Paper Sharing

2024.03.28

Paper Overview

- Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. **Probing Pretrained Language Models for Lexical Semantics.** In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7222–7240, Online. Association for Computational Linguistics.
- Boleda G. **Distributional semantics and linguistic theory**[J]. Annual Review of Linguistics, 2020, 6: 213-234.

Probing Pretrained Language Models for Lexical Semantics

Ivan Vulić[♠] **Edoardo M. Ponti**[♠] **Robert Litschko**[♢] **Goran Glavaš**[♢] **Anna Korhonen**[♠]
[♠]Language Technology Lab, University of Cambridge, UK
[♢]Data and Web Science Group, University of Mannheim, Germany
{iv250,ep490,alk23}@cam.ac.uk
{goran,litschko}@informatik.uni-mannheim.de

EMNLP'20 Cited by 201

LTL Overview

- Language Technology Lab: https://ltl.mmll.cam.ac.uk/
- Major areas: NLP fundamentals; NLP for health/multilingualism/education/digital humanities/quantitative linguistics
- Selected Publications on Semantics:

Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. (ACL'19)

Unseen Word Representation by Aligning Heterogeneous Lexical Semantic Spaces. (AAAI'19)

Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. (EMNLP'19)

Acquiring Verb Classes Through Bottom-Up Semantic Verb Clustering. (LREC'18)

Background

- Pretrained language models offer unmatched results on NLP tasks
- SoTA PLMs (at that time): BERT, RoBERTa, T5
- Context-sensitive learnable embedding vs. static type-level embedding
- Why they work? or what do they learn?
- Probing: linguistic knowledge or structure (syntax or morphology)

Motivation

- The paper focuses on how and where lexical semantic knowledge is encoded in PLMs?
- [Note] Lexical Semantics -> type-level; content words; concept

  NOT token-level/Context-specific, like polysemy; metonymy etc.

  NOT functional/grammatical words, like pronouns, determiners, conjunctions (lexicalization vs. grammaticalization)

  NOT grammatical/functional/structural meaning

Motivation

- Previous work suggest that PLMs have rich lexical knowledge
- *But* a systematic empirical study across different languages is currently lacking.

Contributions

Systematic experiments

- Different extraction configurations
- Several models
- Six typologically diverse languages
- Five diverse lexical tasks

Contributions

Research Questions:

- Q1: language-specific or universal for lexical extraction strategies?
- Q2: Is lexical knowledge concentrated or scattered in NNs?
- Q3: how well does “BERT-based” static word embedding?
- Q4: Do monolingual LMs learn similar representations for words denoting similar concepts (i.e., translation pairs)?

Q1,Q4: consistency; Q2: distribution Q3: aggregation

Related Work

- Word embeddings: static (Word2Vec, FastText) vs. contextual PLMs (BERT-based)

  Differences from static word embeddings:

  (1) type-level vs. token-level (2) complete or subword strings
- Probing English-only tasks using learned classifiers on POS tags [1], word senses [2], or dependency structures [3].
- Multi-Simlex[4] considers multiple languages but only one lexical task.

Experimental Setup

- PLMs: monolingual BERT Base & multilingual BERT
  - 12 768-dimensional Transformer layers
  - from bottom (L1) to top (L12) plus embedding layer (L0)
  - 12 attention heads

  fastText (FT) vectors (on Wiki)
- Languages: English (EN), German (DE), Russian (RU), Finnish (FI), Chinese (ZH), and Turkish (TR)
- Corpus: 1M sentences from Europarl (EN, DE, FI), UNPS (RU, ZH), WMT17 (TR)

Extraction Configuration

Component	*Label*	*Short Description*
Source LM	MONO	Language-specific (i.e., monolingually pretrained) BERT
	MULTI	Multilingual BERT, pretrained on 104 languages (with shared subword vocabulary)
Context	ISO	Each vocabulary word w is encoded *in isolation*, without any external context
	AOC-M	*Average-over-context*: average over word's encodings from M different contexts/sentences
Subword Tokens	NOSPEC	Special tokens [CLS] and [SEP] are excluded from subword embedding averaging
	ALL	Both special tokens [CLS] and [SEP] are included into subword embedding averaging
	WITHCLS	[CLS] is included into subword embedding averaging; [SEP] is excluded
Layerwise Avg	AVG(L≤n)	Average representations over all Transformer layers up to the n-th layer L_n (included)
	L=n	Only the representation from the layer L_n is used

Table 1: Configuration components of word-level embedding extraction, resulting in 24 possible configurations.

Extraction Configuration

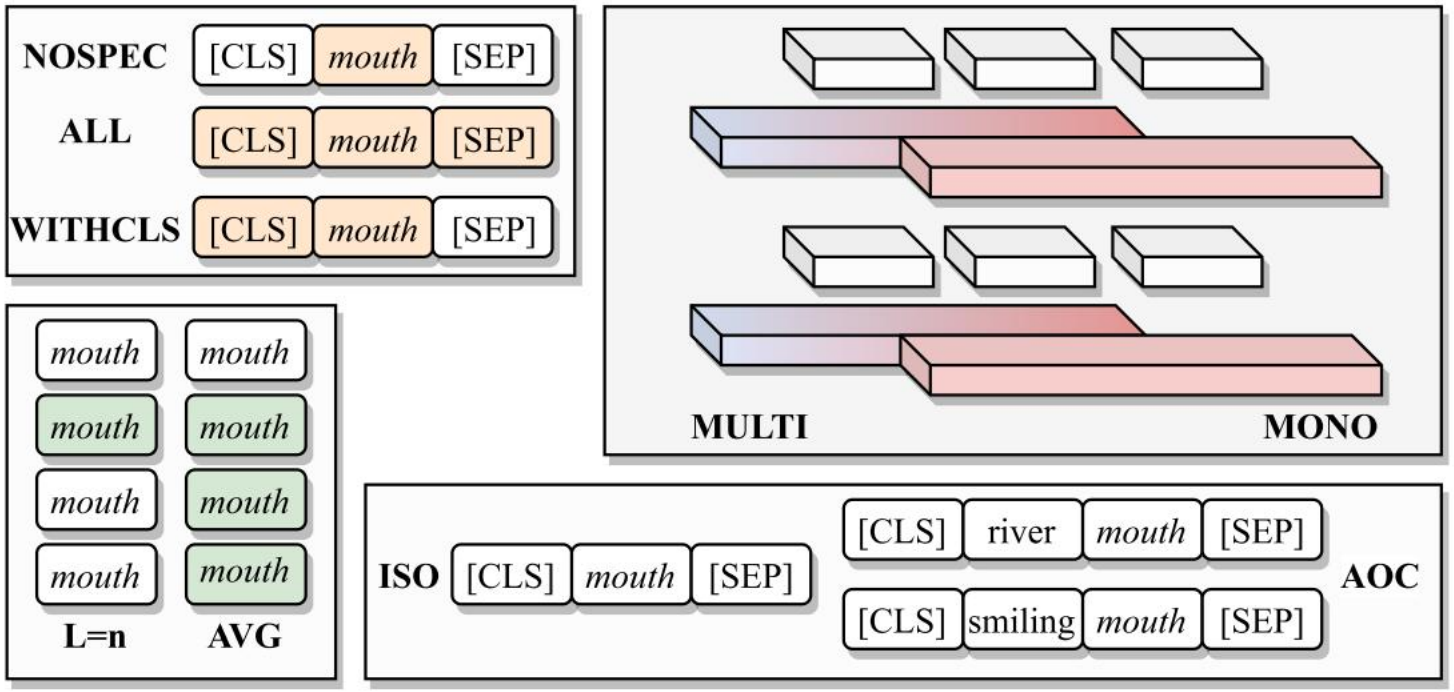


Figure 1: Illustration of the components denoting adopted extraction strategies, including source LM (top right), presence of context (bottom right), special tokens (top left), and layer-wise averaging (bottom left).

Tasks

- T1: Lexical Semantic Similarity (LSIM)
  Metrics: Correlation between similarity scores from human evaluation and word vectors for word pairs
  Dataset: Multi-SimLex; 1,888 pairs with 13 languages;
- T2: Word Analogy (WA)
  Dataset: Bigger Analogy Test Set (BATS) with 99,200
  Example: man:king=woman:?
  Metrics: Precision@1

$$\text{argmax}_d(\cos(c - a + b)) \ s.t. \ w_a : w_b = w_c : x$$

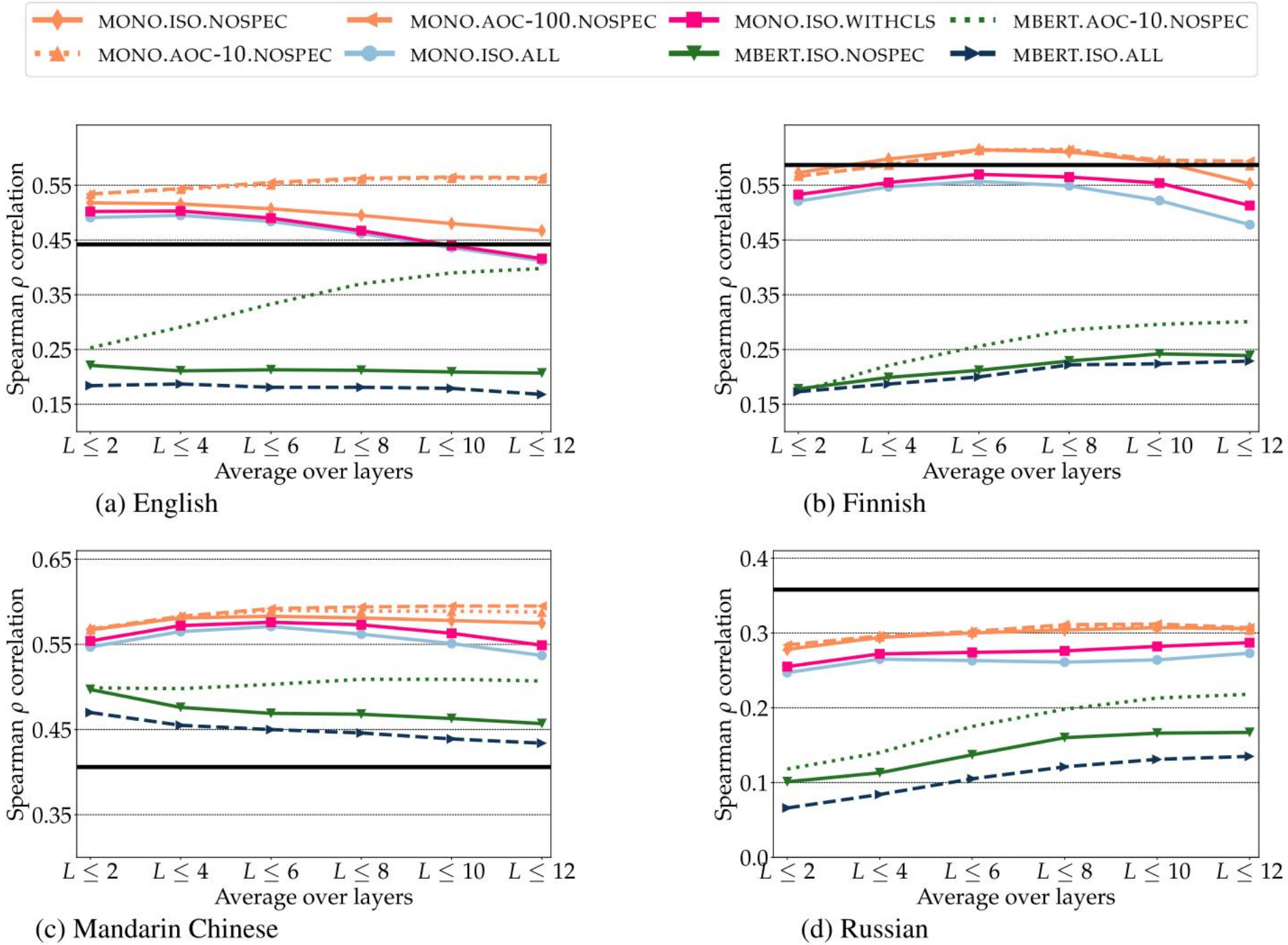
- T3: Bilingual Lexicon Induction (BLI)
  Dataset: XLING, 5K training for the mapping, 2K for test
  Goal: to retrieve target language translations for a (test) set of source language word; MRR

Tasks

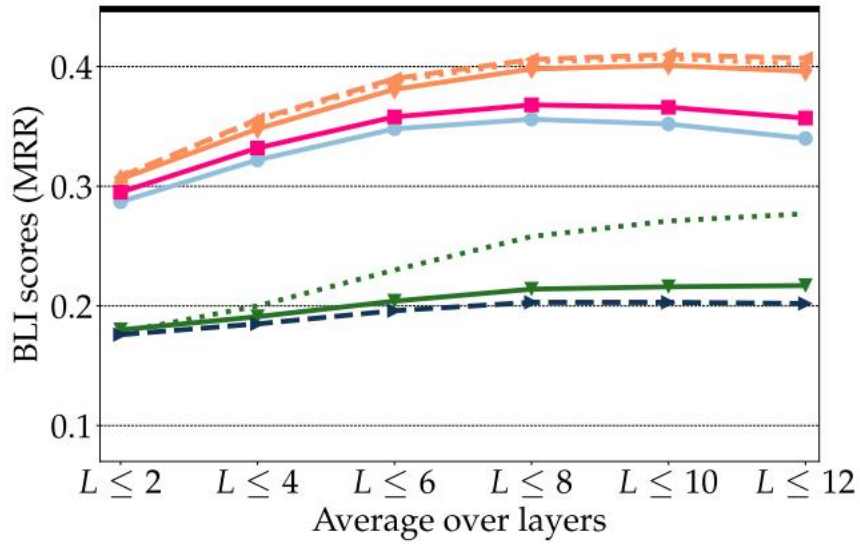
- T4: Cross-Lingual Information Retrieval (CLIR)
  - Dataset: CLEF 2003 in a document-level retrieval task
  - Metrics: MAP
- T5: Lexical Relation Prediction (RELP)
  - Goal: Relation Prediction of synonymy, antonymy, hypernymy, metonymy, plus no relation
  - Dataset: WordNet-based 10K word
  - Metrics: micro-averaged F1 score

Re

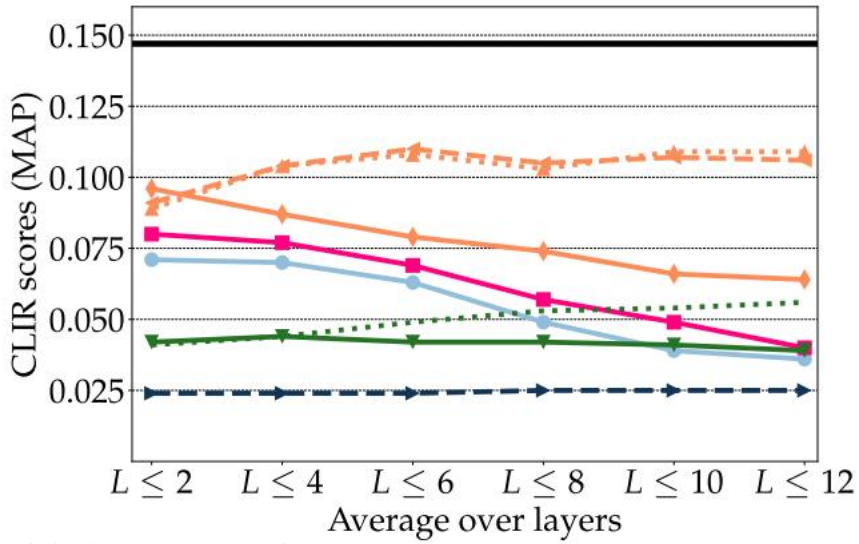
- LSIM


MONO.ISO.NOSPEC
MONO.AOC-100.NOSPEC
MONO.ISO.WITHCLS
MBERT.AOC-10.NOSPEC
MONO.AOC-10.NOSPEC
MONO.ISO.ALL
MBERT.ISO.NOSPEC
MBERT.ISO.ALL
Spearman ρ correlation
0.55
0.45
0.35
0.25
0.15
L ≤ 2
L ≤ 4
L ≤ 6
L ≤ 8
L ≤ 10
L ≤ 12
Average over layers
(a) English
(b) Finnish
0.65
0.55
0.45
0.35
(c) Mandarin Chinese
0.4
0.3
0.2
0.1
0.0
(d) Russian

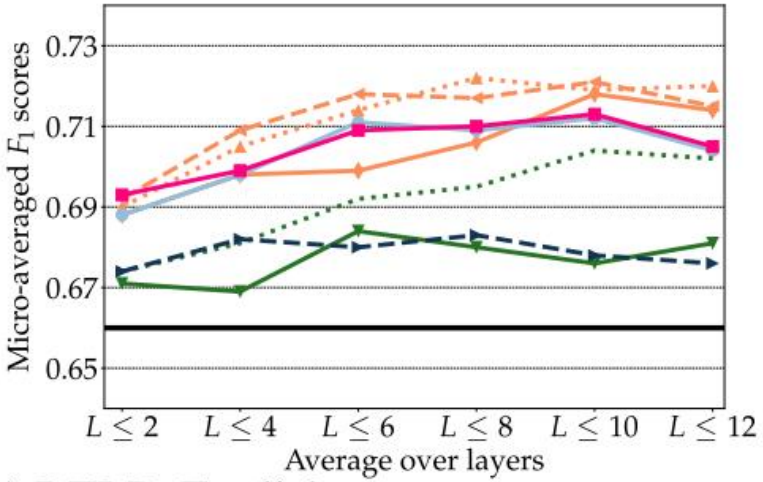
Results



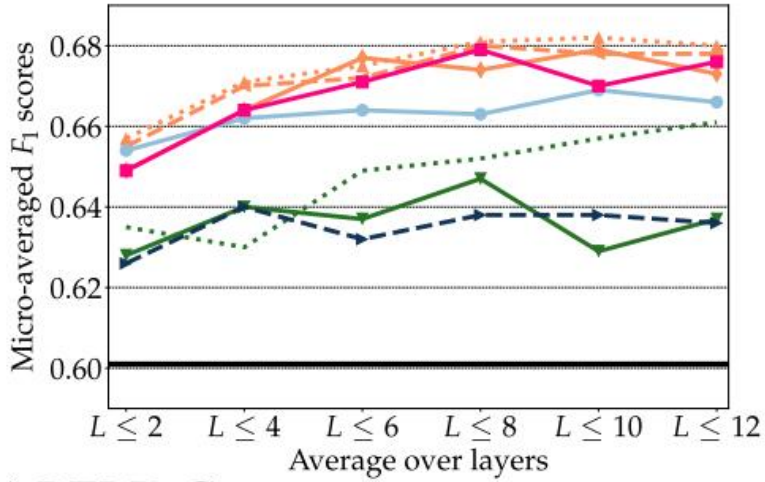
(a) Summary BLI results



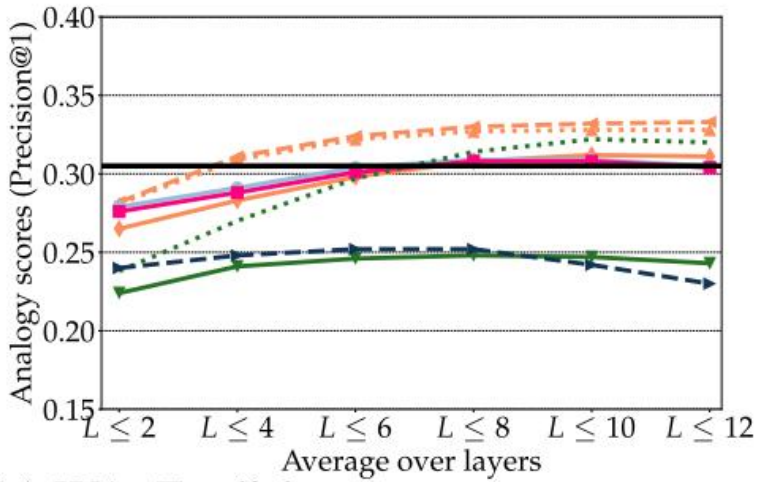
(b) Summary CLIR results



(a) RELP: English



(b) RELP: German



(c) WA: English

Contributions

Research Questions:

- Q1: language-specific or universal for lexical extraction strategies?
- Q2: Is lexical knowledge concentrated or scattered in NNs?
- Q3: how well does “BERT-based” static word embedding?
- Q4: Do monolingual LMs learn similar representations for words denoting similar concepts (i.e., translation pairs)?

Q1,Q4: consistency; Q2: distribution Q3: aggregation

Results

- monolingual vs. multilingual LMs: [Q1] MULTI> MONO
- How important is context? [Q3] yes, less instances is okay!
- How important are Special Tokens? [Q3] no, better without them!
- How important is layer-wise Averaging? [Q3]

  marginally better from bottom to top (lower layer may have type-level lexical knowledge); better than L0 and average all;
- Comparison to Static Word Embeddings. [Q3]

  better (RELP); worse (BLI&CLIR) and mixed (LSIM&WA)
- Differences across Languages and Tasks. [Q1,4] Some variation

Lexical Information in Individual Layers

- Cross-lingual and cross-layer consistency
- Similarity Metrics: centered kernel alignment (CKA)

$$\mathrm{CKA}(X, Y) = \frac{\|Y^\top X\|_{\mathrm{F}}^2}{(\|X^\top X\|_{\mathrm{F}} \|Y^\top Y\|_{\mathrm{F}})}$$

- Self-similarity: CKA similarity among different layers
- Bilingual layer correspondence: CKA for translation pair for the same layer

Cross-layer

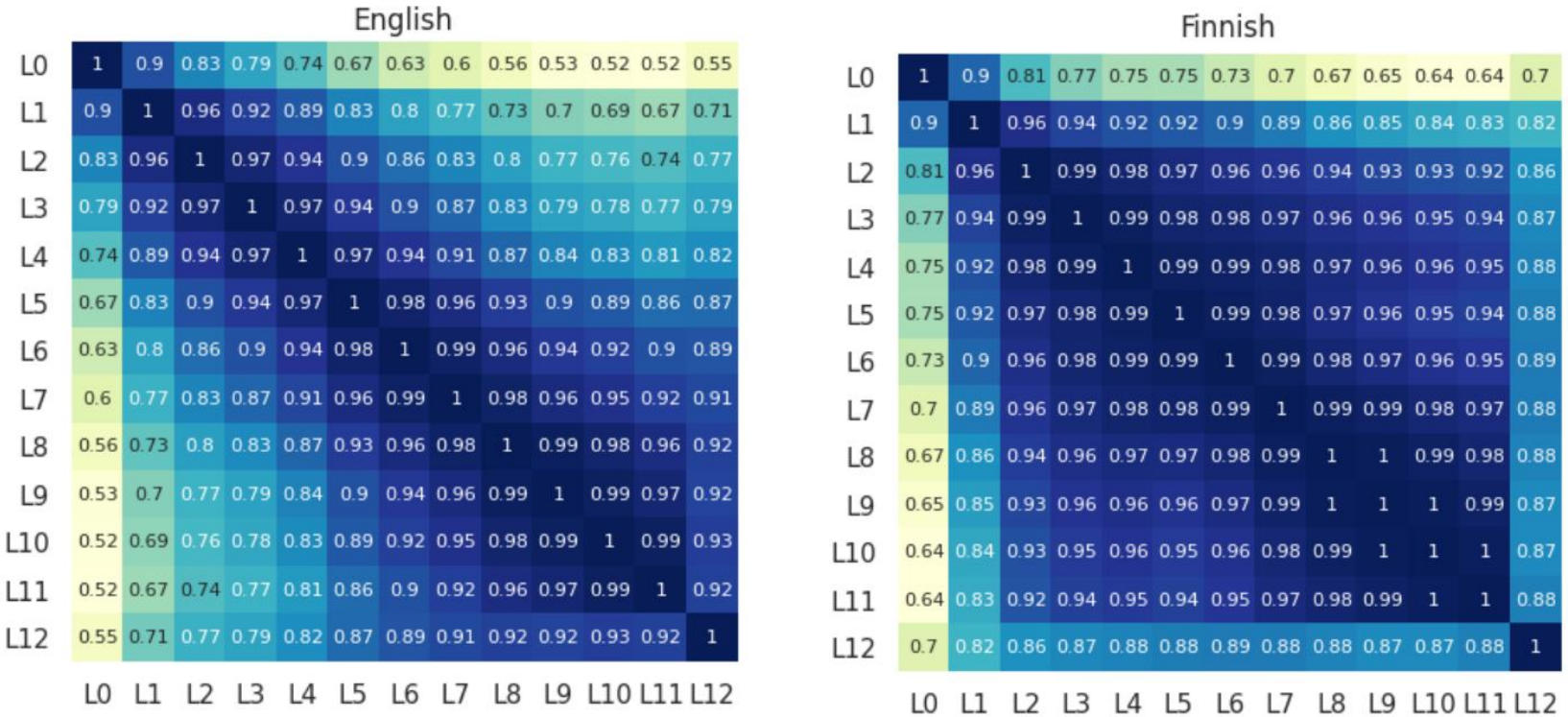


Figure 7: Self-similarity heatmaps: linear CKA similarity of representations for the same word extracted from different Transformer layers, averaged across 7K words for English and Finnish. MONO.AOC-100.NOSPEC.

Bilingual

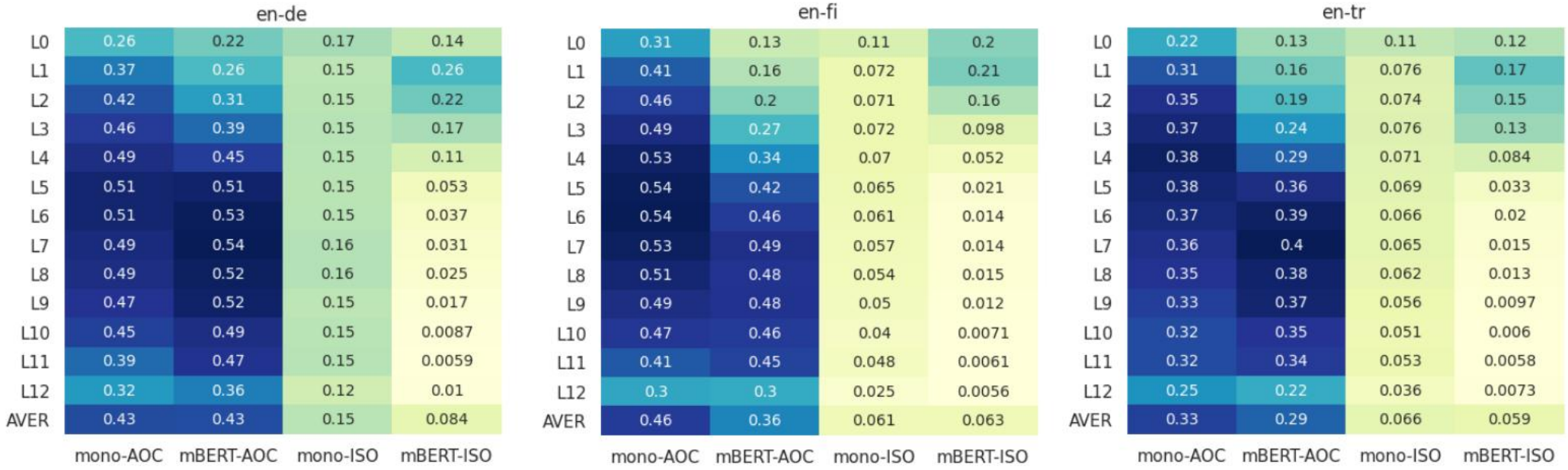
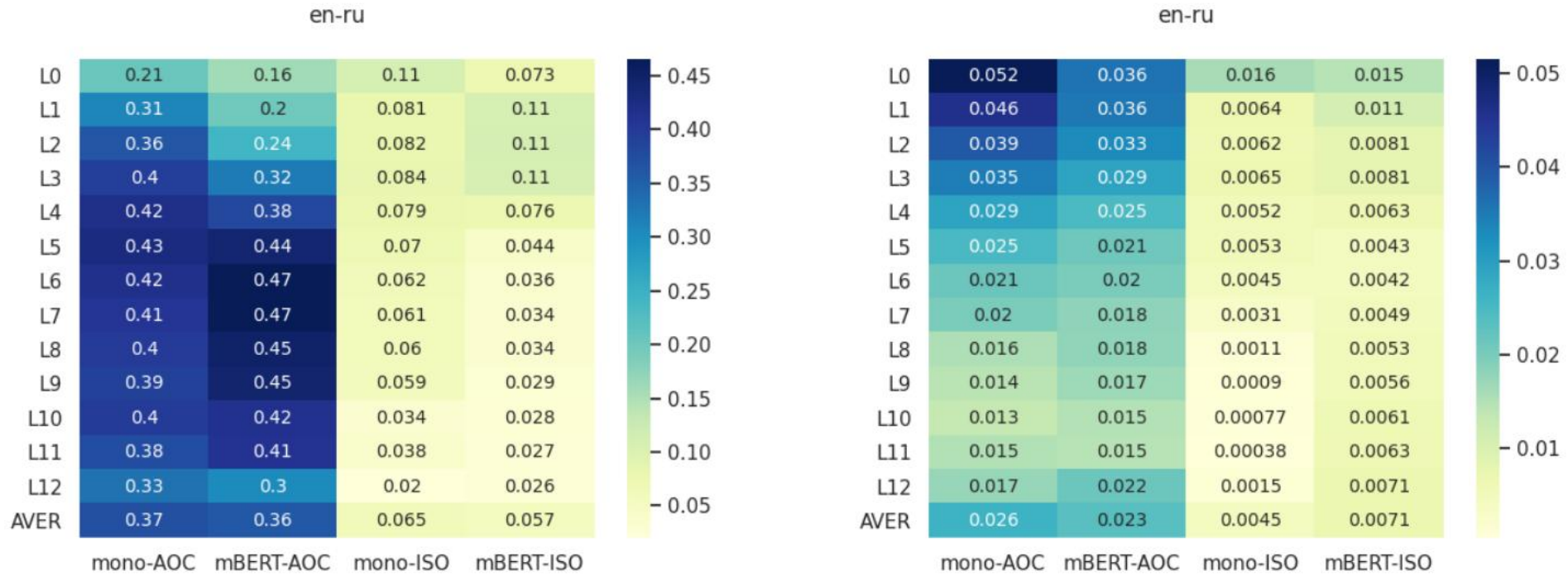


Figure 5: CKA similarity scores of type-level word representations extracted from each layer (using different extraction configurations, see Table 1) for a set of 7K translation pairs in EN–DE, EN–FI, and EN–TR from the BLI dictionaries of Glavaš et al. (2019). Additional heatmaps (where random words from two languages are paired) are available in the appendix.

Bilingual



(a) EN–RU: Word translation pairs (b) EN–RU: Random word pairs

Figure 6: CKA similarity scores of type-level word representations extracted from each layer for a set of **(a)** 7K EN–RU translation pairs from the BLI dictionaries of Glavaš et al. (2019); **(b)** 7K random EN–RU pairs.

Layer-wise performance

		L_0	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	L_9	L_{10}	L_{11}	L_{12}
LSIM	EN	.503	**.513**	.505	.510	.505	.484	.459	.435	.402	.361	.362	.372	.390
	FI	.445	**.466**	.445	.436	.430	.434	.421	.404	.374	.346	.333	.324	.286
WA	EN	.220	.272	**.293**	.285	**.293**	.261	.240	.217	.199	.171	.189	.221	.229
BLI	EN–DE	.310	.354	.379	**.400**	.394	.393	.373	.358	.311	.272	.273	.264	.287
	EN–FI	.309	.339	.360	.367	**.369**	.345	.329	.303	.279	.252	.231	.194	.192
	DE–FI	.211	.245	.268	.283	.289	**.303**	.291	.292	.288	.282	.262	.219	.236
CLIR	EN–DE	.059	**.060**	.059	**.060**	.043	.036	.036	.036	.027	.024	.027	.035	.038
	EN–FI	.038	**.040**	.022	.018	.011	.008	.006	.006	.005	.002	.003	.002	.007
	DE–FI	.054	**.057**	.028	.015	.016	.022	.017	.021	.020	.023	.015	.008	.030

Table 2: Task performance of word representations extracted from different Transformer layers for a selection of tasks, languages, and language pairs. Configuration: MONO.AOC-100.NOSPEC. Highest scores per row are in bold.

Type-level lexical information is available in *lower layer*. Other work [Kawin Ethayarajh. 2019] suggests higher layers are more *context-specific*

Conclusion

- Thorough experiments to analyze representations and lexical semantics across different languages.
- Some recommendations:

(1) monolingual LMs (2) encoding words with multiple contexts

(3) excluding special tokens (4) averaging over lower layers

- Future directions:

(1) Larger models? GPT-like models? [A1]

(2) How corpora affect AOC configurations? (Why beyond what)

https://arxiv.org/abs/2403.01509

Fantastic Semantics and Where to Find Them: Investigating Which Layers of Generative LLMs Reflect Lexical Semantics

Annual Review of Linguistics

Distributional Semantics and Linguistic Theory

Gemma Boleda[1,2]

[1]Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona 08018, Spain; email: gemma.boleda@upf.edu

[2]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona 08010, Spain

Annual Review of Linguistics, 2020; Cited by 228

Authors and affiliation

- Gemma Boleda, ICREA Research Professor in the Department of Translation and Language Sciences of the Universitat Pompeu Fabra (Barcelona, Spain)
- ICREA: 278 researchers in all fields of knowledge, from philosophers to astrophysicists, that perform their research in 48 different host institutions in Catalonia.

Catalan Institution for Research and Advanced Studies, is a foundation supported by the Catalan Government and guided by a Board of Trustees.

https://gboleda.github.io/

Overview

- Distributed semantics provides multidimensional, graded, empirically induced word representations.
- Limited impact in theoretical linguistics
- This paper reviews methods and results relevant for the areas
  1. Semantic Change
  2. Polysemy and composition
  3. Grammar-semantics interface

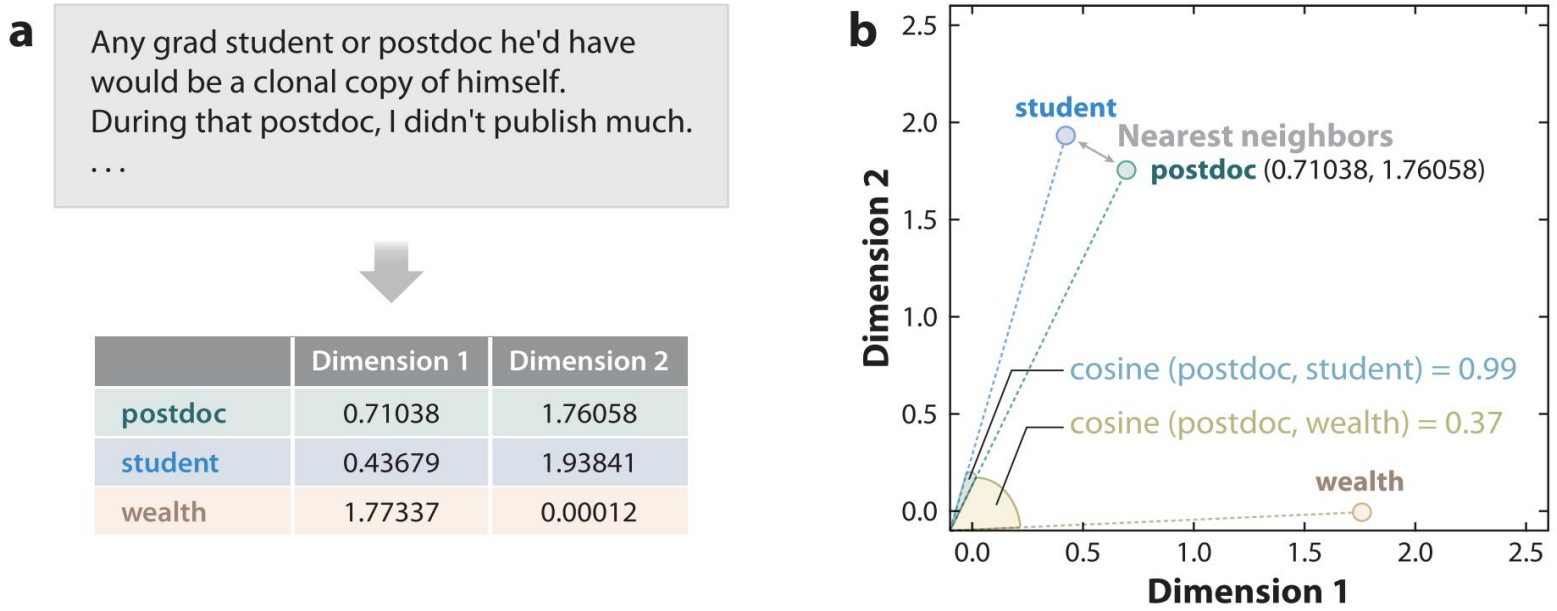
Distributional Semantics

- Distributional Hypothesis

Similarity in meaning results in similarity of word distribution (Harris, 1954)

Reverse engineer: from word distribution to a meaning (representation)

- Multidimensional and continuous space with geometric relations



Distributional Semantics

- How to get the representations?
  Co-occurrence statistics -> machine learning type of algorithms
- Difference with structuralist representation (feature, like ±editable)
  automatic (learnable) vs. manual
  multidimensionality vs. fewer features
  gradedness vs. concreteness

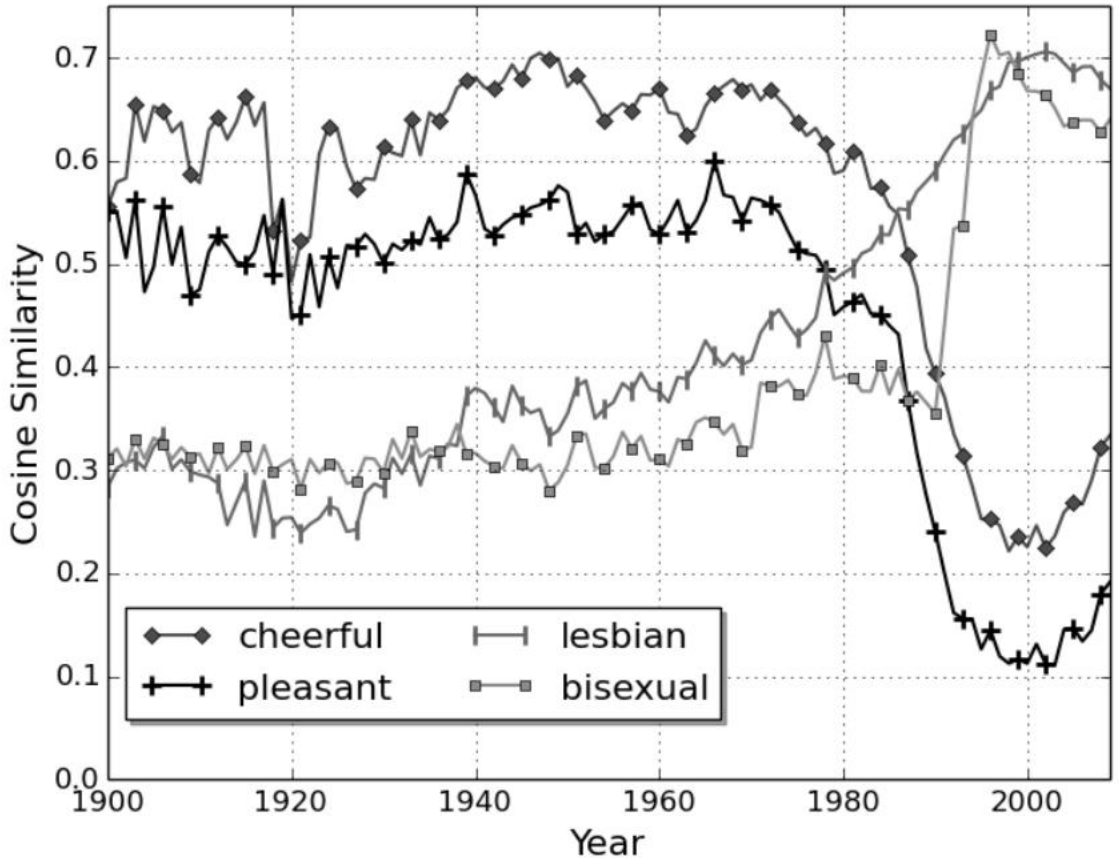
Table 1 Near-synonyms in semantic space: the words closest to *man*, *chap*, *lad*, and *guy*

Word	Nearest neighbors[a]
man	woman, gentleman, gray-haired, boy, person
lad	boy, bloke, scouser, lass, youngster
chap	bloke, guy, lad, fella, man
guy	bloke, chap, doofus, dude, fella

Semantic Change

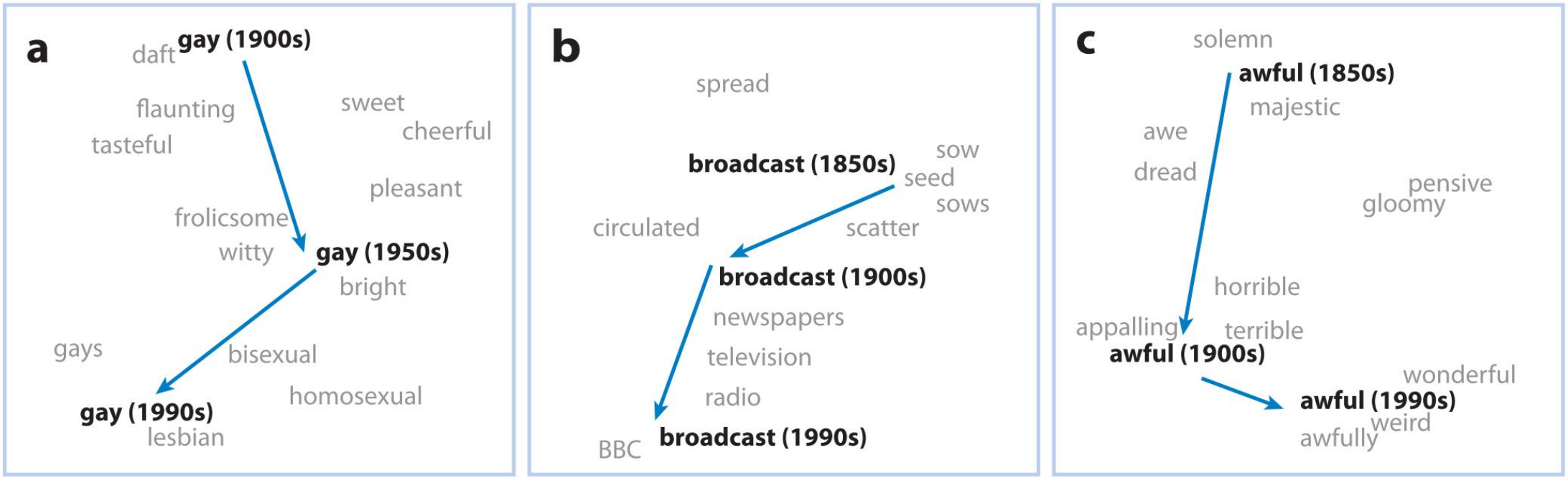
- Hypothesis: a change in context of use mirrors a change in meaning (a "diachronic" version of distribution hypothesis)
- The inference process is typically carried out by building word representations at different points in time [Kim et al. 2014]
- Traced by similarity scores or the nearest neighbors [Hamilton et al. 2016]

Semantic Change



- Similarity to “gay” in different time periods
- Traced by similarity scores
- [Kim et al, 2014]

Semantic Change



Traced by nearest neighbors [Hamilton et al. 2016]

Semantic Change

- Detection of types of semantic shift
  - Narrowing and broadening [Sagi et al. 2009]
  - grammaticalization (e.g., do) [Sagi et al. 2009]
- systematically exploring data and advancing the theory

[Xu&Kemp, 2015] assessed two previously proposed laws:

1) co-evoluation: pairs of similar words tend to evoluate together

2) differentiation: synonyms tend to evolve differently due to efficiency

Semantic Change

- Promising directions
  - detecting; locating; tacking; testing theory
  - NOTE: There are several (if not many) papers, workshops, etc
- Challenges
  - data hungry
  - spurious effects
  - functional words

Polysemy and composition

- Single Representation, Polysemy via Composition

  The (nuanced) meaning of some words is affected by context,

  For example: bake a cake (creation) vs. bake the potato (change)

  Resolution: Composition of a larger constituent

Polysemy and composition

a

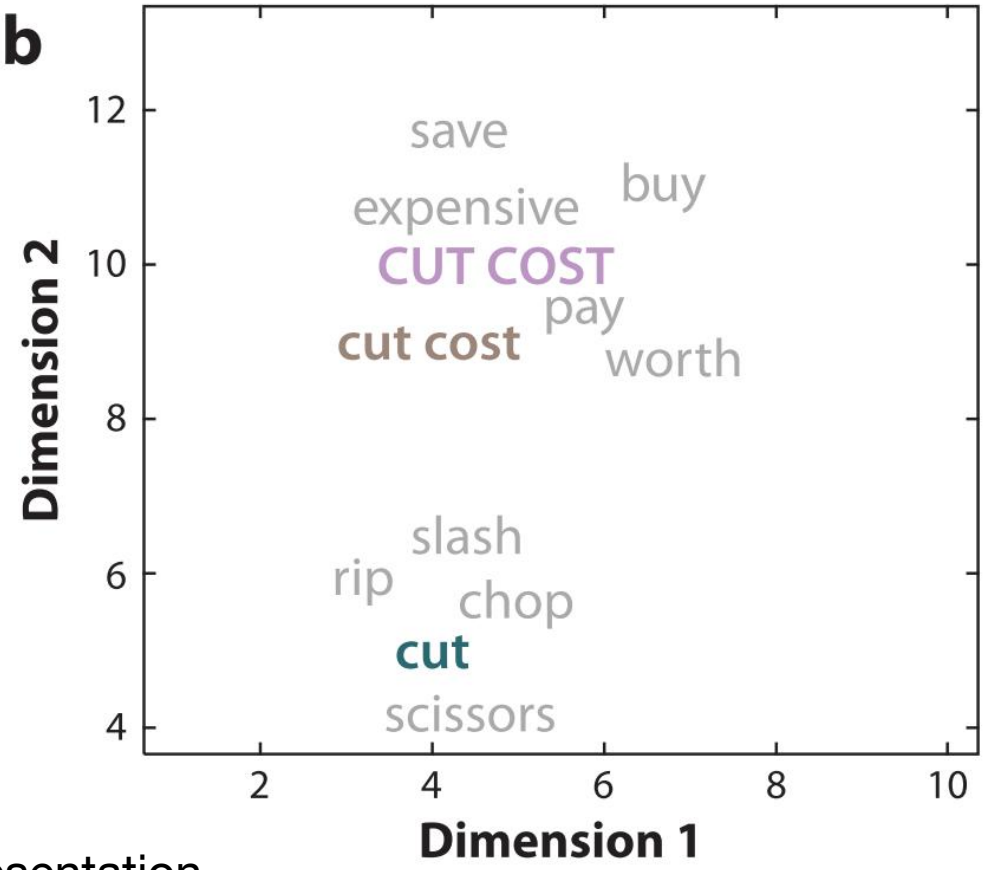
Corpus based

	Dimension 1	Dimension 2
cut	4	5
cost	1	5
cut cost	4	9

Synthetic

	Dimension 1	Dimension 2
CUT COST	5	10

b



Synthetic is approximate to the corpus based representation
Dimension 2 may represent an abstract meaning

Polysemy and composition

- Single Representation, Polysemy via Composition

  The (nuanced) meaning of some words is affected by context,

  For example: bake a cake (creation) vs. bake the potato (change)

  Resolution: Composition of a larger constituent

  Maybe “bake + cake” can represent “bake” more by dragging it into a “creation” direction/dimension.

Polysemy and composition

- Different Representation, Polysemy via Word Senses
  Sense-specific word representation
  word sense induction
  vector per word use [Schutze, 1998] (contextual representation)

Discussion

- Distributional semantics offer an elegant framework:

Multidimensionality: common+specific

gradedness: degree of “synchronic” change

Probabilistic? Modeling polysemy as uncertainty [Liu & Liu, 2023, A2]

- To make predictions and test specific hypothesis driven by linguistic theory

e.g., [Boleda et al. 2013] in “Adj + Noun” phrase, if N is more typical to A, then it is easier to predict

-> conceptual aspect vs. referential aspect

- [NOTE] Chinese has countless compound words [A3]

Grammar-semantics interface (Brief)

- Syntax-Semantics Interface

  e.g., Verbs with different arguments may mean differently

  A recent work on revertible SVO [A4]

- Morphology-Semantics Interface

  e.g., Disambiguation of affix, such as “-er”

  Compositionality

  Semantic opacity and semiopacity

  A more theoretical work: derivational affix with emotional valence [Lapesa et al. 2017]

Conclusion

- Combination of two areas
  - (1) the connection of use, meaning and grammar is relevant
  - (2) semantic relationship via geometric relationship (similarity)
  - (3) gradedness
  - (4) abstractions of the relevant semantic classes by averaging representations
  - (5) Compositionality via simple operations on representations
- Challenges
  - Data quality and quantity
  - [Note] Representational view from a black box: what a vector really tells us?

Thank you for listening!

https://juniperliuzhu.netlify.app/

A.1 Our work 1

- Fantastic Semantics and Where to Find Them: Investigating Which Layers of Generative LLMs Reflect Lexical Semantics

Large language models have achieved remarkable success in general language understanding tasks. However, as a family of generative methods with the objective of next token prediction, the semantic evolution with the depth of these models are not fully explored, unlike their predecessors, such as BERT-like architectures. In this paper, we specifically investigate the bottom-up evolution of lexical semantics for a popular LLM, namely Llama2, by probing its hidden states at the end of each layer using a contextualized word identification task. Our experiments show that the representations in lower layers encode lexical semantics, while the higher layers, with weaker semantic induction, are responsible for prediction. This is in contrast to models with discriminative objectives, such as mask language modeling, where the higher layers obtain better lexical semantics. The conclusion is further supported by the monotonic increase in performance via the hidden states for the last meaningless symbols, such as punctuation, in the prompting strategy.

https://arxiv.org/abs/2403.01509 Under Review

A.2 Our work 2

Ambiguity Meets Uncertainty: Investigating Uncertainty Estimation for Word Sense Disambiguation

- Word sense disambiguation (WSD), which aims to determine an appropriate sense for a target word given its context, is crucial for natural language understanding. Existing supervised methods treat WSD as a classification task and have achieved remarkable performance. However, they ignore uncertainty estimation (UE) in the real-world setting, where the data is always noisy and out of distribution. This paper extensively studies UE on the benchmark designed for WSD. Specifically, we first compare four uncertainty scores for a state-of-the-art WSD model and verify that the conventional predictive probabilities obtained at the end of the model are inadequate to quantify uncertainty. Then, we examine the capability of capturing data and model uncertainties by the model with the selected UE score on well-designed test scenarios and discover that the model reflects data uncertainty satisfactorily but underestimates model uncertainty. Furthermore, we explore numerous lexical properties that intrinsically affect data uncertainty and provide a detailed analysis of four critical aspects: the syntactic category, morphology, sense granularity, and semantic relations.

https://aclanthology.org/2023.findings-acl.245/ Findings: ACL 2023

A.3 Our work 3

- 基于词向量的汉语复合词内部语义关系的量化研究

复合构造是汉语词汇最常见的构词方式，它使得词汇内部的语义关系更为明显，也同时具有较强的可分析性。另一方面，计算语言学领域通过大规模语料学习词向量来作为词汇的语义表征。前人的研究大多关注于词向量如何对下游任务起作用，或者词汇之间的语义依赖，却较少关注词汇内部的语义关系，这对于以复合词构词为主、内部结构可分析的汉语来说，不失为一个重要缺失。本文探究了复合词向量可否如实反映两种语义关系：主导性和可组合性。前者表明复合词的哪一部分从语义上讲更加重要；后者体现了整体的词义多大程度上可以通过部分的意义推导出来。本文的研究发现通过词向量对于这两种关系的判断基本与语言学中的吻合。同时，通过对大规模词汇词向量的词义发掘，可以推断出主导性和可组合性受到多种因素影响，这些因素对于新词预测、语言教学、词典编撰等具有一定的实用参考价值。

Script

A.4 Our work 4

- 基于大语言模型的汉语主宾可逆句语义与施事程度评估

主宾可逆句是现代汉语语法体系中较为特殊的语法现象，其基本特征是句子中的主语和宾语可以相互交换位置而不影响基本句义。换句话说，交换成分的语义角色并未发生明显改变，从而产生了“格配置变动”。与之相对的主宾不可互逆句则往往由于论元与核心动词的典型施受关系，交换主宾语后语义发生反向改变。大语言模型在大规模语料上进行训练，并取得了卓越的文本理解能力。一个值得探究的问题是，它能否正确区分这两种情况，以及理解背后的施受语义关系？本文收集了相关的语料，并对现有的汉语大语言模型进行句子等义性和各个成分的施事度进行评估。并得出如下结论：1）一般取末层所有词向量的平均方式不足以区别两种情况；2）通过采用不同层的信息，大语言模型可以在一定程度上反映不同位置上的施事程度。

Script