Idioms, Probing and Dangerous Things: Towards Structural Probing for Idiomaticity in Vector Space

Filip Klubička, Vasudevan Nedumpozhimana, John D. Kelleher

6th May 2023

19th Workshop on Multiword Expressions @ EACL, Dubrovnik, Croatia

ADAPT Engaging Content Engaging People

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH TU DUBLIN TECHNOLOGICAL UNIVERSITY DUBLIN

A World Leading SFI Research Centre

Science Foundation Ireland SFI For what's next

HOST INSTITUTION Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin

HOST INSTITUTION DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

PARTNER INSTITUTIONS TU Dublin; University College Dublin An Coláiste Ollscoile, Baile Átha Cliath Ireland's Global University; MTU Ollscoil Teicneolaíochta na Mumhan Munster Technological University; TUS; Maynooth University National University of Ireland Maynooth; Ollscoil na Gaillimhe University of Galway

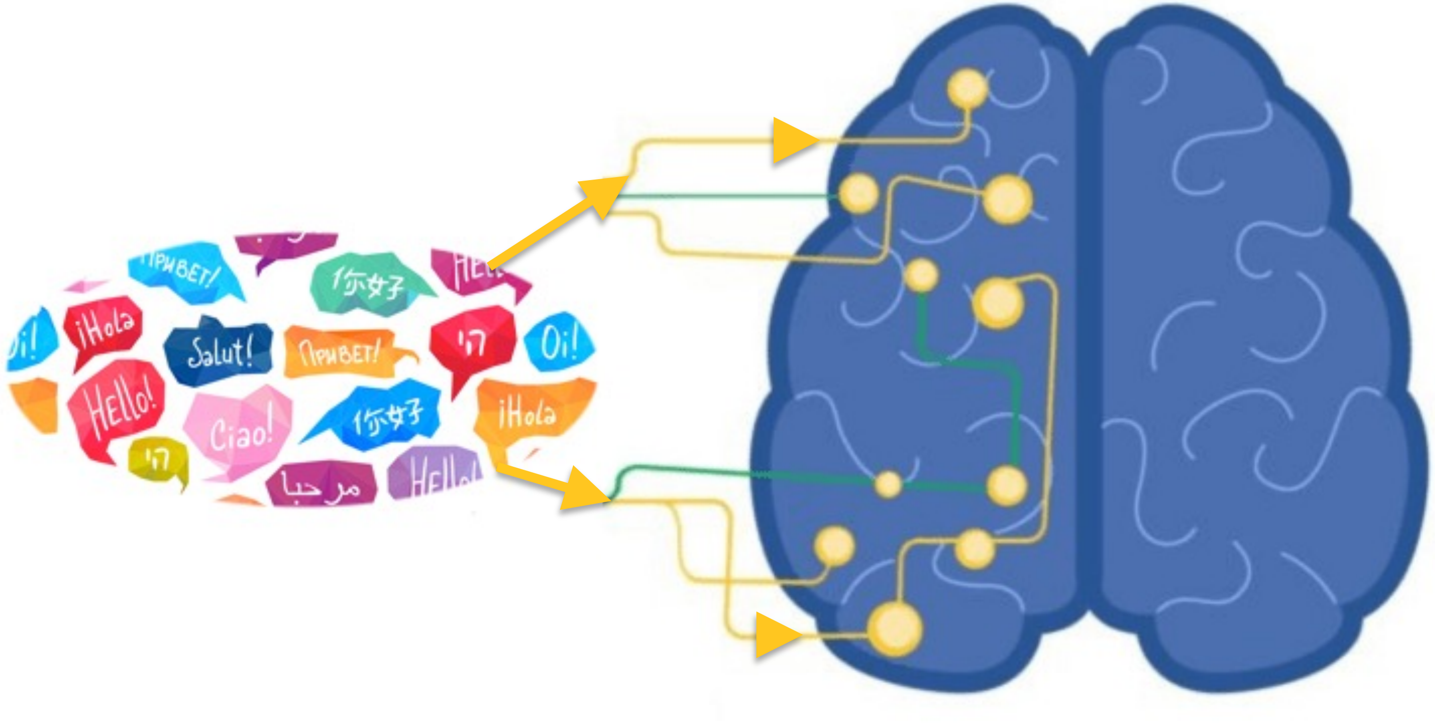
I. Introduction

Engaging Content
Engaging People

Semantics

- semantics
- meaning
- literal vs. figurative
- idiomaticity

MEANING



Natural Language Processing

- computational semantics
- meaning representations
- vector space models
- embeddings (word2vec, GLOVE…)
- language models (BERT, GPT-4…)

Interpretability

- explainable AI
- interpretable models
- BlackboxNLP (Alishahi et al. 2019)
- probing framework

Research questions

- How is idiomaticity structurally encoded in vector space?
- Does the vector norm play a role in encoding idiomatic information?
- Is idiomatic usage encoded similarly to contextual incongruity?

Approach:

- apply the *probing with noise* framework
- repurpose existing idiom dataset into a probing dataset
- examine structural properties of a static and contextual encoder

II. Probing with Noise

Engaging Content
Engaging People

Probing with Noise (Klubička and Kelleher, 2022)

1. Choose a linguistic property of interest, e.g. verb tense
2. Choose or design an appropriate dataset
3. Choose a word/sentence representation, e.g. BERT
4. Choose a probing classifier (i.e. the probe), e.g. MLP
5. Train the probe on the embeddings as input
6. Evaluate the probe's performance on the task (vanilla baseline)
7. Introduce systematic noise in the embedding
8. Repeat training, evaluate and compare

Information containers

Embeddings = Vectors

- vectors = direction + magnitude
- direction (coordinates) defined by dimension values
- magnitude (length) defined by vector norm

vector								norm
10	5	-2	4	-8	1	2	5	37

- two information containers
  - vector **dimensions**
  - vector **norm**

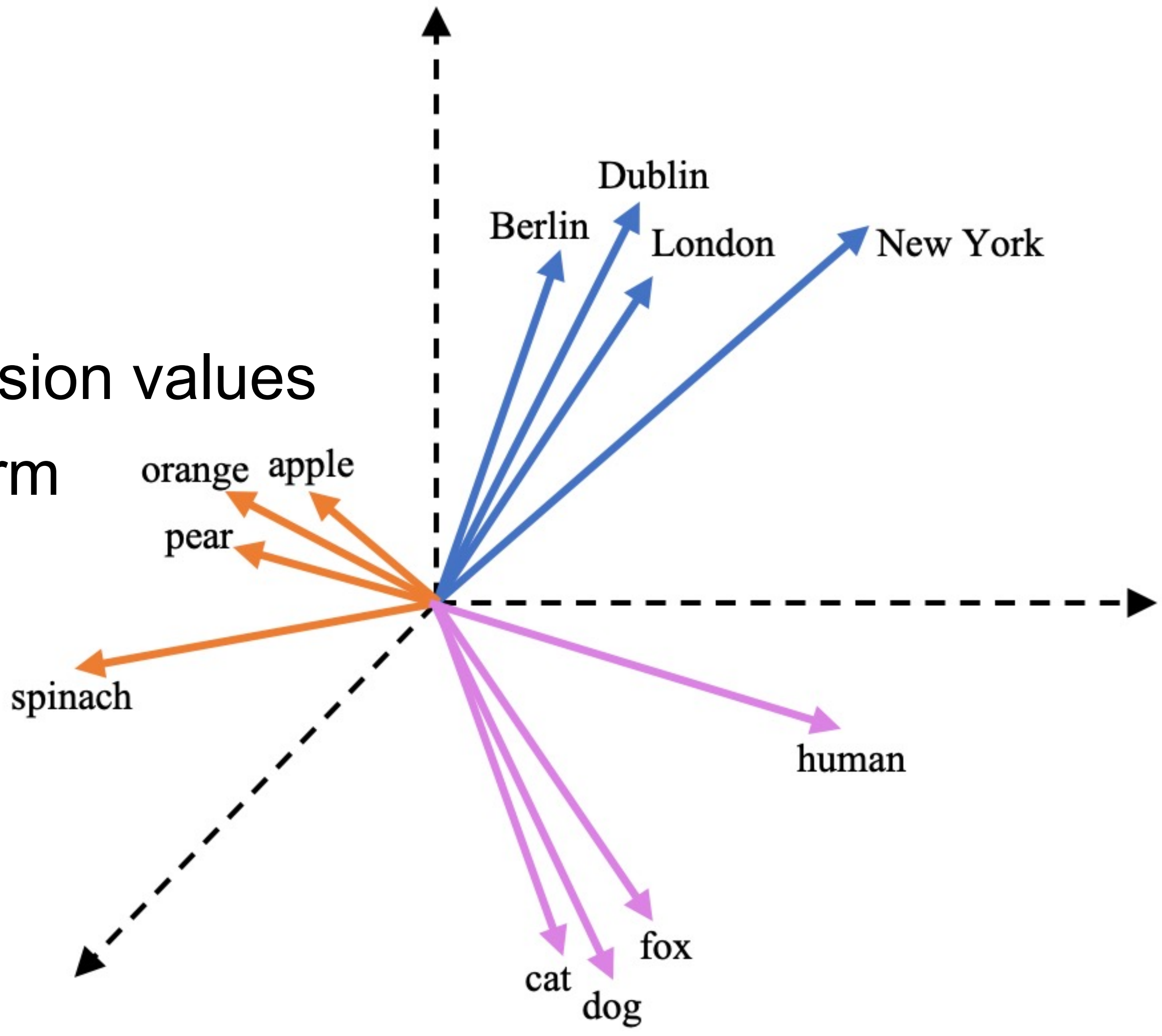


Figure 1. An illustrative example of a vector space model.

III. Idiom Probing Dataset

Engaging Content
Engaging People

Motivation

- a probing task needs to ask a simple, non-ambiguous question
- probing for idiomatic usage:
  - requires a simple task that can directly tease out idiomaticity
  - requires sentence-level instances
  - requires same idiomatic phrase used both literally and idiomatically
- VNC-Tokens dataset (Cook et al., 2008)
  - English Verb-Noun (Idiomatic) Combinations
    - *e.g. hit road, pull plug, make mark*
  - 1205 sentences, 28 VN(I)Cs
    - *749 Idiomatic usage*
    - *456 Literal usage*

Chosen train-test split

VNC	Train set Total	Train set Idiomatic	VNC	Test set Total	Test set Idiomatic
blow top	28	23			
blow trumpet	29	19			
blow whistle	78	27			
get sack	50	43			
get nod	26	23			
get wind	28	13			
hit road	32	25			
hit roof	18	11	cut figure	43	36
hit wall	63	7	find foot	53	48
lose head	40	21	have word	91	80
lose thread	20	18	hold fire	23	7
make face	41	27	kick heel	39	31
make hay	17	9	see star	61	5
make hit	14	5	take heart	81	61
make mark	85	72			
make pile	25	8			
make scene	50	30			
pull leg	51	11			
pull plug	64	44			
pull punch	22	18			
pull weight	33	27			
Total:	814	481		391	268
Ratio:		0.5909			0.6854

Table 1. A breakdown of VNCs and idiomatic instances in the chosen train and test split.

IV. Experiments

Engaging Content
Engaging People

Models and evaluation

- GloVe (Pennington et al., 2014)
  - common crawl (2.2M tokens), cased
  - sentence embedding = average of word embeddings
  - 300-dimensional sentence embedding
  - off-the-shelf
- BERT (Devlin et al., 2018)
  - pytorch-pretrained-bert; bert-base-uncased
  - sentence embedding = average of final layer word embeddings
  - 768-dimensional sentence embedding
  - no fine-tuning

- probe model: Multi-Layered Perceptron (MLP)
- problem: binary classification
- evaluation metric: AUC_ROC score (0.5 = model does not discriminate)

Results

GloVe				
Model	IU_F		IU_R	
	auc	±CI	auc	±CI
rand. pred.	.4994	.0015	.4998	.0013
rand. vec.	.4997	.0015	.5	.0013
vanilla	.7485	.0003	.7717	.0022
abl. N	.7445	.0006	.7687	.0021
abl. D	.5012	.0018	.4993	.0015
abl. D+N	.4991	.0018	.5005	.0015

BERT				
Model	IU_F		IU_R	
	auc	±CI	auc	±CI
rand. pred.	.4997	.0015	.4998	.0013
rand. vec.	.4997	.0015	.5013	.0013
vanilla	.8411	.0002	.8524	.0016
abl. N	.8413	.0003	.8532	.0016
abl. D	.4991	.0019	.4978	.0015
abl. D+N	.4999	.0018	.5004	.0015

Tables 2 and 3. Probing results on GloVe and BERT models and baselines, including both the setting where the VNC's in the hold-out test set are fixed (IU_F) and the setting where they are resampled each time (IU_R). Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs. Note that cells shaded light grey belong to the same distribution as random baselines, as there is no statistically significant difference between the different scores; cells shaded dark grey belong to the same distribution as the vanilla baseline; and cells that are not shaded contain a significantly different score than both the random and vanilla baselines, indicating that they belong to different distributions.

Findings

- both vanilla GloVe and BERT significantly outperform random baselines
  - both GloVe and BERT encode a non-zero amount of idiomatic usage information
- vanilla BERT significantly outperforms vanilla GloVe
- IU_R outperforming IU_F indicates that predicting on IU_F is more challenging
  - the model is forced to rely on VNC-independent features to make predictions,
- no conclusive indication that the norm encodes idiomaticity information on this task
  - surprising – contextual incongruity?

Post hoc analysis: correlation

Task	Vectors	GloVe		BERT	
		L1	L2	L1	L2
IU	vanilla	-0.2231	-0.1786	-0.1490	-0.1756
	abl. N	-0.0074	0.0276	-0.0397	-0.0167

Table 4. Pearson correlation coefficients between class labels and L1 and L2 norms for vanilla vectors and vectors with ablated norms. For this analysis the Idiomatic label was mapped to 1 and the Literal label to 0.

- vanilla GloVe and BERT both norms have a weak negative correlation with IU labels
- correlation drops to ≈0 when ablating norm information, indicating information loss
  - does not align with previous experimental results ?
- negative correlation means that sentences containing idiomatic usage are positioned closer to the origin relative to sentences that contain literal usage
  - both GloVe and BERT vectors containing idiomatic usage are slightly shorter

Post hoc analysis: dimension deletion

Engaging Content
Engaging People

Expectations:

- deleting half the vector's dimensions should cause a performance drop
- this should happen regardless of which half of the vector is deleted

GloVe				
Model	**IU_F**		**IU_R**	
	auc	±CI	auc	±CI
rand. pred.	.4994	.0015	.4998	.0013
rand. vec.	.4997	.0015	.5	.0013
vanilla	.7485	.0003	.7717	.0022
del. 1h	**.7737**	.0005	.7553	.0023
del. 2h	.7043	.0005	.7545	.002

BERT				
Model	**IU_F**		**IU_R**	
	auc	±CI	auc	±CI
rand. pred.	.4997	.0015	.4998	.0013
rand. vec.	.4997	.0015	.5013	.0013
vanilla	.8411	.0002	.8524	.0016
del. 1h	**.8668**	.0002	**.8576**	.0016
del. 2h	.8137	.0003	.8368	.0016

Tables 5 and 6. Probing results on GloVe and BERT dimension deletion experiments, including both the setting where the VNC's in the hold-out test set are fixed (IU_F) and the setting where they are resampled each time (IU_R). Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs. Note that cells shaded light grey belong to the same distribution as random baselines, as there is no statistically significant difference between the different scores; cells shaded dark grey belong to the same distribution as the vanilla baseline; and cells that are not shaded contain a significantly different score than both the random and vanilla baselines, indicating that they belong to different distributions.

V. Limitations and Conclusion

Engaging Content
Engaging People

Limitations

- VNC-tokens dataset is not ideally suited for a probing scenario:
  - it is small (two orders of magnitude smaller than established datasets (Conneau et al., 2018))
  - limited in scope, focusing only on verb-noun compounds
  - a relatively older benchmark
  - imbalanced in terms of idiomatic/literal usage
  - does not control for sentence length, contains niche literary language and the occasional typo
- To do:
  - align dataset with PARSEME annotation guidelines
  - update it with additional example sentences

Conclusion

- both GloVe and BERT encode some idiomatic information to varying degrees
  - BERT encodes more
- both GloVe and BERT store idiomatic information in the second half of their vectors
  - the first half is even detrimental to the vector's overall idiomaticity encoding
- experiments yield inconclusive evidence as to whether idiomaticity is encoded in the vector norm: still an open question
- we also identify some limitations of the used dataset and highlight important directions for future work in improving its suitability for a probing analysis

Thank you for your attention!

e-mail: filip.klubicka@adaptcentre.ie
twitter: @lemoncloak
github: github.com/GreenParachute

School of Computer Science
Technological University Dublin
www.adaptcentre.ie