谁说发 paper 一定要追快打新？2021年，研究 word2vec 也能中顶会！

原创 jxyxiangyu 夕小瑶的卖萌屋 2021-10-10 17:00

文 | jxyxiangyu

前言

> “小夕，小夕，你关注的任务sota又被刷新了！”
> “什么？！”
> 还在跑实验的小夕默默流下了辛酸泪

不得不说nlp领域的发展真的太快了，炼丹师们不光要时刻关注前沿热点，还要快速做出实验，高强度堪比996：导师，臣妾真的做不到啊(っ╥╯﹏╰╥c)

正巧，小编我最近看到一篇研究词向量 word2vec 的论文，中了今年的EMNLP。What？！依稀记得头一次听说word2vec还在三年前。这么古老的东西还有人在研究吗？现在不都是XX-BERT、XX-transformer的时代了吗？

今天让我们一起来看看，到底是咋回事。

论文标题:
Analyzing the Surprising Variability in Word Embedding Stability Across Languages

论文链接:
https://arxiv.org/abs/2004.14876

词向量稳定性

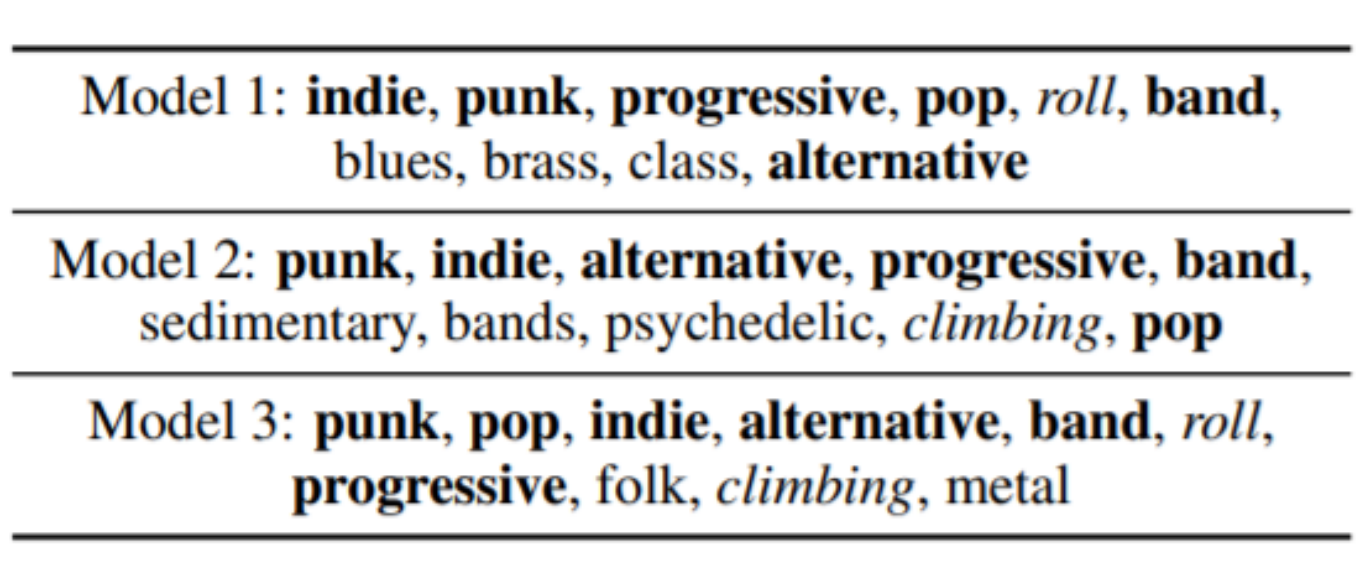
在介绍论文之前，先让我们来了解下词向量的稳定性。词向量的稳定性指的是一个词在不同的向量空间中的最近邻的重叠程度，常用来衡量由数据集、算法和词的属性特征的变化引起的词向量的变化。

这时候一定有小伙伴要问了，都1202年了，还有研究静态词向量的必要吗？No，no，no，如果这么想，格局就小了，我们常用的BERT、GPT这些模型都是建立在大规模语料上预训练得到的，如果面对的是**小语种**，没有像汉语、英语这么**丰富的语料库**，是很难喂饱预训练语言模型的，另外，为了某些小语种专门花费大量的资源训练预训练模型，从工业的角度来看，**成本**也是**非常高**的。这时，自然而然就会想到利用上下文无关的静态词向量来解决这类问题。

稳定性的计算

文章研究的是在多个不同的语言中的词向量稳定性的变化，首先提出了稳定性的计算方式。给定两个向量空间集合X，Y，那么两个向量空间的组合可以表示成(x,y)，其中，$x \in X$，$y \in Y$，对于某个词W在向量空间x，y的稳定性可以用W在这两个空间中最近的10个邻居的重叠百分比来表示，而在X和Y这两个集合中，任意两个向量空间的组合下的稳定性均值，就被定义为词W在这两个向量空间集合的稳定性。

举个栗子，下面的图展示的是词“**rock**”在三个向量空间下最近的10个邻居词，粗体表示向量空间重叠的词，可以看到$Model\ 1$和$Model\ 2$有6个邻居是重叠的，$Model\ 1$、$Model\ 3$和$Model\ 2$、$Model 3$分别有7个词重叠，那么词“**rock**”在这三个向量空间的稳定性就是这三个值的均值（0.667）。



Model 1: **indie**, **punk**, **progressive**, **pop**, *roll*, **band**, blues, brass, class, **alternative**

Model 2: **punk**, **indie**, **alternative**, **progressive**, **band**, sedimentary, bands, psychedelic, *climbing*, **pop**

Model 3: **punk**, **pop**, **indie**, **alternative**, **band**, *roll*, **progressive**, folk, *climbing*, metal

实验

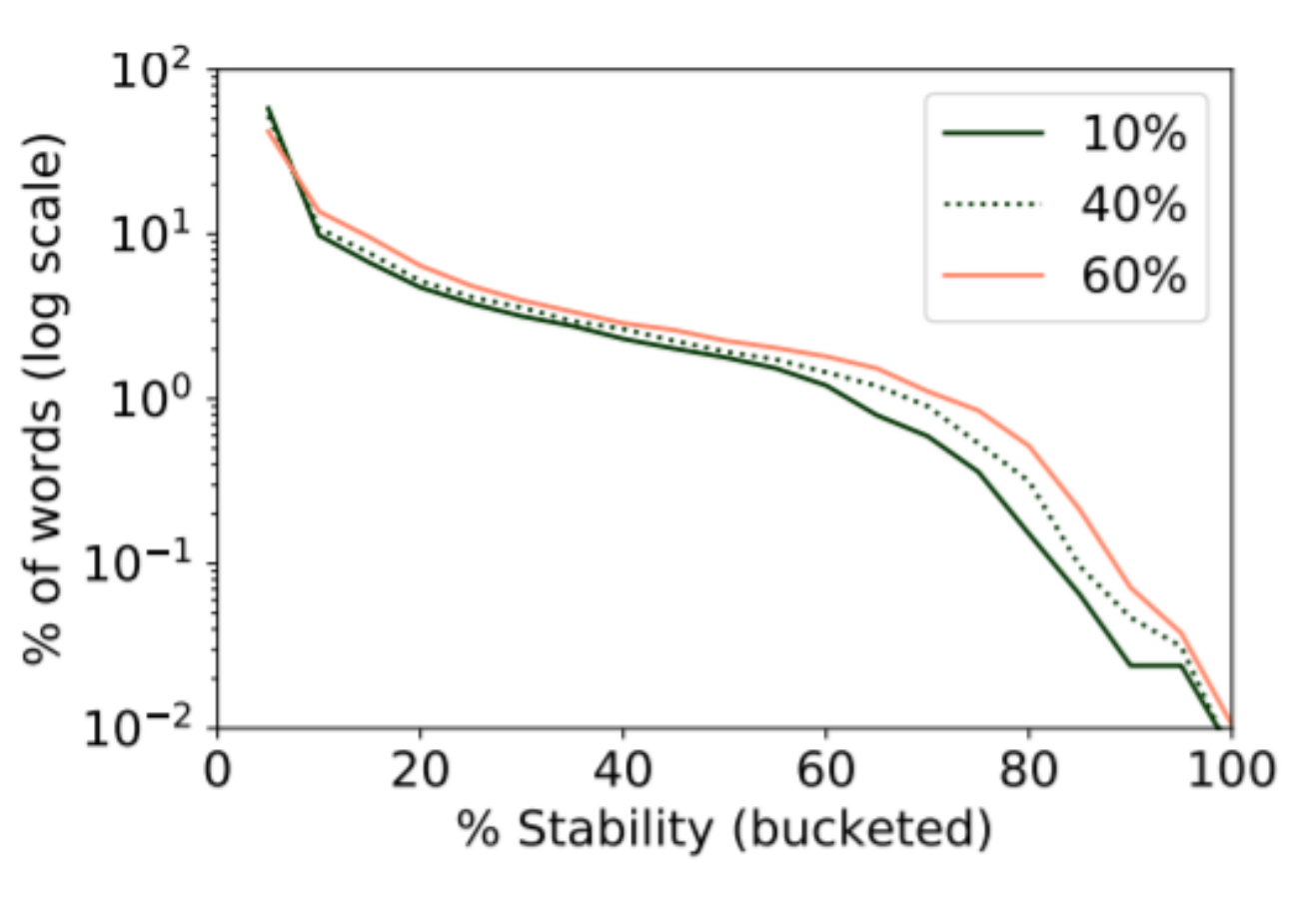
数据集

作者采用的是Wikipedia和Bible两个数据集，其中，Wikipedia包含40种语言，Bible包含97种语言，以及世界语言结构图谱（World Atlas of Language Structures，WALS），包含了近两千种语言属性知识。

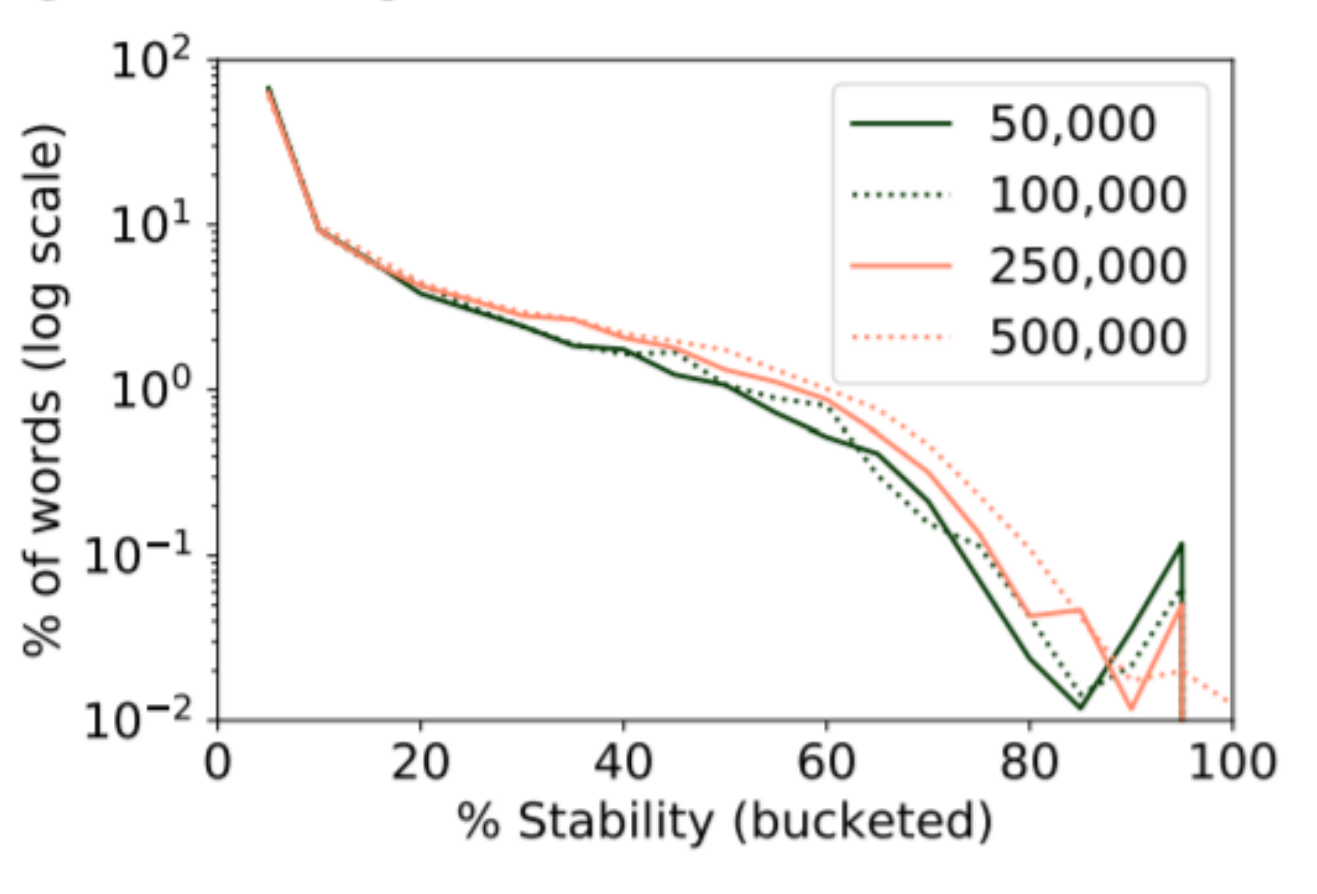
数据集下采样

为减小不同语言数据量对词向量稳定性的影响，论文对原始的数据集做了下采样处理，具体方法是对数据集**不重复地**下采样（downsampling without replacement）。

为研究不同的下采样方法对稳定性的影响，用作者的话来说，希望通过下采样得到跨语言且有可比性的稳定性结果。为此，作者专门对比了可重复采样和不可重复采样两种下采样方法对稳定性的影响。



(a) Sampling *with* replacement, varying percentage overlap between samples.



(b) Sampling *without* replacement, varying sample size.

可以看到可重复采样的方法导致下采样数据集无论采样覆盖比例如何，稳定性基本趋于一致，没有区分度，而不可重复采样的方法得到的数据集有较为明显的稳定性区分，这也是作者选择不重复下采样方法的原因。

数据集上的稳定性

作者针对Wikipedia和Bible两个数据集重叠的26种语言，研究了**不同语言**，**不同词向量生成算法**和**数据**对词向量稳定性的影响，总共三种情况：

1. 由五个下采样的数据集训练得到的GloVe词向量的稳定性
2. 由五个下采样的数据集训练得到的word2ve词向量的稳定性
3. 由一个下采样的数据集随机五次训练得到的word2ve词向量的稳定性

由于Bible数据集过小，因此，只对Bible数据集研究了情况3下稳定性的分布

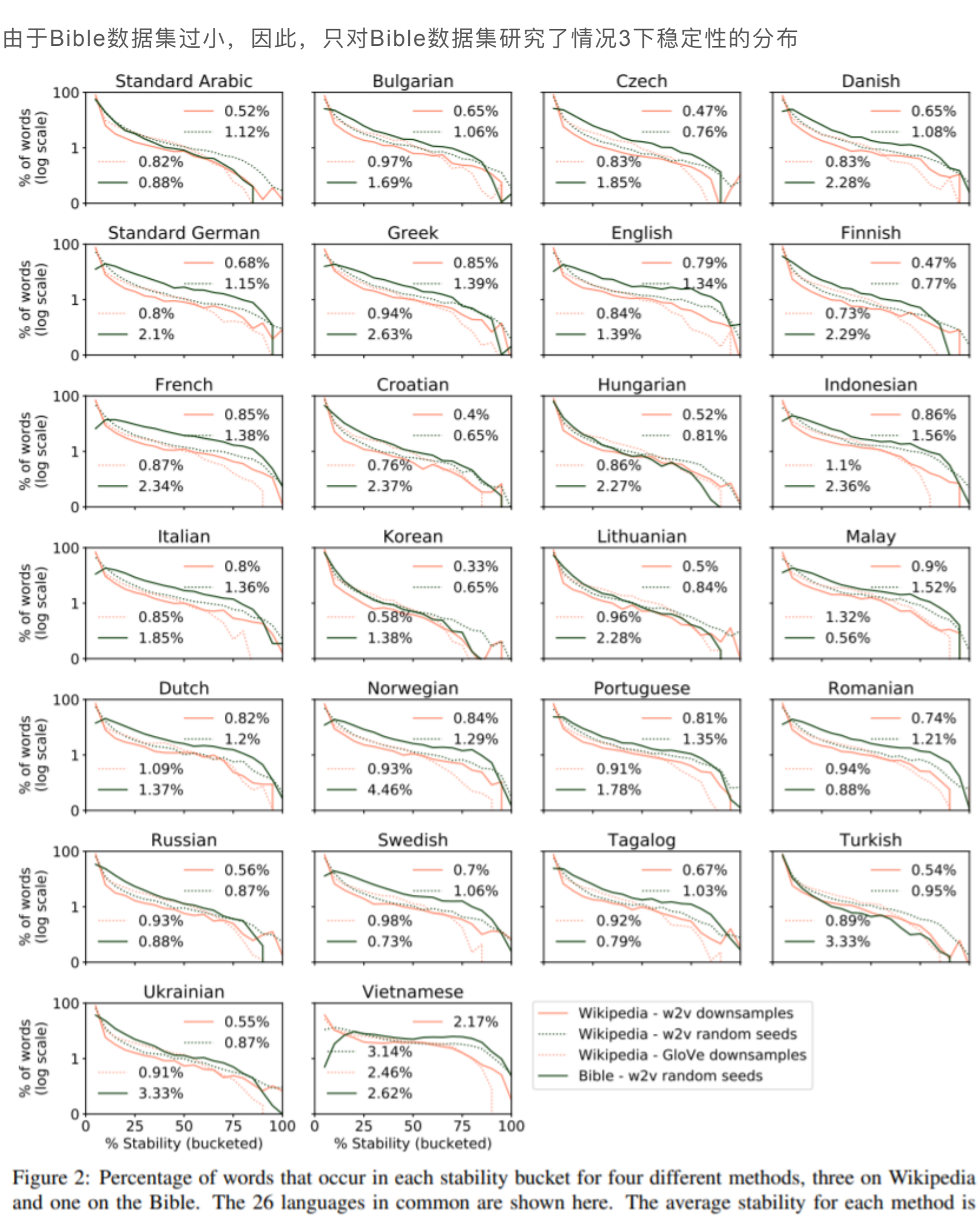
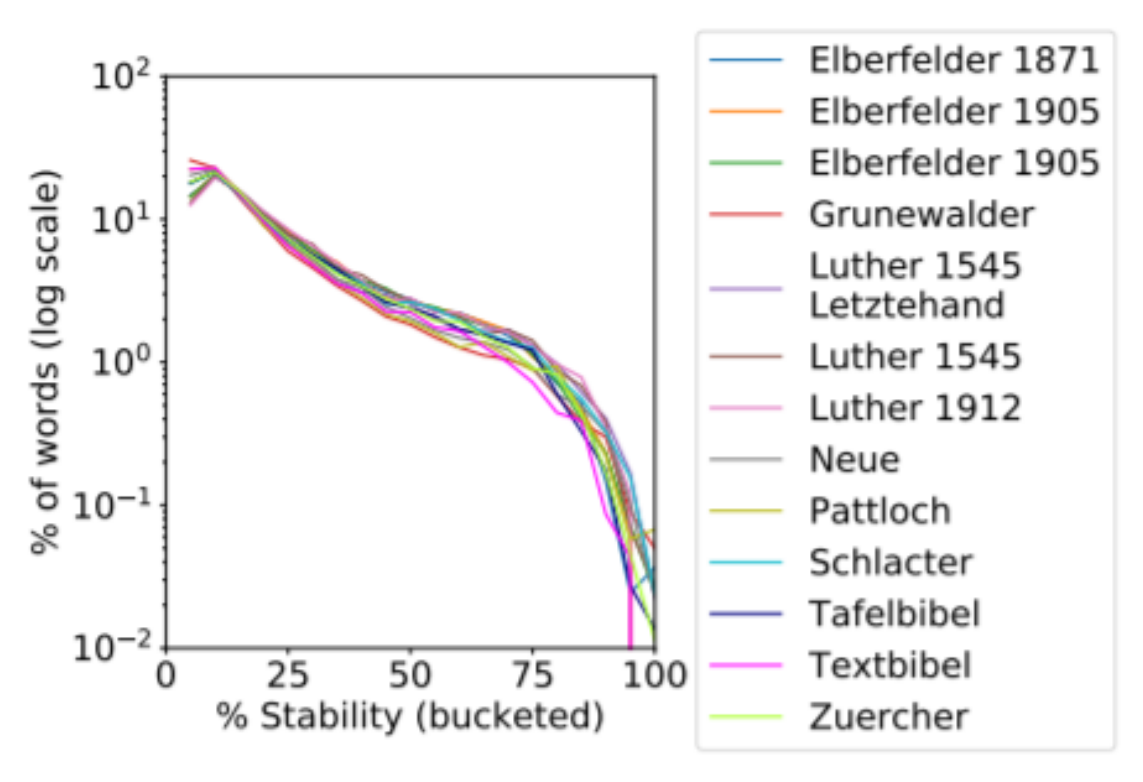


Figure 2: Percentage of words that occur in each stability bucket for four different methods, three on Wikipedia and one on the Bible. The 26 languages in common are shown here. The average stability for each method is shown on the individual graphs.

可以看到在稳定性25%~75%之间，稳定性分布和变化较为平缓，低稳定性和高稳定性的词数量变化明显。

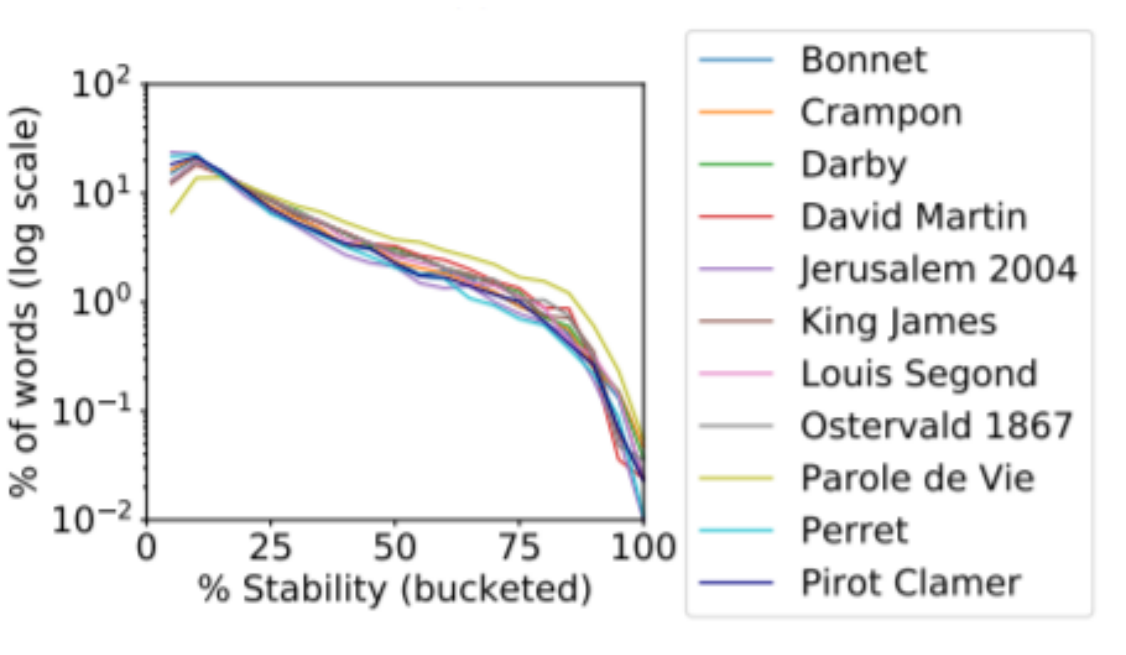
实验结果也表明在相同的训练数据下，不同的训练算法得到的词向量稳定性分布和变化趋于一致，相比而言，训练语料的不同对稳定性有较大的影响。因此，在对比不同语言下的词向量稳定性时，应该**减小语料的内容**对稳定性的影响。

为了研究**同一数据集的不同上下文**对研究不同语言间稳定性的影响程度，作者分别选择了圣经在德语和法语的多个不同译本，在一个下采样数据集上用五个不同的随机数种子训练生成五个word2vec词向量，并取均值作为该译本下的词向量稳定性。可以看到除个别译文外，不同译本之间的稳定性趋势基本一致，基本可以**忽略***相同语义的不同上下文*对词向量稳定性的影响。



(a) German

▲不同bible译本的稳定性分布（德语）



(b) French

▲不同bible译本的稳定性分布（法语）

回归模型

前述的实验对比了多个语言下的稳定性分布与走势，下面作者用岭回预测特定语言下的所有词的平均稳定性的方式，研究**语言属性本身**对词向量稳定性的影响因素。

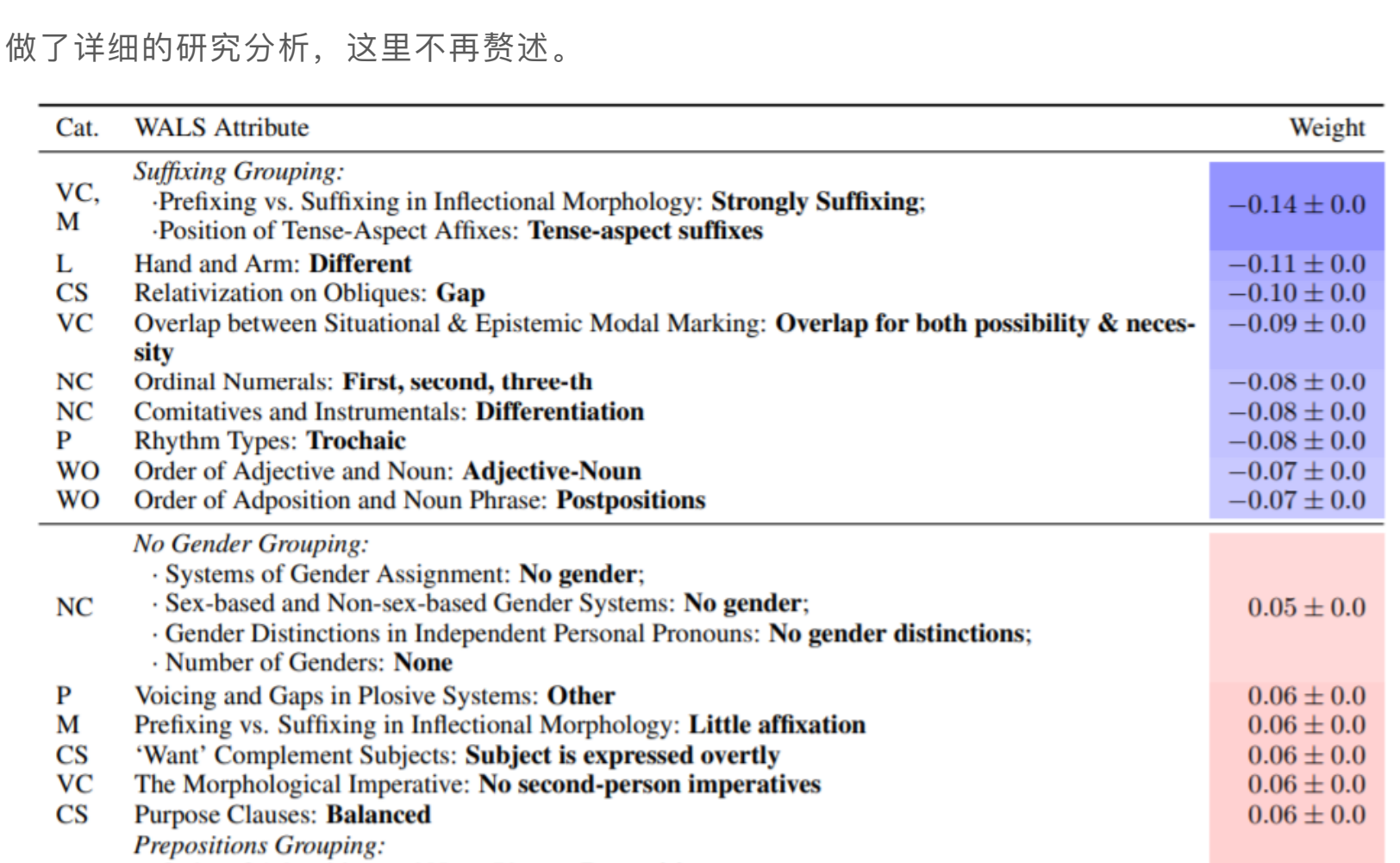
模型的输入是特定语言的语言学特征（属性），输出是稳定性的均值。在讲特征输入模型前，作者做了相应的数据预处理，包括过滤出现频次较低的特征和属性（WALS）以及属性较少的语言，特征分组等，这里就不详细说明了。

评价指标

作者用了两种方式来评估模型：**R^2和留一法交叉验证的绝对误差**。选择拟合效果较好的模型，通过**权重的大小**来确定特征（或属性）对稳定性的贡献度程度。

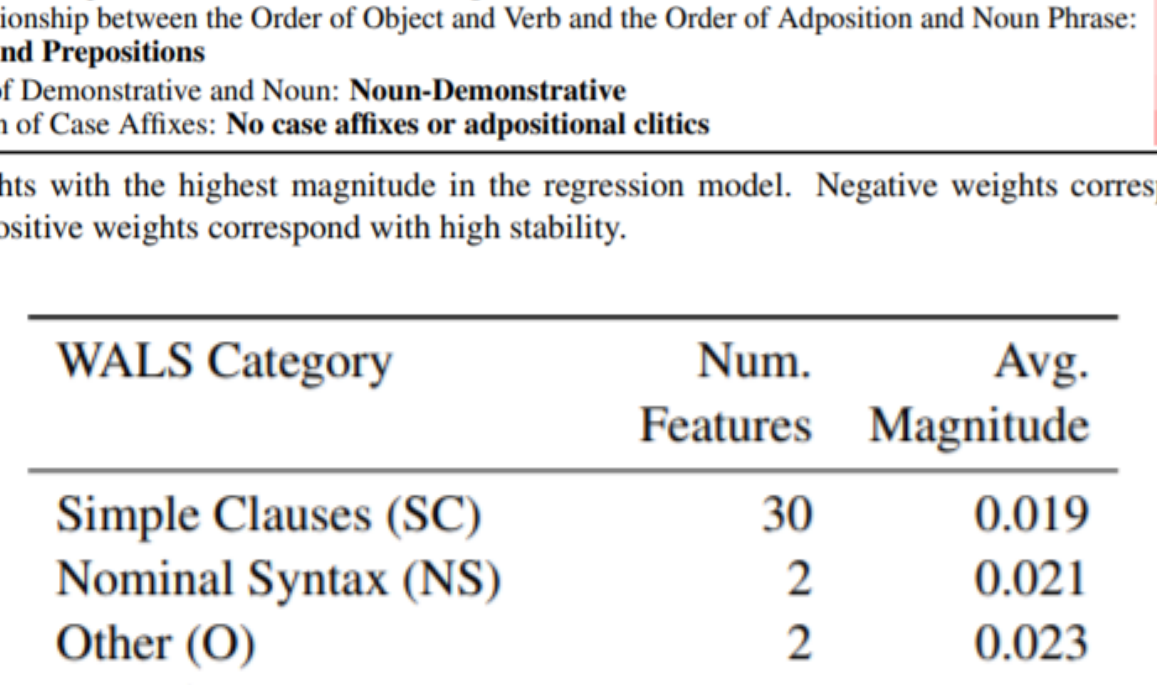
实验结论

作者选择的模型达到了0.96的R^2和0.86 ± 0.55的留一法交叉验证的绝对误差，足以证明模型拟合效果非常好，相应的权重也可以表示属性对稳定性的贡献程度。下面是岭回归模型拟合后得到的属性对稳定性的贡献度权重和对特征分组的平均权重。相应地，作者还对某些属性特征做了详细的研究分析，这里不再赘述。



Cat.	WALS Attribute	Weight
VC, M	*Suffixing Grouping:* ·Prefixing vs. Suffixing in Inflectional Morphology: **Strongly Suffixing**; ·Position of Tense-Aspect Affixes: **Tense-aspect suffixes**	-0.14 ± 0.0
L	Hand and Arm: **Different**	-0.11 ± 0.0
CS	Relativization on Obliques: **Gap**	-0.10 ± 0.0
VC	Overlap between Situational & Epistemic Modal Marking: **Overlap for both possibility & necessity**	-0.09 ± 0.0
NC	Ordinal Numerals: **First, second, three-th**	-0.08 ± 0.0
NC	Comitatives and Instrumentals: **Differentiation**	-0.08 ± 0.0
P	Rhythm Types: **Trochaic**	-0.08 ± 0.0
WO	Order of Adjective and Noun: **Adjective-Noun**	-0.07 ± 0.0
WO	Order of Adposition and Noun Phrase: **Postpositions**	-0.07 ± 0.0
NC	*No Gender Grouping:* · Systems of Gender Assignment: **No gender**; · Sex-based and Non-sex-based Gender Systems: **No gender**; · Gender Distinctions in Independent Personal Pronouns: **No gender distinctions**; · Number of Genders: **None**	0.05 ± 0.0
P	Voicing and Gaps in Plosive Systems: **Other**	0.06 ± 0.0
M	Prefixing vs. Suffixing in Inflectional Morphology: **Little affixation**	0.06 ± 0.0
CS	‘Want’ Complement Subjects: **Subject is expressed overtly**	0.06 ± 0.0
VC	The Morphological Imperative: **No second-person imperatives**	0.06 ± 0.0
CS	Purpose Clauses: **Balanced**	0.06 ± 0.0
WO	*Prepositions Grouping:* ·Order of Adposition and Noun Phrase: **Prepositions**; ·Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase: **VO and Prepositions**	0.06 ± 0.0
WO	Order of Demonstrative and Noun: **Noun-Demonstrative**	0.07 ± 0.0
NC	Position of Case Affixes: **No case affixes or adpositional clitics**	0.11 ± 0.0

Table 3: Weights with the highest magnitude in the regression model. Negative weights correspond with low stability, and positive weights correspond with high stability.



WALS Category	Num. Features	Avg. Magnitude
Simple Clauses (SC)	30	0.019
Nominal Syntax (NS)	2	0.021
Other (O)	2	0.023
Complex Sentences (CS)	11	0.028
Morphology (M)	18	0.031
Word Order (WO)	32	0.031
Phonology (P)	21	0.032
Nominal Categories (NC)	40	0.036
Verbal Categories (VC)	27	0.036
Lexicon (L)	6	0.039

Table 4: Number of binary features and average magnitude of weights in the regression model for different WALS categories. Grouped features are included in each category that they cover.

小结

与常见的在某个任务上提模型、刷sota不同，这篇论文着眼于词向量在不同语言之间的差异的研究，本质上更像是数据分析。文章从数据采样方式入手，分别研究了数据集、训练算法对不同语言的稳定性的分布和走势的影响，并使用岭回归模型拟合了语言的属性特征对稳定性的贡献程度，分析不同属性特征对稳定性的影响。相比提出一个新的模型刷sota而言，可复现性和解释性更高，对词向量的应用有不小的贡献。

当然，这篇文章研究的是经典的静态词向量，和主流的transformer架构相比，确实显得有点“out”，但文章**投了七次才中**，不也证明了只要是金子都会发光吗？小编认为，谁说nlp一定要追快打新，只要是真正有益于nlp领域发展的研究工作，都值得发表，都值得中。（无卡党和少卡党狂喜~~bushi~~）

▲狂喜