论文分享

2025.05.08

刘柱

Rethinking Word Similarity: Semantic Similarity through Classification Confusion

Kaitlyn Zhou, Haishan Gao, Sarah Chen, Dan Edelstein, Dan Jurafsky, Chen Shani
Stanford University
{katezhou, hsgao, sachen, danedels, jurafsky, cshani}@stanford.edu

Embedding derived animacy rankings offer insights into the sources of grammatical animacy

Vivian G. Li
Yale University, New Haven, CT, USA
liguo.vivian@gmail.com

NAACL 2025

- NAACL: Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics
- 每年一次，自然语言处理领域顶会之一，CCF B类会议
- 2025年开会地点：新墨西哥州阿尔伯克基； 4.29-5.4
- https://2025.naacl.org/
- https://aclanthology.org/events/naacl-2025/

NAACL 2025 Overview

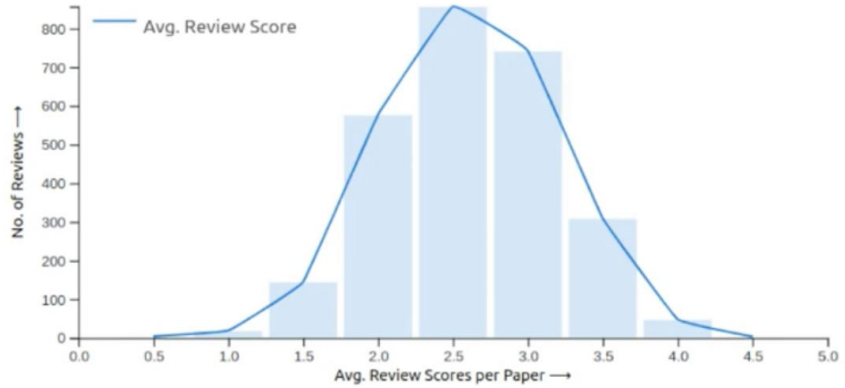
Review Statistics for NAACL

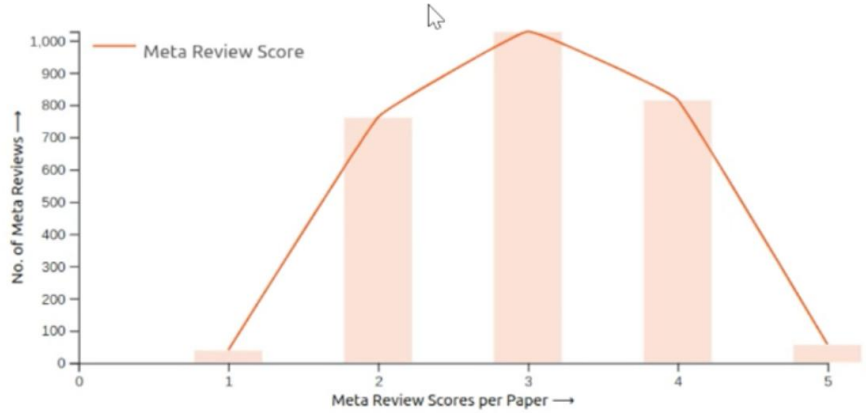
ARR 2024/10
3185 submissions
~3000 reviewers, ~550 area chairs assigned
82 submissions: additional ethics review

Big thanks to ARR team, esp. Viviane Moreira, Anna Rogers, Michael White

NAACL Commitment
1647 submissions
91% from ARR 2024/10
Ranked by 98 senior area chairs

Accepted: 719 Main Conference & 477 Findings





NAACL 2025 Overview

Acceptance Rate for Main Conference Papers

Multi-stage ARR review + commitment
Following precedents in *ACL conferences, we used:

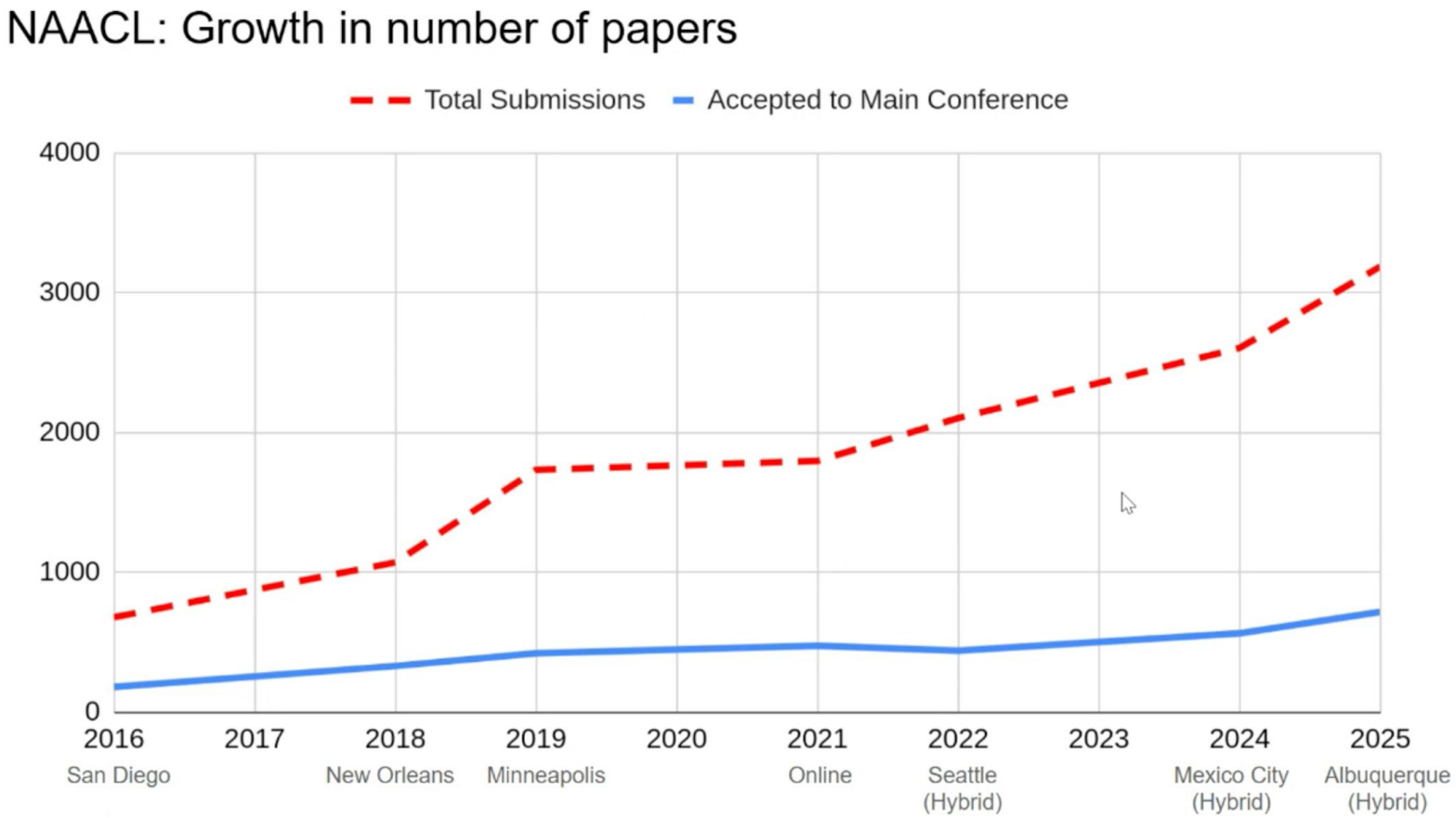
Acceptance rate = (#accepted) / (#contenders)

#contenders =

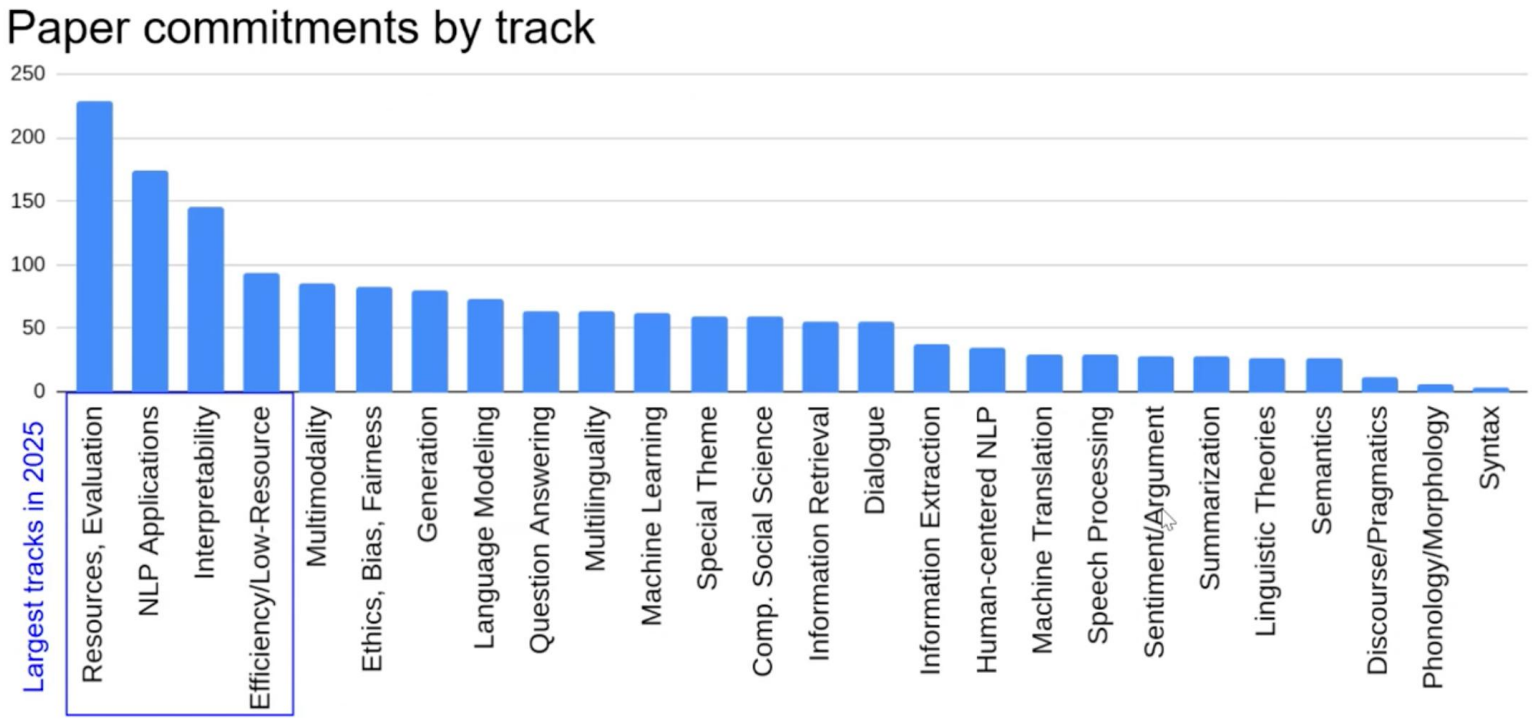
- submissions to ARR 2024/10 that specified NAACL as preferred venue, no preferred venue, or committed to NAACL (3099 submissions)
- + committed to NAACL from previous cycles (147 submissions)

Acceptance rate = 719 / 3246 = 22.15%

NAACL 2025 Overview



NAACL 2025 Overview



Largest tracks by submissions in

2024: NLP Applications, Resources/Evaluation, Interpretability, Efficiency/Low-Resource

2021: NLP Applications, Machine Learning, Information Extraction, Dialogue

2018: Information Extraction, Semantics, Machine Translation, Summarization

Special Theme

Special Theme: NLP in a Multicultural World

- Representative topics (not an exhaustive list):
  - **Cultural localization** of language models.
  - New NLP applications to **support people from diverse cultures**.
  - Analysis of **cultural biases** in language models.
- We received **71** theme track submissions this year
  - 31 papers accepted to the main conference
  - 23 accepted to “Findings”
  - 2 paper awards (best theme paper and runner up)
- Special Theme Sessions:
  - ST.1: Today - **2pm to 3:30**
  - ST.2: Tomorrow - **2pm to 3:30**

Three Keynotes

Keynote Speaker: Rada Mihalcea

LOCATION: BALLROOM A - C

Title: The Power of Many: How Cross-Cultural NLP Will Drive Innovation and Lead to Better Systems

Abstract: In recent years, NLP has made remarkable strides, with language and language-vision models transforming applications across a wide range of domains—including healthcare, education, social sciences, humanities, the arts. Yet, many of these models and the datasets they rely on reflect only a small fraction of the world's population, which often leads to gaps in performance, biases, and missed opportunities for impact. In this talk, I will draw on insights from over a decade of cross-cultural research to highlight key limitations in current benchmarks and models, from their narrow linguistic and cultural scope to challenges in evaluation and application design. I will share key lessons learned along the way and make the case for cross-cultural NLP—one that more effectively captures the diversity of behaviors, beliefs, and linguistic expressions across different communities—not only as a way to build stronger and more robust systems but also to foster an even more innovative research community.

Keynote Speaker: Mike Lewis

LOCATION: BALLROOM A - C

Title: Science and Scaling: How (really) to Pre-train a Llama

Abstract: Pre-trained language models form the basis for much of modern NLP, and while the basic pre-training recipe is well known, many details of model development are hidden in secretive research labs. Based on experience training Meta's Llama models, I will shed light on the evolving science of pre-training. The central challenge of pre-training research is in designing small scale experiments that can confidently be extrapolated to main training runs at orders of magnitude greater scale - which has led to an approach that differs from much of academic research. While careful experiments can disentangle many of the factors that do and don't matter for large scale training, I will also discuss gaps in the science that mean researcher judgements and intuitions remain critical to decision making.

Scaling Intelligence the Human Way

Josh Tenenbaum
MIT

Other interesting papers

- Keywords: Semantics, meaning,

Rethinking Word Similarity: Semantic Similarity through Classification Confusion

Kaitlyn Zhou, Haishan Gao, Sarah Chen, Dan Edelstein, Dan Jurafsky, Chen Shani
Stanford University
{katezhou, hsgao, sachen, danedels, jurafsky, cshani}@stanford.edu

NAACL 2025, Long main

About the authors

- Kaitlyn Zhou, 6-year-old phd in Stanford, incoming AP at Cornell
- focus on: model overconfidence; context-aware evaluation; human-LM interactions
- supervisor： Dan Jurafsky
- Professor both in Humanities (Linguistics) and Computer Science
- Standford NLP

Professors

Chris Manning

Linguistics & Computer Science

Dan Jurafsky

Linguistics & Computer Science

Percy Liang

Computer Science

Chris Potts

Linguistics

Tatsunori Hashimoto

Computer Science

Monica Lam

Computer Science

Diyi Yang

Computer Science

Yejin Choi

Computer Science & HAI

背景

- 语义相似度计算在很多领域都有应用
  - 计算社会科学
  - 数字人文领域
  - NLP应用中对于词义共时和历时的分析
  - 文化分析、历时文本分析
- 向量空间模型
  - 使用向量表征词义
  - 使用余弦相似度来衡量向量/语义的相似性

余弦相似度的局限性

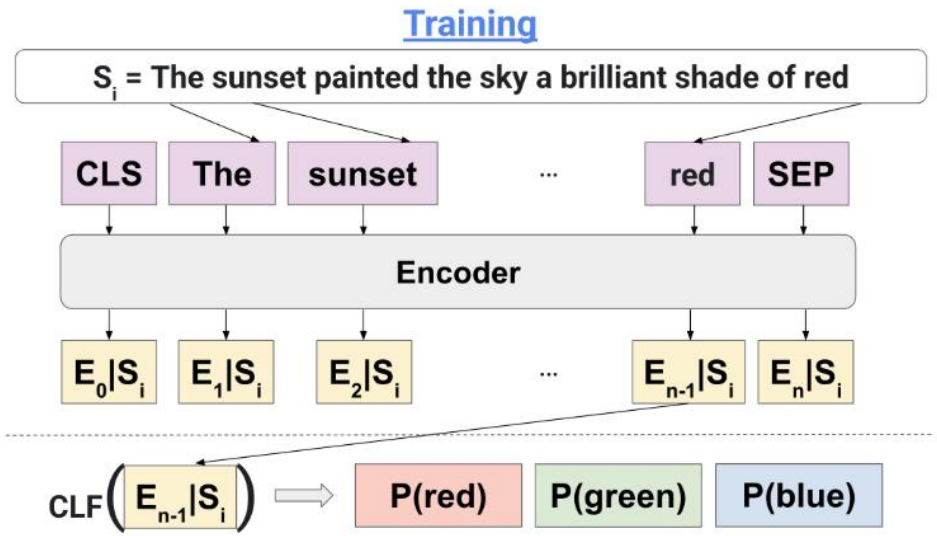
- 由于上下文嵌入空间的非各向同性（anisotropy），余弦相似度容易被少数“异常维度”所主导
  - 向量集中在空间的某个区域(Timkey and Van Schi- jndel, 2021; Ethayarajh, 2019)
- 低估高频词的相似性(Zhou et al., 2022a)
- 无法捕捉语义关系的不对称性（Vilnis and McCallum, 2014)
  - 猫和动物的相似度 比 动物和猫的相似度 更大
- 常常不能匹配人类的语义判断(Nematzadeh et al., 2017,Sitikhu et al., 2019).
- 可解释性较弱，难以捕捉各个方面（语义特征）的相似性(Tversky, 1977; Ettinger and Linzen, 2016; Zhou et al., 2022a, inter alia)
  - 例如：面包和饮料在“可食用”方面相似，但是在质态上不同

创新点

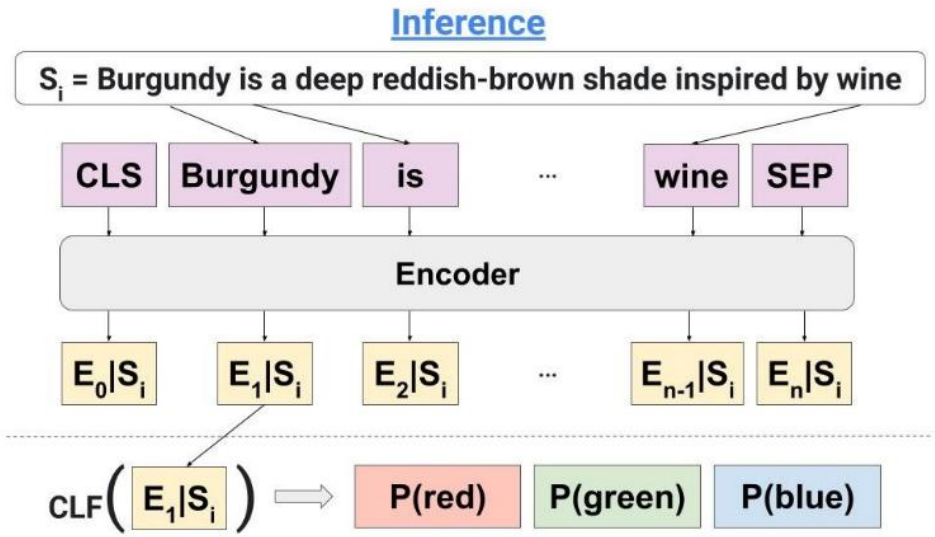
- 重新提出一个度量相似性的方法
  - 词混淆（word confusion）的方式：混淆目标词和干扰词的向量，能否还原回原来的词（根据分类的概率来表示还原程度）
  - I like a cat. -> I like a {dog, animal, table...}.
  - 干扰词：特征
  - 词义的相似性取决于特征的相互交换性
- 在基准实验上进行验证
  - 与余弦相似度进行对比
- 在真实的文化分析数据集上进行验证
  - 一个语义随时间变化的实例

方法——词混淆

- 干扰词集：特征/种子词。例如{red, green, blue}
- 目标词：想要研究词义的词，可以在干扰词集里面，也可以不在
- 分类器训练
  - 找到所有干扰词集中词所在的上下文
  - 提取上下文向量作为分类器输入
  - 输出为干扰词集对应的词，每一个词就对应一个类别
  - 分类损失：重构任务（类似于BERT的掩码预测任务）
- 测试阶段
  - 找到目标词所在上下文，并提取向量
  - 送入到训练阶段学习到的分类器中，分类概率作为与相应词的“相似度”
  - “相似度”体现为混淆后还可以找出原始类别的“概率”



(a) Training *Word Confusion*: The classifier is trained in a self-supervised manner, after selecting the desired features (in this example the classes red, green blue). We extract sentences containing those 3 "feature" words. The input to the classifier is the contextual embedding of the class token, e.g., the BERT embedding of the word "red" in the sentence "The sunset painted the sky a brilliant shade of red". The classifier is trained to predict a class ("red") from that contextual embedding.



(b) *Word Confusion* inference: We are given the classifier and the predetermined set of classes, which will act as features, in this case red, green, blue. Given a target word in a sentence, e.g., *"burgundy"*, we extract its contextual embedding in that sentence e_{burgundy} and compute $P(w_i|e_{\text{burgundy}})$ for each class i. The classifier's confusion matrix then define the similarity of the burgundy with each class. The input word can be one of the feature words (red, green, blue) or not (burgundy).

优点

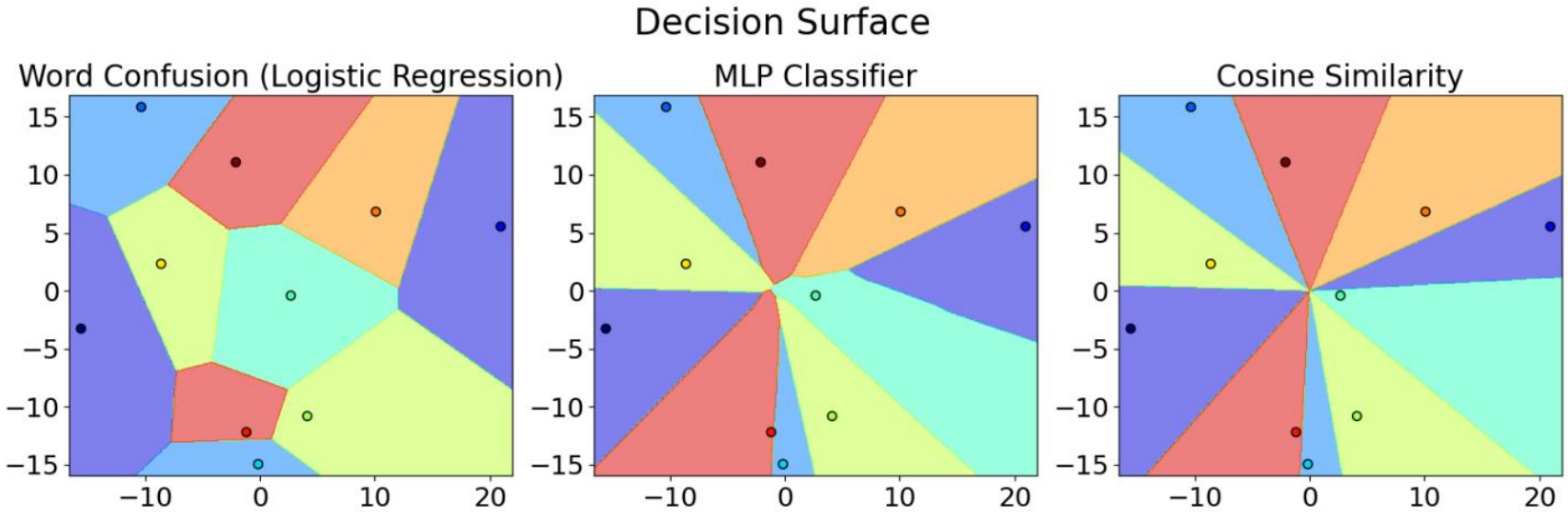


Figure 2: Differences in decision boundaries between *Word Confusion* and cosine similarity. The x and y axes represent two dimensions of an artificially constructed set of data points. Note how cosine similarity's boundaries originate from the origin whereas *Word Confusion*'s are not limited in the same way.

- 非对称性
- 可解释性
  - school: {positive, negative}, or {fun, work}

实验

- MEN: 3000 word pairs annotated by 50 humans. for example, {berry, seed}, {game, hockey}, and {truck, vehicle} with 1-7 scores, 0.68
- WordSim353 (WS353): 2000 pairs, 84% ITA
- SimLex 1000 word pairs, 0.67

Method \ Dataset	MEN	WS353	SimLex
Cosine	0.59	0.54	0.39
Word Confusion	**0.66**	**0.67**	**0.44**

Table 1: Spearman's ρ correlation between *Word Confusion* and cosine similarity results as compared to humans. These three benchmarks focus on slightly different aspects of word similarity. We measure the correlation between human scores and cosine similarity between the language model embeddings versus *Word Confusion*'s similarity scores. As can be seen, our method slightly outperforms cosine similarity.

实验：特征分类

- 情感分类
  - NRC corpus
  - 种子词{positive, negative}
- 语法性别分类
  - “flower” is feminine in French and masculine in Italian.
  - Italian and French nouns (Sahai and Sharma, 2021)
- 语义类别分类
  - ConceptNet class
  - Fashion-Gaming; Sea-Land

结果

Experiment	*Word Confusion*	Cosine 1	Cosine 2	Cosine 3	Ave. Cosine
Sentiment Classification	**0.79**	0.75	0.71	0.84	0.73
Grammatical Gender (Italian)	**0.93**	0.80	0.80	0.71	0.77
Grammatical Gender (French)	0.85	**0.86**	**0.86**	0.83	0.85
ConceptNet Domain (Fashion-Gaming)	0.90	**0.93**	**0.93**	0.90	0.92
ConceptNet Domain (Sea-Land Animals)	**0.83**	0.79	0.80	0.61	0.73
Average	**0.86**	0.83	0.82	0.76	0.80

Table 2: Macro-F1 for *Word Confusion* and cosine similarity across a variety of feature classification tasks. We operationalize cosine similarity in three ways: 1) the distance between the centroids of the seed words and the target words 2) the average distance each of the target word to the centroid of the seed words 3) the average distance of each target word to each seed word (no centroids).

实验三：What Is A Revolution?

- Revolution在法语文献中法国大革命后发生了语义演变
  - 一开始是和人民语义相关
  - 之后转移到了政府
- 种子集{people, government}
- 目标词：revolution & counter-r
- 实验表明：
  - 证明了上述路径
  - 发现了一开始government先和反革命联系在一起的（new）

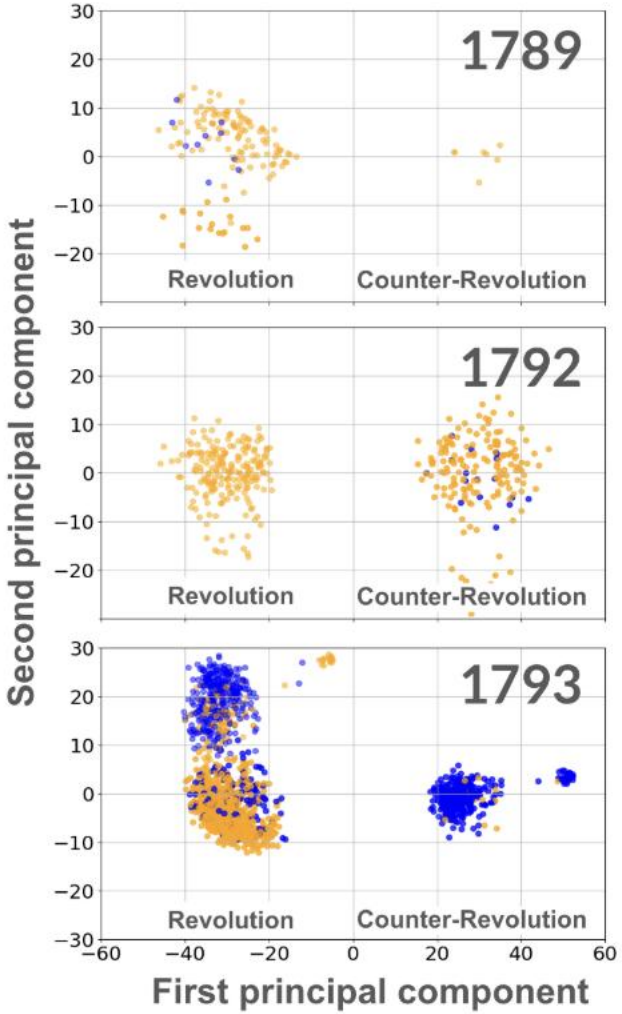


Figure 3: In 1789, the word *"revolution"* was primarily associated with popular action (represented in orange). In 1792 *"revolution"* was now also seen as something that the government should lead (represented in blue) found in the *"counter-revolution"* cluster. In 1793, this new governmental meaning had spread back to the word *"revolution"* itself.

结论

- 使用词混淆的方式来定义语义相似度
- 可解释性：可以通过设置种子来评估不同的方面
- 让相似度更加非对称
- 可以应用于人文或者社会学研究中。

Embedding derived animacy rankings offer insights into the sources of grammatical animacy

Vivian G. Li
Yale University, New Haven, CT, USA
liguo.vivian@gmail.com

NAACL'25 Long

关于作者

- Vivian G. Li

背景

- 语法是如何形成的?
  - 语用 vs. 认知 vs. 语法
  - 语用：基于语言事实和具体用例的 -》 分布式假说，频率、语言模型
  - 认知：人类的感知，与心理、认知、文化息息相关 （区分：底层认知）
  - 语法：语言中出现的形式规则，往往有不规则的部分
  - 语用决定语法 vs. 底层认知决定（塑造）语法(Chomsky, 1965, a.o.)
- 三者常常相关但并非总是一致的
  - 语法体现在语用中，往往受到认知影响，并影响认知
  - 例如：主语-动词-宾语 的语序是汉语的常规语序，大量的语句为这种规则提供支撑，同时主语的部分往往是施事，宾语部分往往是受事
  - 存在反例：受事主语句（便道走行人）
  - 语法性别和认知的性别：German Mädchen (‘girl’) as grammatically neuter
  - 语法格和认知（语义角色）的差异：受事用主格标记，日语的が一般标记主格（施事），但是在“喜欢”等心理动词的受事也用が标记

背景

- 支撑生成语言学派的观点需要分布式无法解释语法形式的例子
- 生命度等级（grammatical animacy）是其中的一个例子
- 语法层面的生命度等级 - 规则总结
- 感知层面的生命度等级 - 人类评分
- 分布层面的生命度等级 - 语言模型

贡献

- 使用语义映射（semantic projection）的方法通过向量的方式来表征生命度等级
- 实验表明，模型通过分布式假说学习到的生命度与人类感知很一致
- 但是它与语法生命度等级不一致

相关工作

- 语义映射方法
  - 通过词向量的方式可以表征类比关系(Mikolov et al., 2013)
  - 不仅仅表示离散的特征，还可以表示17个连续语义特征（例如大小、危险程度），然后不包含生命度等级
- 基于人类评分的生命度等级
  - 生命度：类人的、可以运动思考生殖等(VanArsdall et al., 2017; VanArsdall and Blunt, 2022)
  - 人类进行评分（Radanovi´c et al. (2016) 126名词，VanArsdall and Blunt (2022) 1200名词）
  - 总体排序：Animals > Humans > Inanimate Entities

相关工作

Animate		
Human	Nonhuman	Inanimate
-man	*-ma*	-

Table 1: Animacy modulates plural marking in the Gudandji dialect of Wambaya (Aguas, 1968:5-6; cited in Santazilia, 2020: Table 7)

- 基于语法规则的生命度等级
  - 在一些语言中不同生命度可能会用不同的格标记，同时也会影响性数格以及一致性等(Corbett, 2006, 2012; Comrie, 1989; Croft, 1990; Ortmann, 1998; Santazilia, 2019, 2020; Silverstein, 1976; de Swart et al., 2008, a.o.)
  - 一些语言中是间接影响，存在与其他结构的蕴含共性
  - 例如在英语中，有生名词的属格倾向于使用’s，无生名词的则使用of
  - 有生名词常出现在主格的位置，无生名词常出现在宾格的位置，如果无生名词出现在主格时候，往往使用被动
  - 序列：

Speaker (1st person pronoun) >

Addressee (2nd person pronoun) > 3rd person >

Kin > Human > Animate > Inanimate

相关工作

- 等级差异
  - 在语法规则等级中，人类高于动物，而在人类评分中，动物不低于人类
  - 语法规则的等级更加细致，例如区分不同的人称、代词等
- 研究问题
  - 语用、认知和语言形式之间的关系
  - 分布式假说多大程度可以反映生命度等级?
  - 与两类等级的相关程度?

实验设定&方法

- 第一类目标词
  - 50个普通名词，分为三类
  - 人（例如艺术家等）、动物（例如猫）、物体（例如小山）
  - 从Wordnet中的相关类别中选择高频出现的词
  - 单复数同在
- 第二类目标词
  - 31个代词，包括第123人称的单复数代词以及它们不同的格
  - 例如主格I，宾格me，形容性物主代词my，名词性物主代词mine，反身代词myself
- 通过目标词选择Brown语料中的上下文
  - 每一个目标词选择了30条上下文

实验设定&方法

- 词向量模型
  - BERT GPT-2
- 生命度表征（语义映射方法）
  - 动物的平均表征减去物体的平均表征作为生命度表征
  - 待预测的表征与单位化后的生命度表征做点积

$$\mathbf{v}^l_{animal} = \frac{1}{N_{animal}} \sum_{i=1}^{N_{animal}} \mathbf{e}^l_{animal,i} \quad (3)$$

$$\mathbf{v}^l_{object} = \frac{1}{N_{object}} \sum_{i=1}^{N_{object}} \mathbf{e}^l_{object,i} \quad (4)$$

$$\mathbf{a}^l = \mathbf{v}^l_{animal} - \mathbf{v}^l_{object} \quad (5)$$

$$\hat{\mathbf{a}}^l = \frac{\mathbf{a}^l}{\|\mathbf{a}^l\|}$$

$$s^l_w = \mathbf{e}^l_w \cdot \hat{\mathbf{a}}^l$$

实验设定&方法

- 统计分析
  - 差异显著性
  - Kruskal-Wallis tests and performed post-hoc Dunn’s tests with Benjamini- Hochberg correction

实验1

- 目标类别：人、动物、物体
- 映射评分越低，生命度越高
- 模型体现的生命度等级：
  - 动物>人>物体
- 与人类评分一致
- 可以有效捕捉生命度

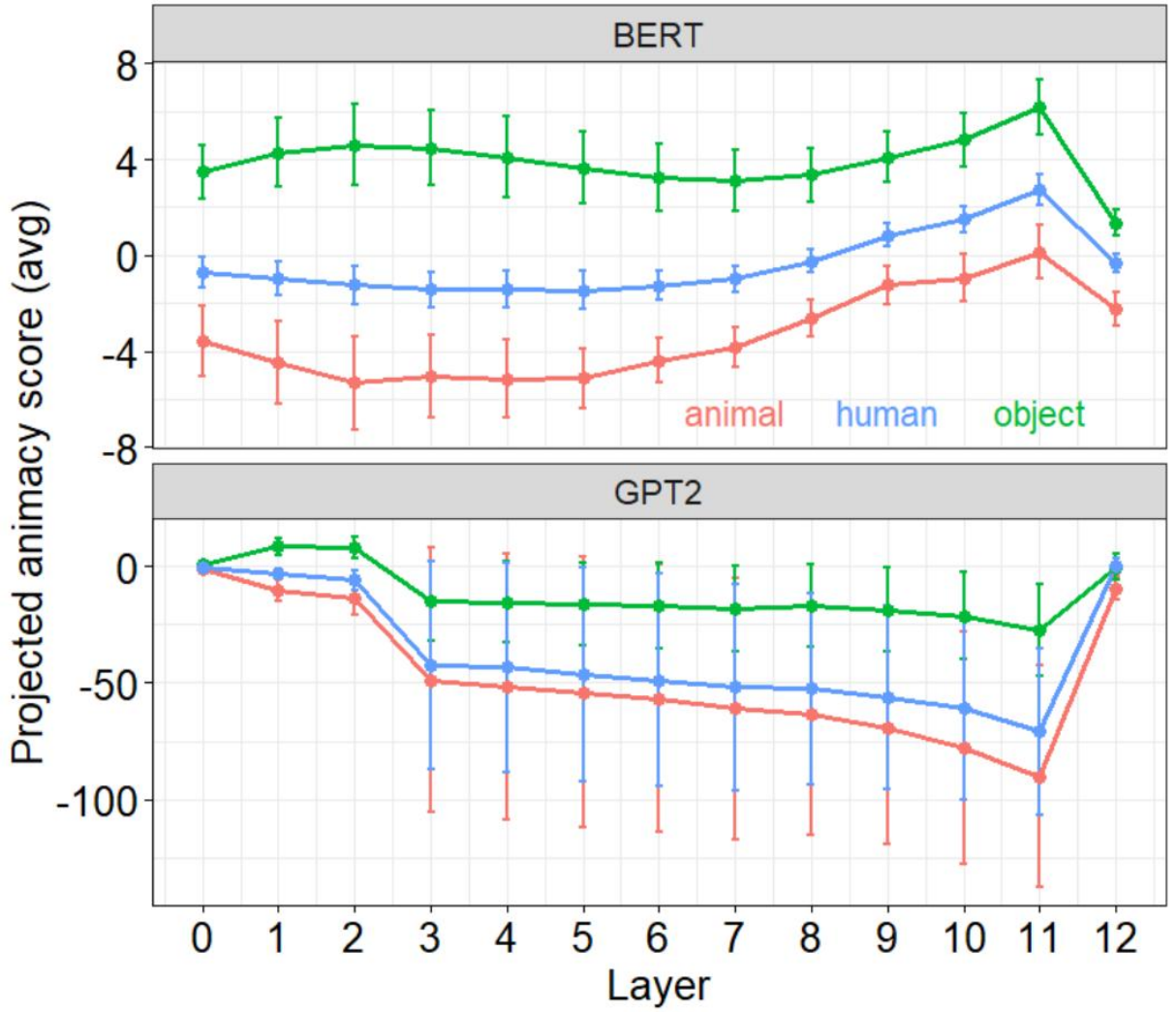


Figure 3: Line graph showing mean and standard deviation (error bars) of animacy scores of categories animals, humans and objects, averaged across words. Within each panel, the lower the score, the higher the animacy.

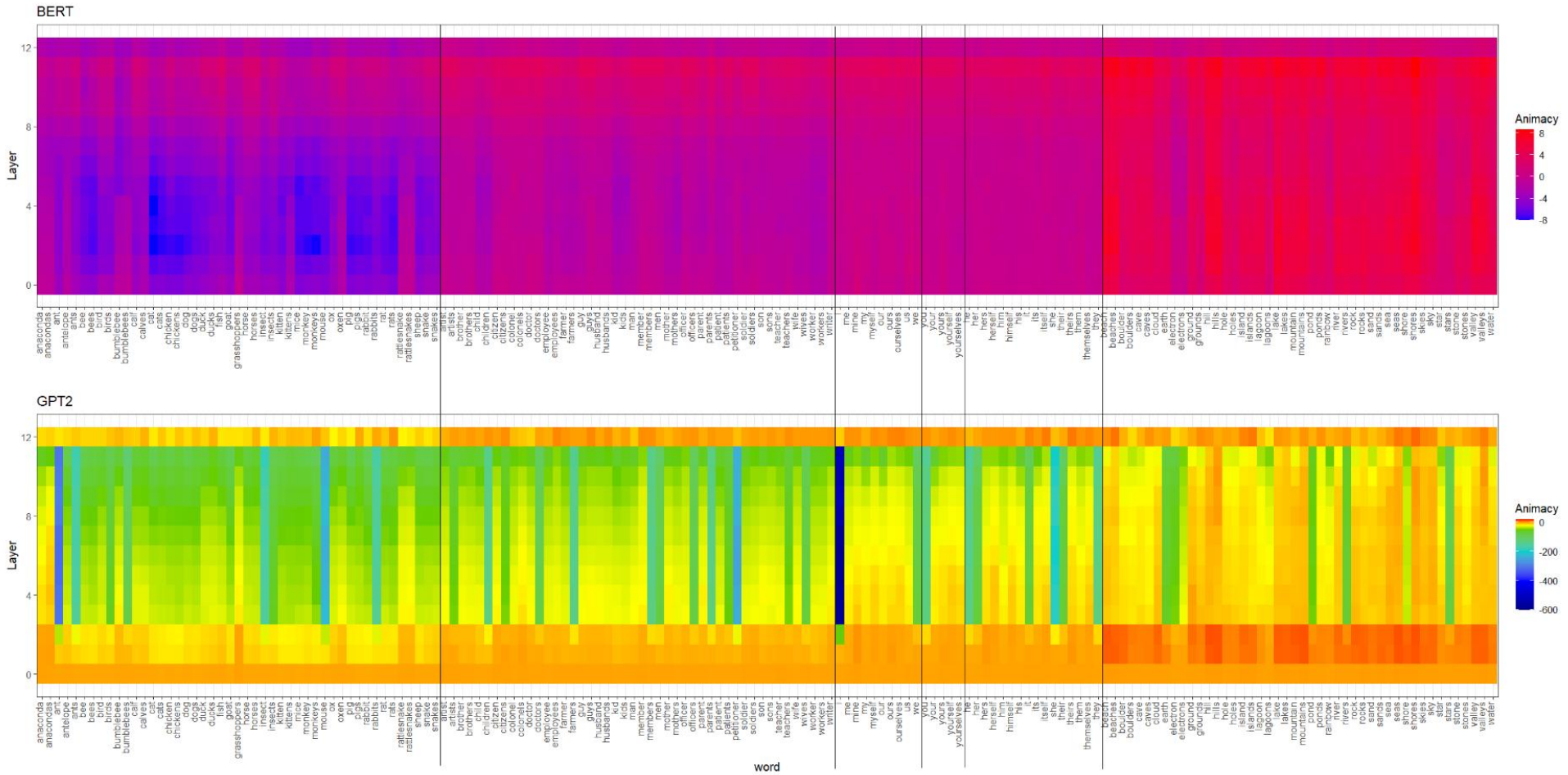


Figure 1: Heatmap showing animacy scores for all words used in the analyses across all layers in BERT (top) and GPT-2 (bottom). The vertical black lines in both panels demarcate different word categories, arranged from left to right: animals, humans, first-person pronouns, second-person pronouns, third-person pronouns, and objects. In both panels, the blue end of the color spectrum represents the highest level of animacy. The full list of words analyzed can be found in Appendix A.1.

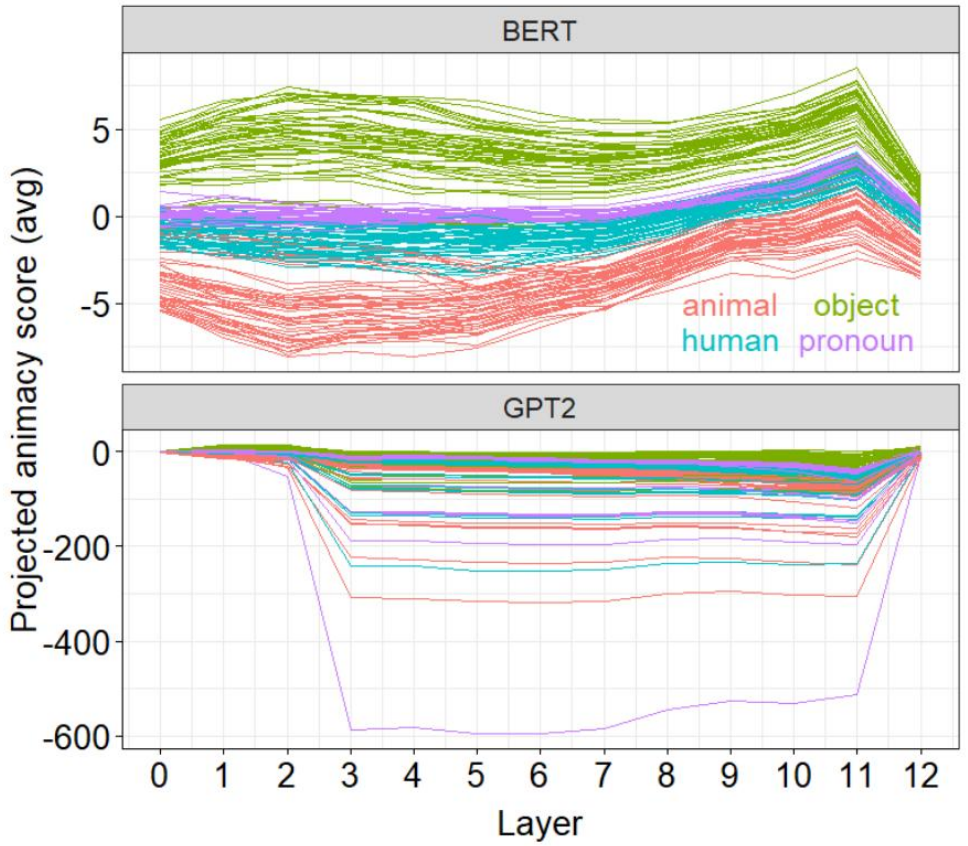


Figure 2: Line graph showing animacy scores of all target words used in the analyses across all layers in BERT (top) and GPT-2 (bottom). Categories (animals, humans, pronouns, objects) are color-coded. Within each panel, the lower the score, the higher the animacy.

实验2 代词 vs 人

- 代词的生命度显著低于人的
- 这些都与语法中的生命度等级有差异

实验3:人称代词之间

- 没有显著差异
- 有显著差异的是BERT的低层，但是第一人称的生命度弱于二、三人称

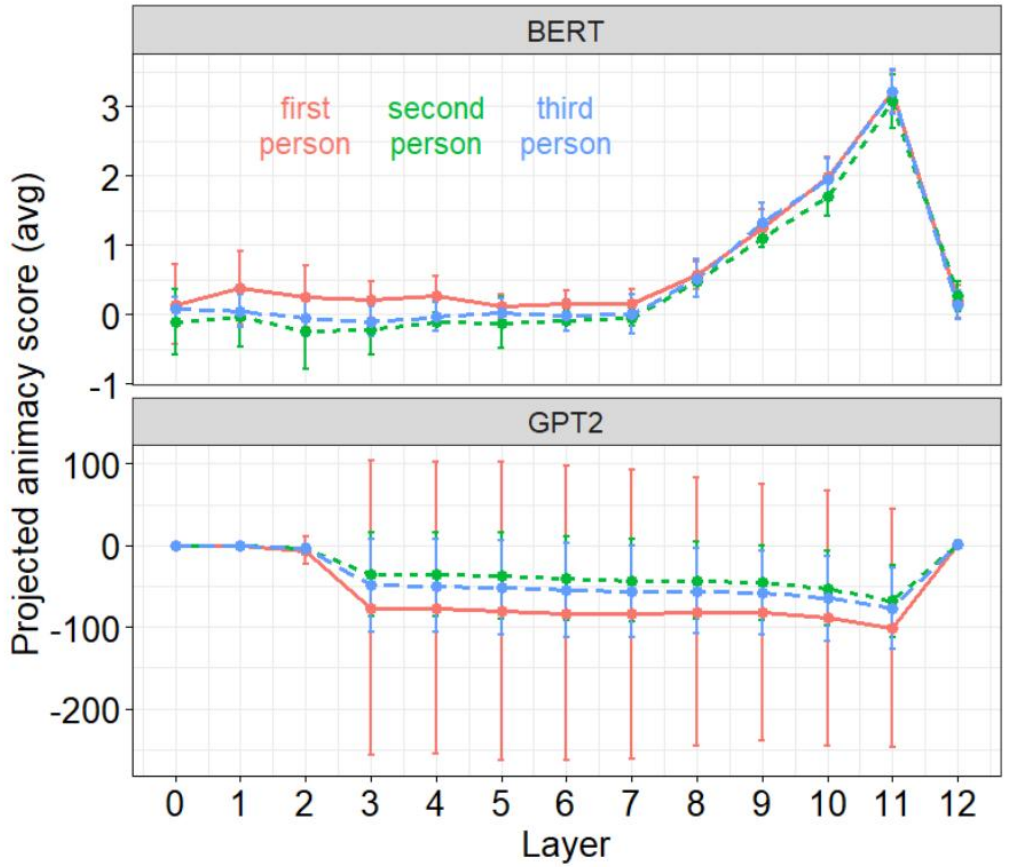


Figure 4: Line graph showing mean and standard deviation (error bars) of animacy scores of first-, second-, and third- person pronouns, averaged across words. Within each panel, the lower the score, the higher the animacy.

讨论

- 语义映射方法是有效的
- 基于词嵌入和语法系统的生命度具有差异
  - 人、动物
  - 代词之间
  - 语言使用和感知中的普遍趋势，并不是语法规则形成的唯一或直接来源；语言学习者是通过“选择性关注”某些语法结构（inductive bias）来建立不同的语法知识。

神经模型中捕捉到的语义知识

- 单向 vs. 双向
  - 单向模型也表现的很好，可能与任务比较简单有关
- 跨层差异
  - 低层编码词类信息，高层编码上下文信息
- 类内差异
  - For example, in both models, ant was more animate than other animal words, whereas rattlesnake(s) and grasshopper were comparatively less animate
  - GPT-2的差异更大，更容易受到格和数的影响，主格、单数要更强
  - 同一个词在不同的模型之间也有差异

结论

- 使用语义映射的方法来量化生命度特征
- 从三个维度比较了生命度序列
  - 人类感知
  - 模型（分布式）
  - 语法系统
- 分析了语法系统与模型的差异

参考

B.1 Seed and Target Words Used

Sentiment Classification

- **Task**: Classifying concepts based on sentiment by using the NRC corpus (Mohammad et al., 2013). Target words: 98 positive and 98 negative words. Seed words: "positive" and "negative".
- **Corpus**: wikitext-103-v1 from HuggingFace. We remove sentences that are shorter than 15 tokens and longer than 200 tokens.
- **Sampling**: We sample 1000 occurrences of "positive" and 1000 occurrences of "negative". For each target word, we sample 30 occurrences.

Grammatical Gender in French and Italian

Experiment 1:

- **Task**: Classifying concepts by the grammatical gender of nouns.
- **Corpus**: Latest Italian Wikipedia abstracts from DBPedia. We removed sentences shorter than 20 tokens and longer than 100 tokens.
- **Sampling**: Target words: 140 Italian nouns. Seed words: 59 Italian masculine and feminine adjectives. For each target word, we sample 30 occurrences. For each seed word, we sample 20 occurrences. Seed and target words have been filtered with respect to frequency. Data comes from Flex-IT (Pescuma et al., 2021).

Experiment 2:

- **Task**: Classifying concepts by the grammatical gender of nouns.
- **Corpus**: Latest French Wikipedia abstracts from DBPedia. We removed sentences shorter than 20 tokens and longer than 100 tokens.
- **Sampling**: Target words: 201 French nouns. Seed words: 65 French masculine and feminine adjectives. Seed and target words have been filtered with respect to frequency. Data comes form Lexique383 (New et al., 2004).

BERT Concept Net Classification Land-Sea

- **Task**: Classifying concepts by classes based on the ConceptNet dataset (Dalvi et al., 2022), predicting if an animal is a sea or land animal.
- **Corpus**: wikitext-103-v1 from HuggingFace. We remove sentences that are shorter than 15 tokens and longer than 200 tokens.
- **Sampling**: Target words: 64 land or sea animals. Seed words: category names: "land" and "sea". We sample 1000 occurrences of each seed word. For each target word, we sample 30 occurrences.

BERT Concept Net Classification Fashion-Gaming

- Task: Classifying concepts by classes based on the ConceptNet dataset (Dalvi et al., 2022), predicting if a concept comes from the fashion domain or the design domain.
- Corpus: wikitext-103-v1 from HuggingFace. We remove sentences that are shorter than 15 tokens and longer than 200 tokens.
- Sampling: Target words: 29 terms related to fashion or gaming. Seed words: category names: "fashion, clothes" and "gaming, games". We sample 500 occurrences of each seed word. For each target word, we sample 30 occurrences.