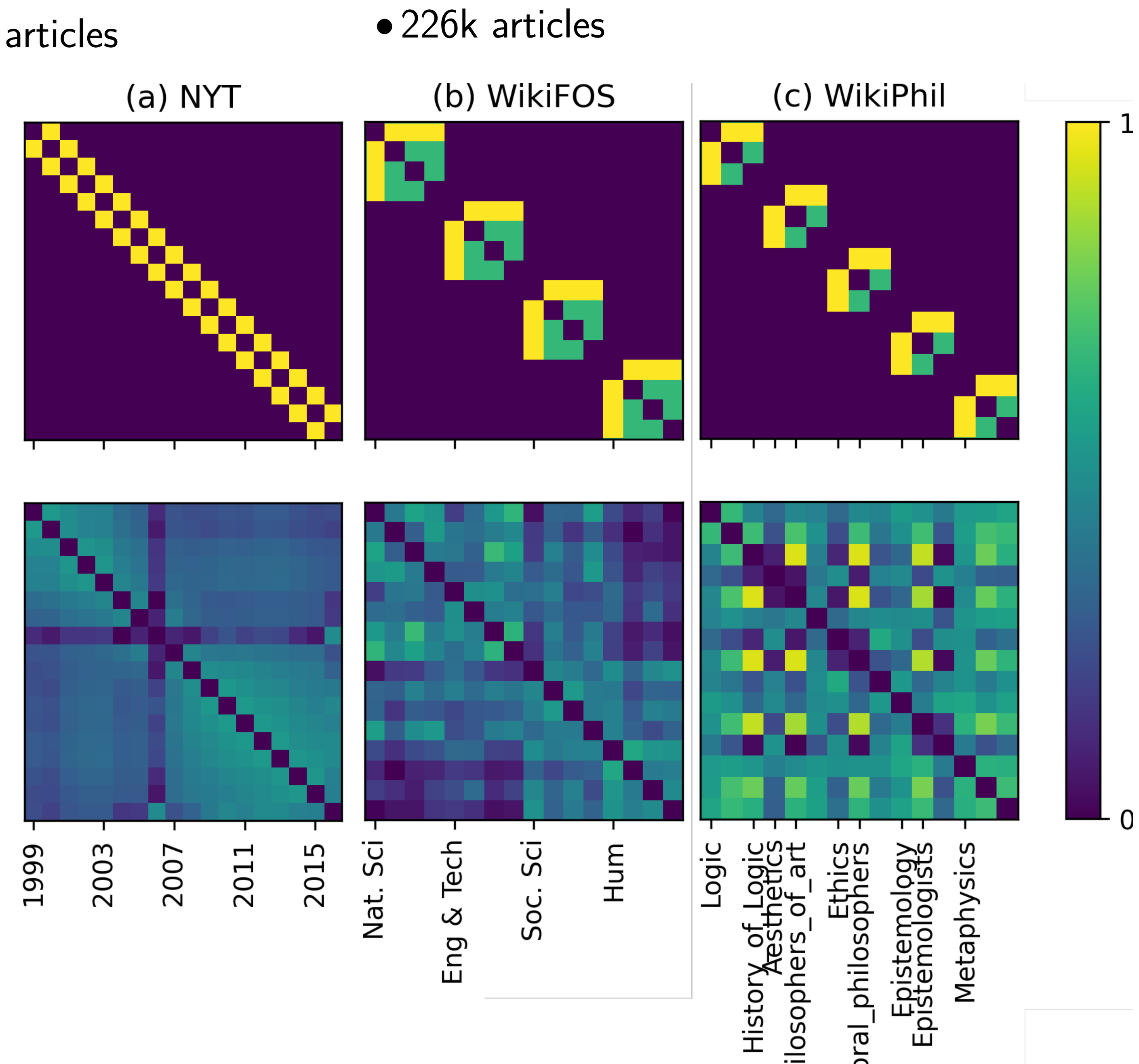


Fig.2: Prior affinity matrix W used for W2VConstr (upper), and the estimated affinity matrix by W2VPred (lower) where the number indicates how close slices are (1: identical, 0: very distant). The estimated affinity for NYT implies the year 2006 is an outlier. We checked the corresponding articles and found that many paragraphs and tokens are missing in that year. Note that the diagonal entries do not contribute to the loss for all methods.

Results

Five nearest neighbors to the word "power"

Natural Science	Eng&Tech	Social Science	Humanities	GloVe	Skip-Gram
generator	generator	powerful	powerful	control	Power
PV	inverter	control	control	supply	inverter
thermoelectric	alternator	wield	counterbalance	capacity	mover
inverter	converter	drive	drive	system	electricity
converter	electric	generator	supreme	internal	thermoelectric

Table 1: Five nearest neighbors to the word "power" in the domain-specific embedding space, learned by W2VPred, of four main categories of WikiFoS (left four columns), and in the general embedding space learned by GloVe and Skip-Gram on the entire dataset (right-most columns, respectively).

New York Times

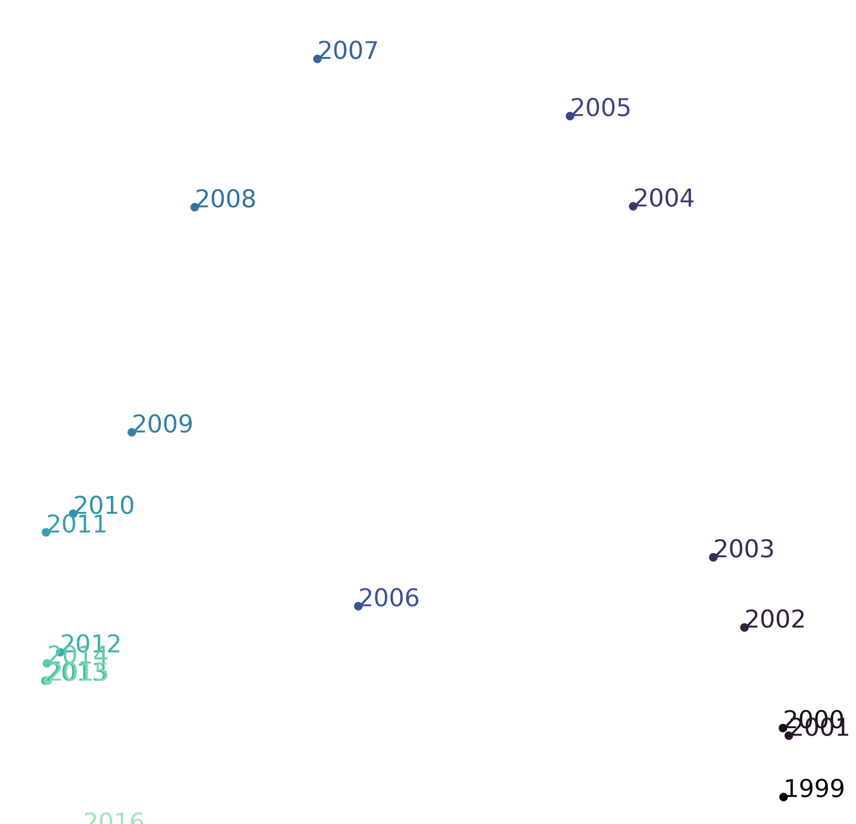


Fig.3: Spectral embeddings of all years in the NYT dataset (after applying W2VPred) shows that 2006 is an outlier. The reason for this might be that the original dataset (1990-2005) has been extended and many paragraphs are missing while the number of articles are the same.

Wikipedia Field of Science

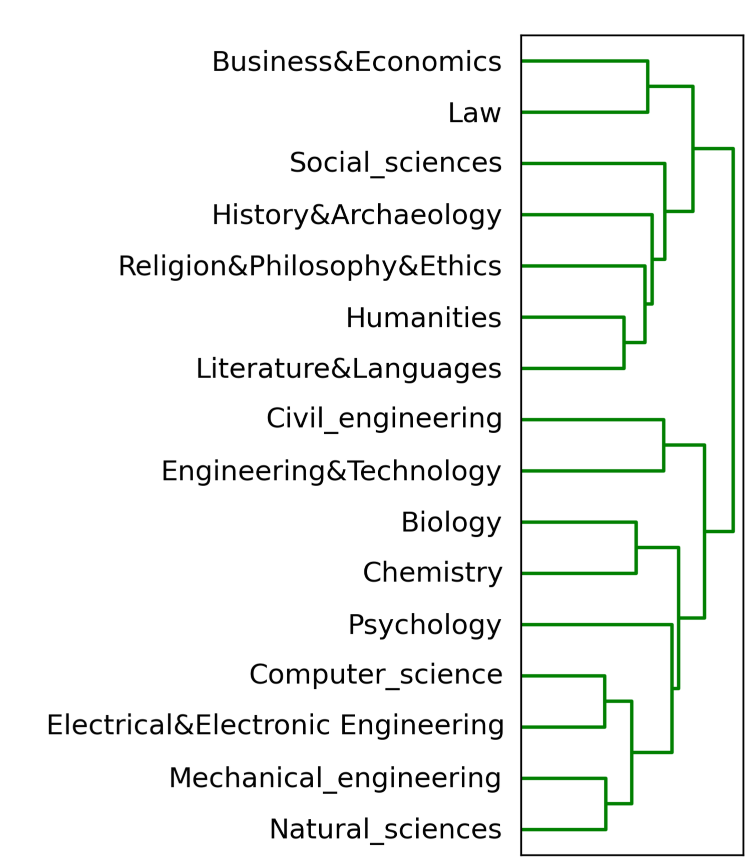


Fig.4 Dendrogram for the categories of Wikipedia FoS. The main two cluster contain the categories from the two related main categories:
#1 Humanities & Social Sciences,
#2 Natural Science & Engineering.

Wikipedia Philosophy

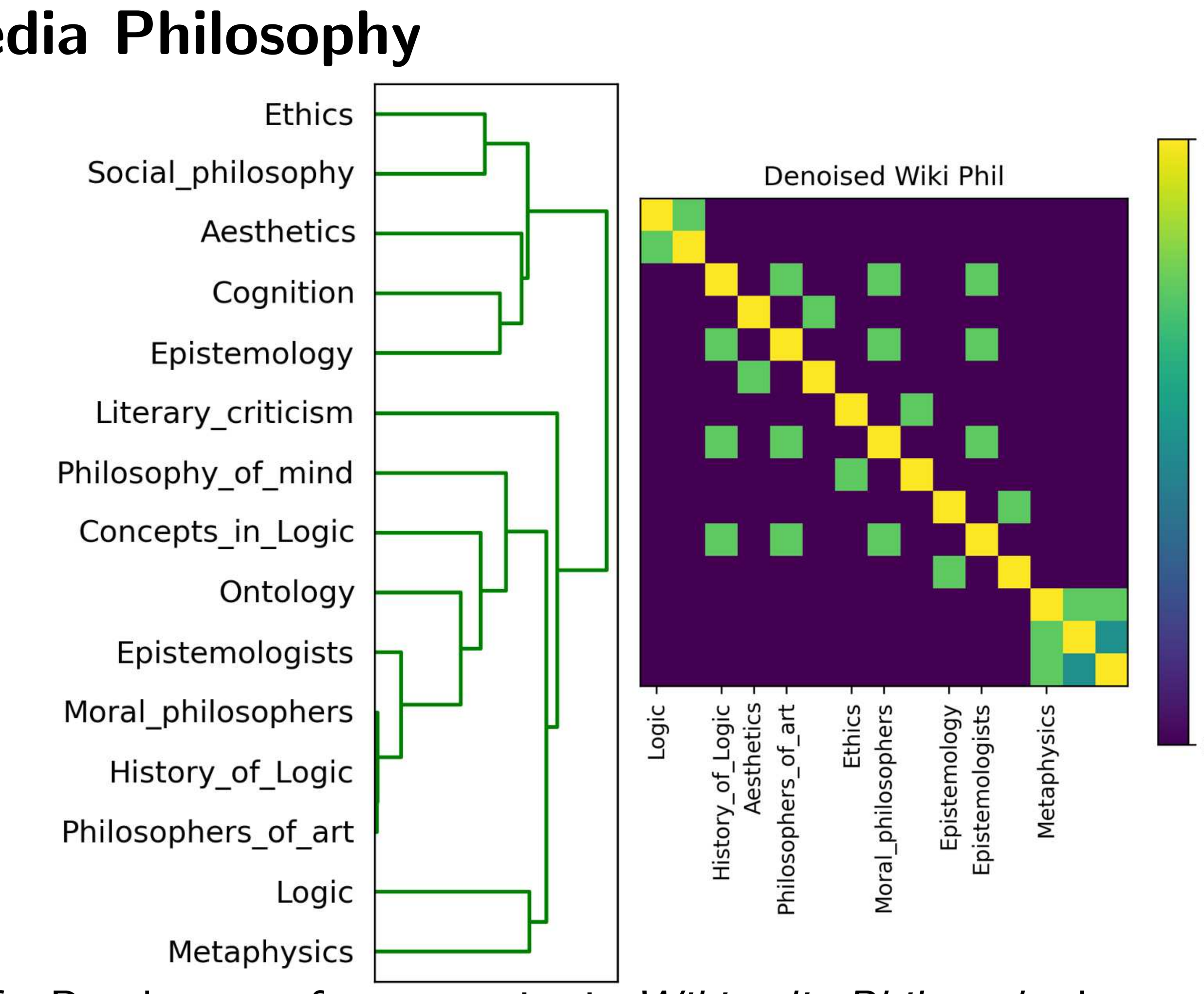


Fig.5 *Left*: Dendrogram for categories in *Wikipedia Philosophy* learned by W2VPred based on the affinity matrix W. *Right*: Denoised Affinity matrix built from the learned structure by W2VPred. Newly formed Cluster includes *History of Logic*, *Moral Philosophers*, *Epistemologists*, and *Philosophers of Art*.

Conclusion

Word2Vec with Structure Constraint (W2VConstr):

– if knowledge about prior structure is known or can be assumed.

Word2Vec with Structure Prediction (W2VPred):

– Can be used to predict structure, including outlier detection.

&: to denoise and update prior structure.