清華大學 Tsinghua University

Paper Sharing

Zhu Liu

2023.12.12

Two papers

- Qing Lyu, Marianna Apidianaki, and Chris Callison-burch. 2023. Representation of Lexical Stylistic Features in Language Models' Embedding Space. In Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), pages 370–387, Toronto, Canada. Association for Computational Linguistics.

- Petersen E, Potts C. Lexical Semantics with Large Language Models: A Case Study of English "break"[C]//Findings of the Association for Computational Linguistics: EACL 2023. 2023: 490-511.

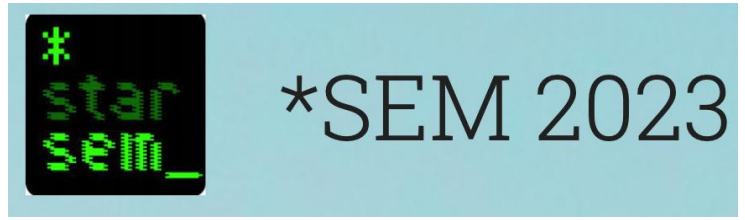
Representation of Lexical Stylistic Features in Language Models' Embedding Space

Qing Lyu **Marianna Apidianaki** **Chris Callison-Burch**
University of Pennsylvania
{lyuqing, marapi, ccb}@seas.upenn.edu

StarSEM 2023 (on ACL 2023)

https://github.com/veronica320/Lexical-Stylistic-Features/tree/main

*SEM



- The 12th Joint Conference on **Lexical and Computational** Semantics (*SEM 2023)
- 29 Long, 16 short, 8 findings

- How Are **Idioms** Processed Inside Transformer Language Models?
- **Syntax and Semantics** Meet in the "Middle": Probing the Syntax-Semantics Interface of LMs Through Agentivity
- Estimating **Semantic Similarity** between In-Domain and Out-of-Domain Samples
- **Representation** of Lexical Stylistic Features in Language Models' **Embedding** Space
- A Tale of Two Laws of **Semantic Change**: Predicting Synonym Changes with Distributional Semantic Models
- „Mann" is to "Donna" as「国王」is to « Reine » Adapting the **Analogy** Task for **Multilingual and Contextual** Embeddings
- Limits for learning with language models
- Adverbs, Surprisingly
- **Figurative** Language Processing: A Linguistically Informed Feature Analysis of the Behavior of Language Models and Humans

Authorship

- Chris Callison-Burch is an associate professor of Computer and Information Science at the University of Pennsylvania.
- His PhD students joke that now whenever they ask him anything his first response is “Have you tried GPT for that?”
- More than 100 publications, which have been cited over 25,000 times. Sloan Research Fellow, and he has received faculty research awards from Google, Microsoft, Amazon, Facebook, etc.

Authorship

Hi! My name is Veronica Qing Lyu (吕晴). I am a fifth-year PhD student in Computer and Information Science at the University of Pennsylvania, advised by Chris Callison-Burch and Marianna Apidianaki.

In the beginning of my PhD studies, I worked on information extraction and script learning. My current research interests lie in the intersection of linguistics and natural language processing, especially probing Language Models (LMs) for interpretability. Specifically, I hope to answer three questions:

- (i) *What* knowledge do LMs encode?
- (ii) *Why* do LMs make certain predictions?
- (iii) *How* can we make interpretability insights from (i) and (ii) actionable?

Before Penn, I was an undergrad in linguistics at the Department of Foreign Languages and Literatures at Tsinghua University.

Background

- Text style by lexical choices

  complexity e.g., *help* vs. *assist*

  formality, e.g., *dad* vs. *father*

  figurativeness, e.g., *heavy* vs. *burdened*

- How to detect these styles in a lexical level?

  Previous: statistics-based features (word length, frequency, affixes…)

  Current: **LMs** have encoded rich lexical semantics (similarity, polysemy…)

Contributions

- To probe whether representation space encodes these properties.
- Various settings with several experimental conclusions

  Settings: static vs. contextual; which layers; model selections; text length

  Conclusions:

  (1) overall expectations

  (2) short texts prefer static vectors, while long prefer contextual ones.

  (3) Anisotropy corrections for a better probation

Related Work

- Probing linguistic and world knowledge encoded in LM representations.

1) Diagnostic classifiers on well-designed tasks

Characteristics: intuitive, understandable but extrinsic (prone to tasks, new data and learning)

Classic Works: BERTology [1]; LLM prompt queries

2) Geometrics-based

Characteristics: robust, intrinsic but limited capability, vague objectives

Limited Operations: algebra; transformations; distance; decomposition…

Works: Analogy task in Word2Vec [2] (king-man+women=queen); Adjective “intensity” [3]; other human-aligned properties (honesty…).

3) Others: information-theory; behavior tests; visualization and so on.

Definitions

- Complexity

”talk to children or non-native English speakers” vs. “used by academics or domain experts”

- Formality

“the way one talks to a superior” vs. “used with friends”

- Figurativeness

Literal vs. non-literal

- Contextual or Lexical?

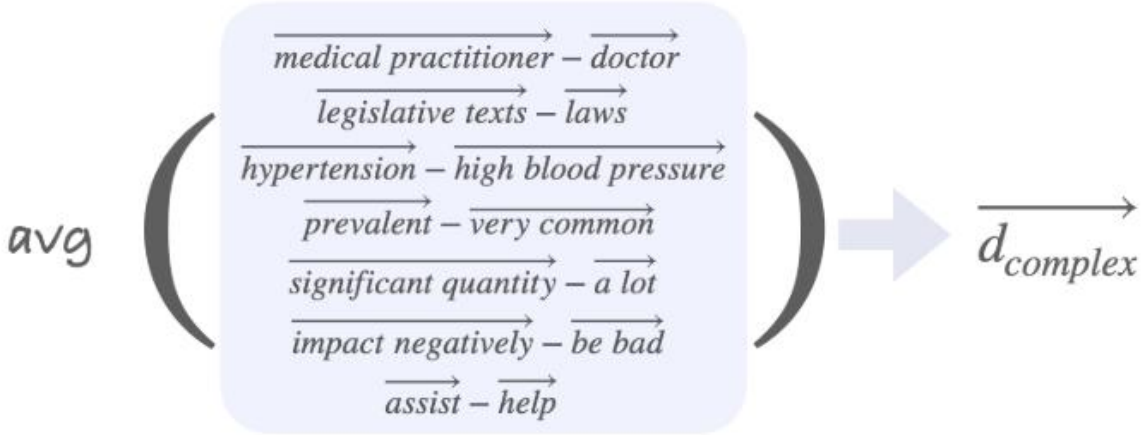
COMPLEXITY	doctor → medical practitioner
	laws → legislative texts
	high blood pressure → hypertension
	very common → prevalent
	a lot → significant quantity
	be bad → impact negatively
	help → assist
FORMALITY	my gosh → jesus
	breathing → respiratory
	yeah → yes
	ten years → decade
	first of all → foremost
	a whole bunch → full
	my dad → father
FIGURATIVENESS	bright → radiant
	heavy → burdened
	unsympathetic → cold-hearted
	fall → plummet
	a lot of → a sea of
	quick → lightning
	hard → ironclad

Method

- This method is widely used in geometric probing experiments [][]
- Feature vector generation (static “training”)

A small set of seed pairs in terms of certain r$\overrightarrow{dVec}$on.

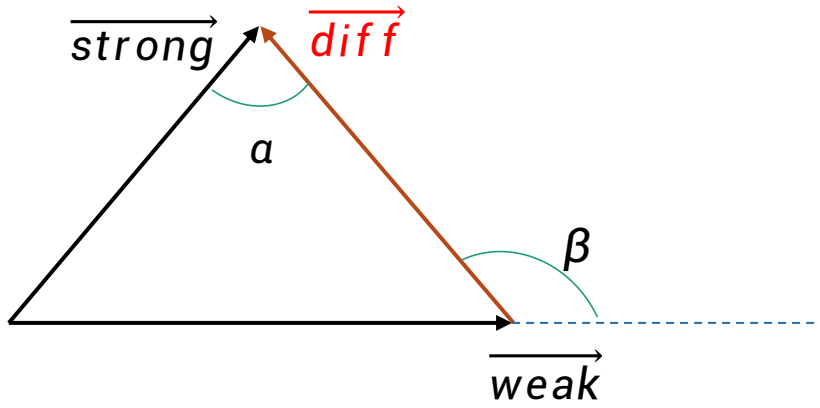
Then, to calcula[illegible]



Method

- Feature vector generation

Cos similarity with difference vecto$\overrightarrow{d_{complex}}$, $\overrightarrow{d_{formal}}$ and $\overrightarrow{d_{fig}}$



Experimental Setup

- Evaluation task and metrics

  binary classification task: a pair of texts that are semantically similar but stylistically different.

“You must adhere to the rules.” (t0) and “You must obey the rules.” (t1)

t0 is more figurative (and formal?)

Metric: Accuracy (with balanced labels)

Experimental Setup

- Seed Pair

  only 7 seed pairs for each notion

Less seed, more powerful. With some work only just 1 seed

COMPLEXITY	doctor → medical practitioner
	laws → legislative texts
	high blood pressure → hypertension
	very common → prevalent
	a lot → significant quantity
	be bad → impact negatively
	help → assist
FORMALITY	my gosh → jesus
	breathing → respiratory
	yeah → yes
	ten years → decade
	first of all → foremost
	a whole bunch → full
	my dad → father
FIGURATIVENESS	bright → radiant
	heavy → burdened
	unsympathetic → cold-hearted
	fall → plummet
	a lot of → a sea of
	quick → lightning
	hard → ironclad

Experimental Setup

Feature	Short-text (word/phrase)	Long-text (sentence)
Complexity	SimplePPDB	SimpleWikipedia
Formality	StylePPDB	GYAFC
Figurativeness	–	IMPLI

- Datasets (short and long)
- Baselines: Majority and Frequency
- Configuration

  Language Models (Static, like Glove, fastText; Contextual, like BERT and RoBERTa)

  Different layers

  Pooling Strategies: Mean vs. Max for more than one token (phrase and sentences)
- Validation for the best configuration and test.

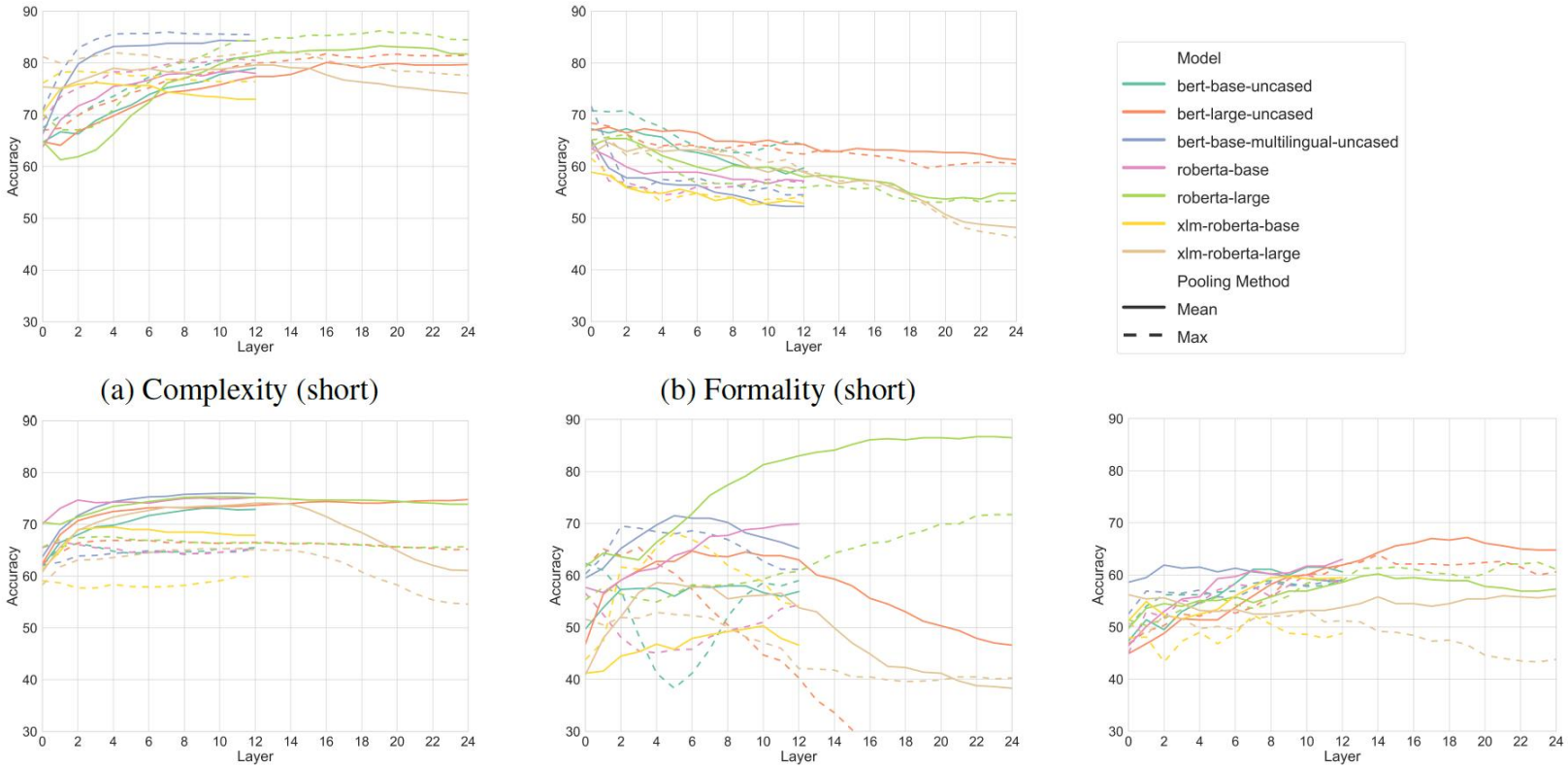
Results and Discussion

Pooling	Model	Complexity		Formality		Figurativeness
		short	long	short	long	long
	majority	55.1	50.6	51.2	51.8	51.4
Mean	frequency	83.2	51.0	61.0	41.4	49.7
	static	84.8 glove	60.0 glove	**76.8** glove	82.8 glove	54.3 glove
	contextualized (single layer)	**86.2** roberta-large (4)	**76.5** mbert-base (1)	68.7 bert-base (1)	82.4 roberta-large (12)	**72.9** bert-large (14)
	contextualized (layer agg)	84.4 mbert-base (10)	76.0 mbert-base (11)	67.6 bert-large (1)	**86.7** roberta-large (23)	67.2 bert-large (19)
Max	frequency	80.7	46.4	57.2	42.5	47.9
	static	**89.4** glove	58.0 glove	**76.0** glove	63.4 glove	56.0 fasttext
	contextualized (single layer)	87.7 roberta-large (4)	**69.4** roberta-base (12)	71.7 mbert-base (0)	**73.6** mbert-base (1)	**64.8** bert-large (11)
	contextualized (layeragg)	86.2 roberta-large (19)	67.6 roberta-large (4)	71.7 mbert-base (0)	71.7 roberta-large (24)	63.9 bert-large (14)

- LMs beat baseline (majority and frequency)
- Mean is better than Max, generally.
- Contextual representation is better than static one, except for "short".

Effects of laye

- Monotonous only for complexity and formality short. (but different trends)
- agg is better than single for complexity and formality, but worse for figurativeness.



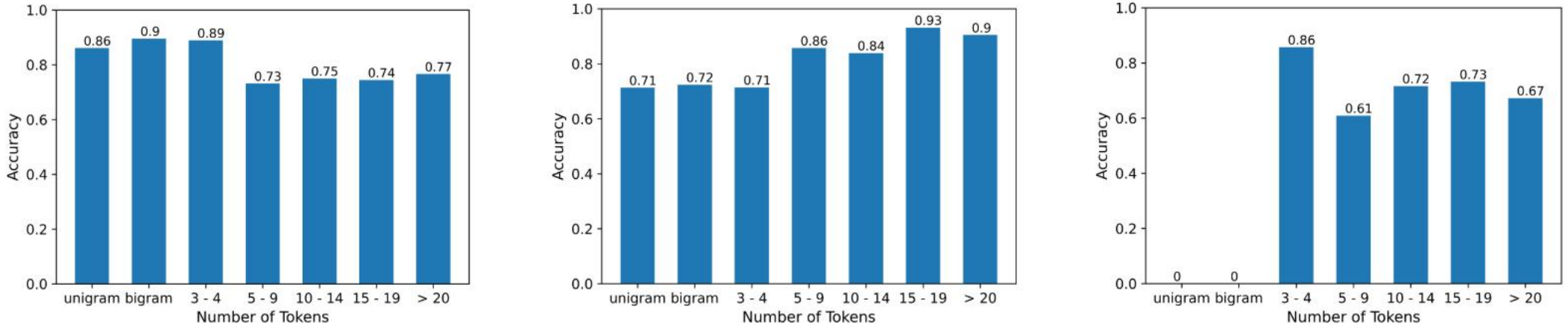
(a) Complexity (short) (b) Formality (short)

(c) Complexity (long) (d) Formality (long) (e) Figurativeness (long)

Figure 3: Performance change across layers of different LMs (under the layer aggregation setting).

Pooling	Stats	Complexity		Formality		Figurativeness
		short	long	short	long	long
Mean	2 beats 1 (%)	63.0	78.0	92.9	72.4	54.3
	acc gain	2.6	4.1	4.3	5.3	0.1
Max	2 beats 1 (%)	66.1	72.4	95.3	64.6	44.9
	acc gain	3.0	3.0	4.4	3.2	-0.5
Average	2 beats 1 (%)	64.6	75.2	94.1	68.5	49.6
	acc gain	2.8	3.5	4.3	4.3	-0.2

Effect of Text Length

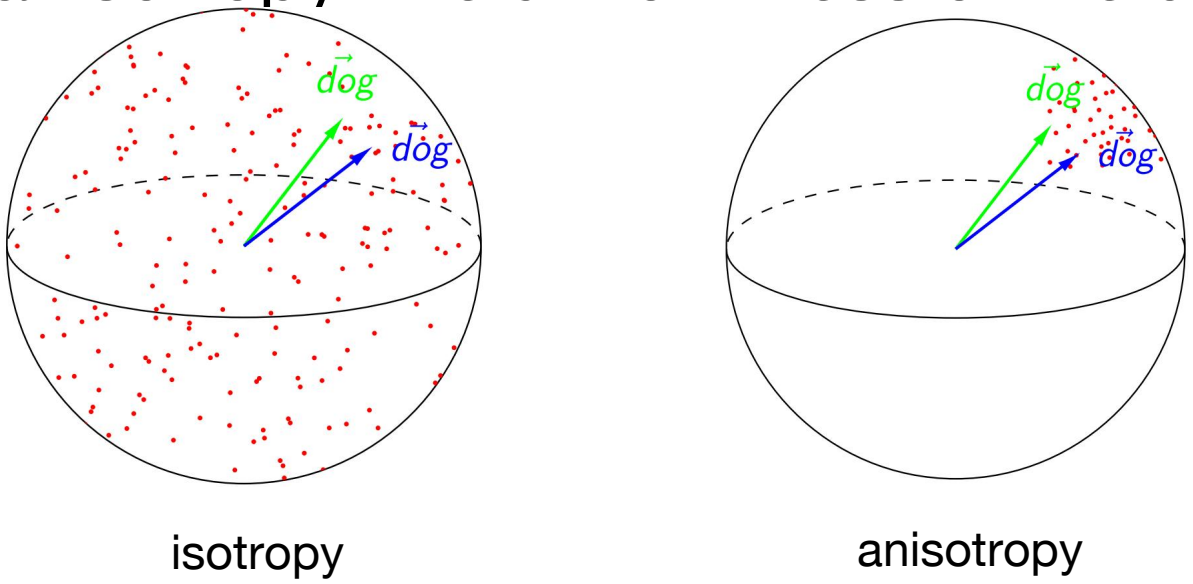


(a) Complexity (b) Formality (c) Figurativeness

Figure 4: Optimal performance over different bins of text length (under the layer aggregation setting).

Anisotropy Reduction Experiments

- Previous study [4] shows that LMs' repres are anisotropy (各向异性)
- Isotropy vs. anisotropy: the uniformness of the distribution



isotropy anisotropy

Anisotropy Reduction Experiments

- Three reduction methods

(1) All-but-the-top (abtt) [5

$$\mu = \frac{1}{|\mathcal{S}|} \cdot \sum_{x \in \mathcal{S}} x$$

$$u_1, ..., u_d = \mathrm{PCA}(\{x - \mu, x \in \mathcal{S}\}).$$

$$x_{abtt} = x - \mu - \sum_{i=1}^{k} \left(u_i^\top x \right) u_i.$$

(2) Standardization

$$\sigma = \sqrt{\frac{1}{|\mathcal{S}|} \cdot \sum_{x \in \mathcal{S}} (x - \mu)^2}$$

$$x_{standard} = \frac{x - \mu}{\sigma}$$

(3) Rank-based (similarities -> correlation)

$$x_{rank} = rank(x).$$

Anisotropy Reduction Experiments

Pooling	Model	Complexity		Formality		Figurativeness
		short	long	short	long	long
Mean	static	84.8	60.0	**76.8**	82.8	54.3
	contextualized (singlelayer)	86.2	**76.5**	68.7	82.4	**72.9**
	contextualized (singlelayer+**abtt**)	80.3	69.3	76.6	76.7	70.9
	contextualized (singlelayer+**standardization**)	**90.4**	73.9	74.1	80.6	68.3
	contextualized (singlelayer+**rank**)	85.6	76.0	70.8	81.7	71.8
	contextualized (layeragg)	84.4	76.0	67.6	**86.7**	67.2
	contextualized (layeragg+**abtt**)	81.7	68.5	76.6	63.0	72.6
	contextualized (layeragg+**standardization**)	**90.4**	73.6	75.2	79.9	67.6
	contextualized (layeragg+**rank**)	83.7	75.7	68.1	82.1	67.0
Max	static	89.4	58.0	76.0	63.4	56.0
	contextualized (singlelayer)	87.7	69.4	71.7	73.6	64.8
	contextualized (singlelayer+**abtt**)	80.6	64.9	78.2	80.8	66.7
	contextualized (singlelayer+**standardization**)	**90.5**	63.8	**80.9**	**81.7**	60.4
	contextualized (singlelayer+**rank**)	87.1	**69.6**	70.3	76.0	66.5
	contextualized (layeragg)	86.2	67.6	71.7	71.7	63.9
	contextualized (layeragg+**abtt**)	81.9	63.9	78.2	72.5	**71.1**
	contextualized (layeragg+**standardization**)	**90.5**	63.7	**80.9**	80.6	61.9
	contextualized (layeragg+**rank**)	86.1	69.3	71.7	71.5	67.4

Reference

Conclusion

- The embedding space of pretrained LMs encodes abstract stylistic notions such as formality, complexity, and figurativeness
- static embeddings are better at capturing the style of short texts (words and phrases) whereas contextual embeddings at longer texts (sentences).
- correcting the anisotropy of contextualized LMs' representation space could close the performance gap from static embeddings on short texts

Lexical Semantics with Large Language Models: A Case Study of English *break**

Erika Petersen
Stanford University
epetsen@stanford.edu

Christopher Potts
Stanford University
cgpotts@stanford.edu

Findings of EACL 2023

https://github.com/epetsen/break-llms

Authorship

- Christopher Potts: Professor and Chair, Department of Linguistics, and Professor, Department of Computer Science

https://crfm.stanford.edu/assets/report.pdf#philosophy

Talk: Lexical semantics in the time of large language models

https://www.youtube.com/watch?v=EbwtZtd8XRo

Background

- Linguistics and neural network research have common ground and common cause. The authors argue that LLMs are powerful devices for studying lexical semantics.
- English verb break has long been central to lexical semantics because it has a range of senses that related to its argument structure.
- Computational models and methods:

1) Lexical/Sense representation
2) Similarity estimation; meaning modulation; polysemy detection; WSD
3) Clustering/Decomposition/Visualization
4) Lexical Semantic Probing

Three tenets of lexical semantics

	Linguistics	Static vectors	LLMs
High dimensionality: Lexical semantic entries consist of many features.	Yes	Yes	Yes
Contextual modulation: A word sense will be influenced by its immediate morphosyntactic context as well as the broader context of use.	Yes	No	Yes
Discreteness: The features in lexical semantic entries are discrete and highly structured.	Yes	No	No

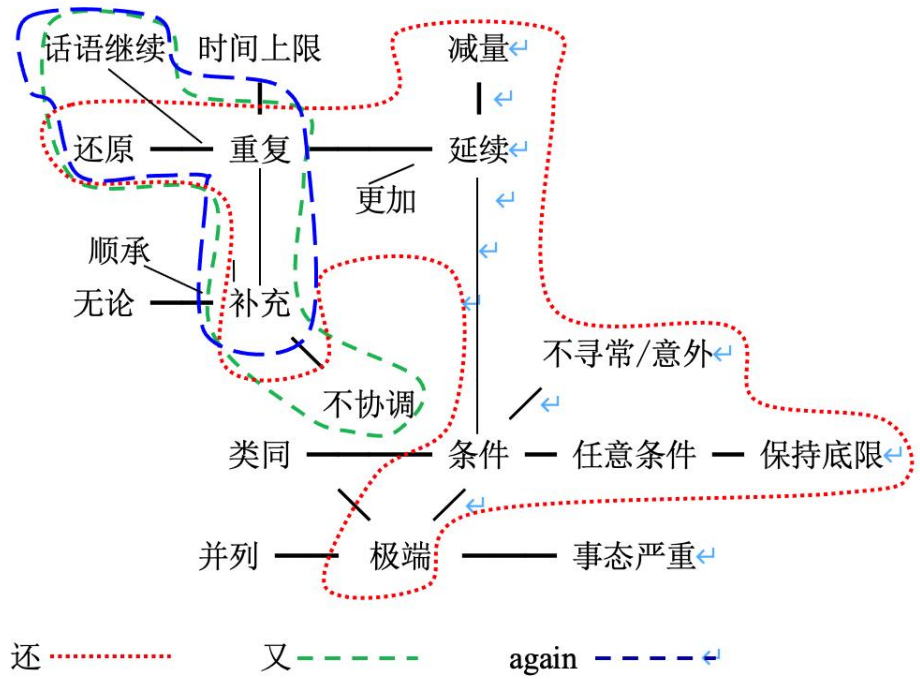
Table 1: Core tenets. Our focus is in particular on the relationship between 'Linguistics' and 'LLMs' in this table.

- Symbolic grammar (early NLP) naturally meets all the three tenets.
- Discreteness was discarded when distributional models become central.

举例

语言	词	类同	补充	重复	延续	更加	增加	减量	还原	条件	任意	极端	严重	无论	让步	上限	顺承	不
英语	also	类	补															
	too	类	补															
	again		补	重					还									
	still				延	更		减		条	任							

补充义副词语义地图：



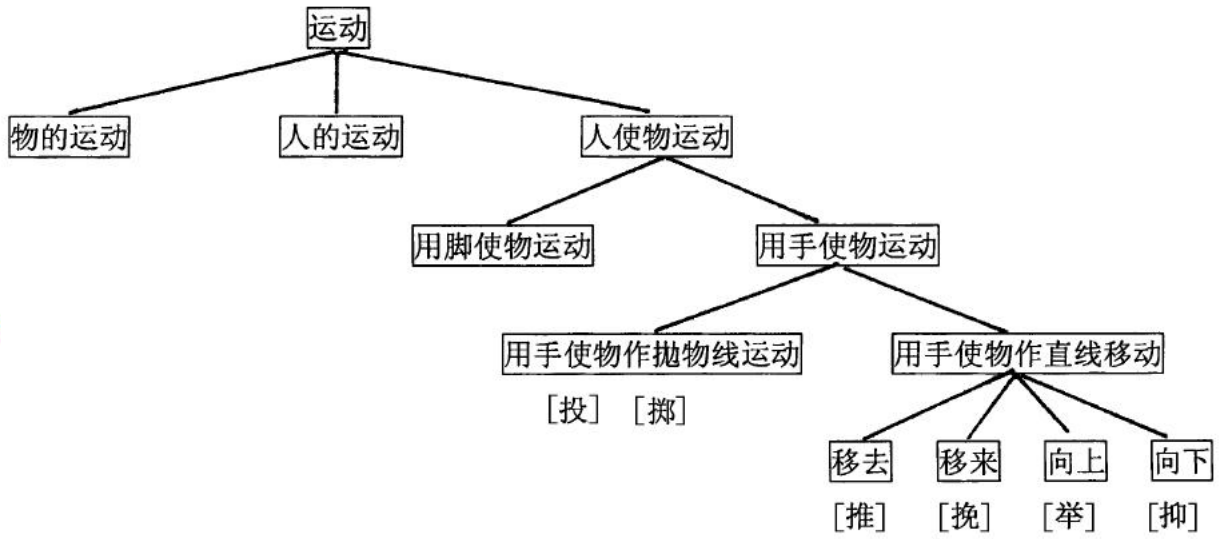
knife *n.* **1.** a cutting tool composed of a blade with a sharp point and a handle. **2.** an instrument with a handle and blade with a sharp point used as a weapon.

Fig. 1. An example of an enumerative entry for noun *knife*.

$$
\begin{bmatrix}
\textbf{knife} \\
\text{TYPESTR} = \left[\text{ARG1} = \boxed{\text{x}}\ \textbf{artifact tool}\right] \\
\text{ARGSTR} = \begin{bmatrix}
\text{D-ARG1} = \boxed{\text{y}}\ \textbf{physical object} \\
\text{D-ARG2} = \boxed{\text{w}}\ \textbf{human} \\
\text{D-ARG3} = \boxed{\text{z}}\ \textbf{human} \\
\text{D-E1} = \boxed{e_1}\ \textbf{transition} \\
\text{D-E2} = \boxed{e_2}\ \textbf{process}
\end{bmatrix} \\
\text{QUALIA} = \begin{bmatrix}
\text{FORMAL} = \boxed{\text{x}} \\
\text{CONSTITUTIVE} = \{\textbf{blade}, \textbf{handle}, \ldots\} \\
\text{TELIC} = \textbf{cut act}(\boxed{e_2}, \boxed{\text{w}}, \boxed{\text{x}}, \boxed{\text{y}}) \\
\text{AGENTIVE} = \textbf{make act}(\boxed{e_1}, \boxed{\text{z}}, \boxed{\text{x}})
\end{bmatrix}
\end{bmatrix}
$$

Fig. 2. An example of a generative entry for noun *knife*.

表 1 “运动”概念域的层级结构



Contribution

- A case study of “break”.
- They argue that “the facts surrounding break should lead linguists to reconsider their commitment to discreteness and embrace a more fluid, usage-based foundation for semantic theory.”

“Break”

- A change-of-state verb that undergoes the causative alternation

(1) The linguist broke the window

(2) The window broke.

alternative change-of-state verbs are referred to *break*-verbs.

unaccusativity hypothesis -> theme-like argument as a subject

“Break”

- Break can take a wide range of senses. (think of wordnet 50+)

	Frame	Sense
1.	break the vase	shatter
2.	break the computer	render inoperable
3.	break the news	reveal
4.	break the silence	interrupt
5.	break the record	surpass
6.	break the code	decipher
7.	break the law	violate
8.	break the habit	end
9.	break the horse	tame
10.	break a $10 bill	make change
11.	break the fall	lessen
12.	the weather broke	changed
13.	the day broke	began

(a) Uses without particles/predicates.

	Frame	Sense
14.	break off the engagement	end
15.	break out	begin
16.	break out of jail	escape
17.	break out in hives	get
18.	break into the building	intrude
19.	break down the problem	analyze
20.	break down the proteins	decompose
21.	break in	enter
22.	break in	interrupt
23.	break free	escape
24.	break even	profit = loss
25.	break forth	emerge
26.	break to the right	turn

(b) Uses with particles/predicates.

Table 2: Senses for *break*. A comprehensive account of senses may not be possible (Section 5.3).

“Break”

- Sense distinctions interact with the causative alternative (and semantic roles).

causative alternative

obligatorily transitive

obligatorily intransitive

	Frame	Sense
1.	break the vase	shatter
2.	break the computer	render inoperable
3.	break the news	reveal
4.	break the silence	interrupt
5.	break the record	surpass
6.	break the code	decipher
7.	break the law	violate
8.	break the habit	end
9.	break the horse	tame
10.	break a $10 bill	make change
11.	break the fall	lessen
12.	the weather broke	changed
13.	the day broke	began

(a) Uses without particles/predicates.

	Frame	Sense
✓ 14.	break off the engagement	end
15.	break out	begin
16.	break out of jail	escape
17.	break out in hives	get
18.	break into the building	intrude
✓ 19.	break down the problem	analyze
✓ 20.	break down the proteins	decompose
21.	break in	enter
22.	break in	interrupt
23.	break free	escape
24.	break even	profit = loss
25.	break forth	emerge
26.	break to the right	turn

(b) Uses with particles/predicates.

Accusative but intransitive (differ from the left)

Table 2: Senses for *break*. A comprehensive account of senses may not be possible (Section 5.3).

Question for lexical semantic theory

- Whether there is a single unifying semantic frame underlying this diverse array of senses.
- “[t]he various meanings of BREAK [. . .] can all be subsumed under a ‘deep’ meaning, ‘(cause) not to continue in existing state’, which links even the most disparate meanings of BREAK’ []
- Another: a few more primitive semantic dimensions that give rise to a combinatorial space of predicted senses.

Feature-based theories (linguistics)

	Abstract Linguistic Features				Semantic features					
	Transitive	*Unaccusative*	*Agent*	*Metaphorical*	*separate*	*violate*	*end*	*appear*	*out_escape*	*out_begin*
1. We broke the vase	1	0	1	0	1	0	0	0	0	0
2. The vase broke	0	1	0	0	1	0	0	0	0	0
3. We broke the law	1	0	1	1	0	1	0	0	0	0
4. The silence broke a procedural rule	1	0	0	1	0	1	0	0	0	0
5. We broke the silence	1	0	1	1	0	0	1	0	0	0
6. The day broke	0	1	0	1	0	0	0	1	0	0
7. The storm broke	0	1	0	1	0	0	0	1	0	0
8. Sweat broke on his forehead	0	1	0	1	1	0	0	1	0	0
9. We broke out (of jail)	0	0	1	0	0	0	0	0	1	0
10. Fighting broke out	0	1	0	1	0	0	0	0	0	1

Table 3: Partial feature-based analysis of *break* in different syntactic contexts.

Static Vector Modeling

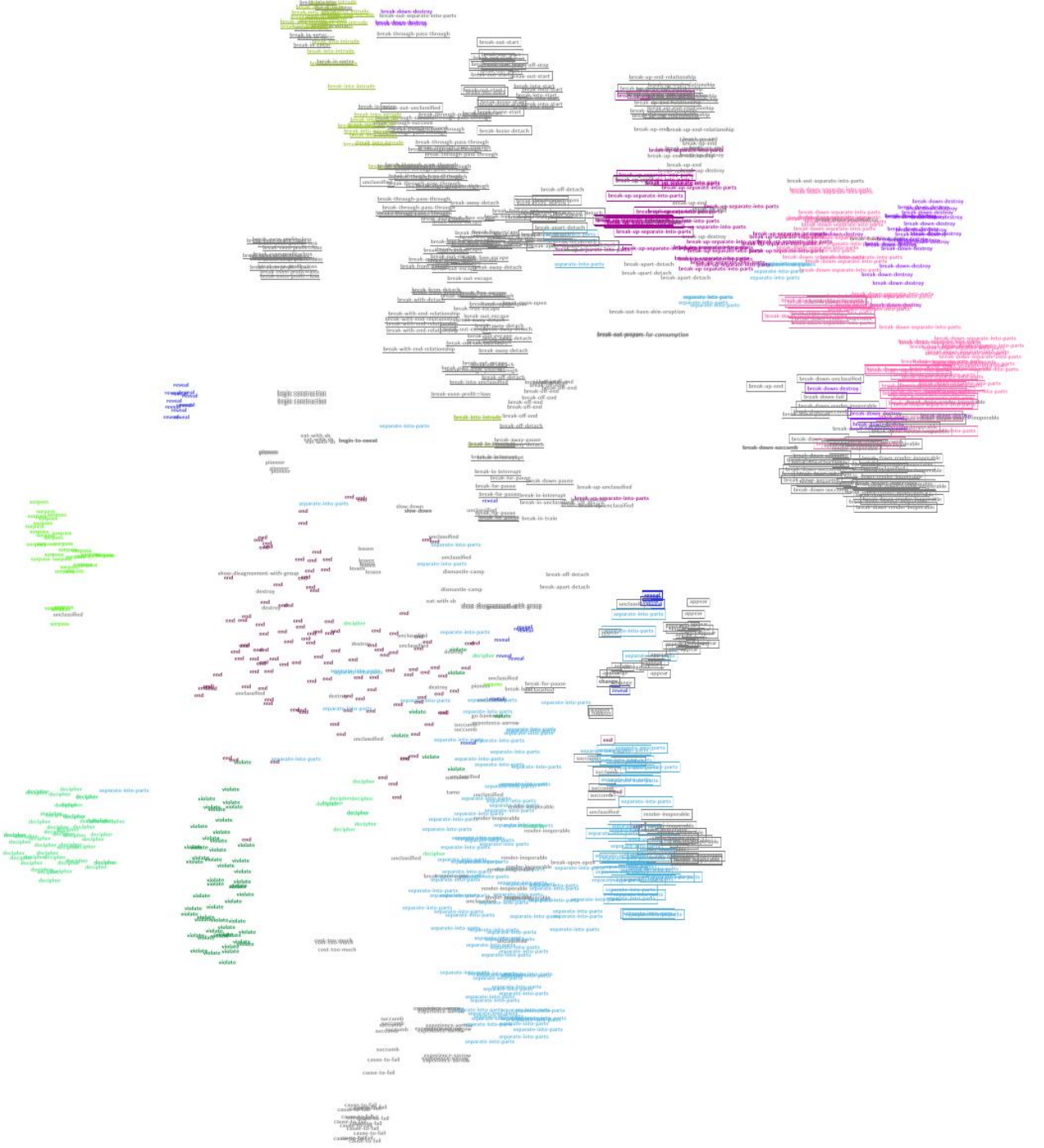
- From sparse features to dense dimensions.
- Models: word2vec, Glove, fastText
- Co-occurrence; PMI (high weights to word occurrence pair by rejecting the null hypothesis that they are independent.)
- No discrete, interpretable features as linguistic theory shows.
- Due to the lack of context modulation, they are incapable to describe the same word type “break”. (more like a weighted average of these senses.)

LLM investigations

- Transformers (objective of MLM); RoBERTa-large
- LLM representations are continuous and highly abstract.
- Probing: supervised classifiers
- Annotated datasets: One M.S. thesis [Peterson, 2020] and more examples (1042 sentences from CoCA) with construction labels (‘unergative’, ‘un- accusative’, ‘causative’) and 72 meaning class.
- Metrics: F1 scores; selective scores (minus a random control task)
- The last layer of RoBERTa.

Results – visualization (app.)

- Higher layer has a more robust result
- Good meaning clustering
- T-sne has information loss and is not stable enough.



Result

Layer	Probe	Control	Selectivity
1	0.33	0.03	0.30
6	0.81	0.03	0.79
12	0.83	0.03	0.80
18	0.80	0.03	0.76
24	0.86	0.03	0.83

(a) Meaning-class probing results.

Layer	Probe	Control	Selectivity
1	0.50	0.33	0.17
6	0.94	0.34	0.60
12	0.96	0.33	0.63
18	0.96	0.35	0.61
24	0.97	0.32	0.65

(b) Construction-type probing results.

Table 5: RoBERTa-large probing results. We report Macro F1 and Selectivity, which is the Macro F1 score for the task minus the Macro F1 for a control task (random assignment of tokens to classes). Results for other models are similar; see Appendix D.

- The construction type probes are nearly perfect
- Error examples could inform lexical semantic theory given its performance

Examples

- E1: “spontaneous”
- E2-E9 Meaning mix, have different readings.
- More detailed analysis in A.F.

		Meaning		Construction	
	Sentence	Gold	Predicted	Gold	Predicted
1.	Patients will sometimes break out in a spontaneous recitation of the rosary	break_out_ start	break_out_ start	unacc.	unerg.
2.	It was like you knew something, like you knew the story was getting ready to break again.	reveal	appear	unacc.	unacc.
3.	@(Soundbite-of-music)@!Mr-GELB: (Singing) Tell me who’s going to pick up the pieces when you start to break down.	break_down_ separate_into_parts	break_down_ succumb	unacc.	unacc.
4.	People have so many problems overcoming the disputes that occur when families break up	break_up_ end_relationship	break_up_ separate_into_parts	unacc.	unacc.
5.	“So why tell the whole story now? Somebody, some male, has got to be willing to break this code of silence,” he says.	violate	end	unacc.	unacc.
6.	So they forwarded the pictures to Madrid, where another officer noticed some printing on a towel that helped break the case.	decipher	end	causative	causative
7.	Then too, stress can also work to break down the immune system, increasing the likelihood of respiratory and creating gastrointestinal and nervous disorders.	break_down_ render_inoperable	break_down_ destroy	causative	causative
8.	Wind, naturally acidic rain, and physical processes such as freezethaw cycles also break down rock.	break_down_ separate_into_parts	break_down_ destroy	causative	causative
9.	It didn’t take being an ICU exec to break the code: trade secret.	decipher	violate	causative	causative

Discussion

- A striking alignment: context + high dimension
- Discreteness is good but really consuming and demanding.
- LLM suggest a theory that is actually more about tokens (instances of use) than types. (usage-based)
- (For me) Can we decompose the “dense myth” in current LMs?

Q & A

THANK YOU

Note