MIT&谷歌大脑用 AI 破解失传的古代文字，被称“现代版罗塞塔石碑” | ACL 2019

人工智能7 月 9 日

漫漫尘埃下，掩藏了许多曾经辉煌灿烂古代文明，但我们现在却无法清晰地知道，这些地方究竟发生了什么。

搞懂这些历史的最佳方式，就是找到他们的文字记载。However，记载文字的石碑可以被考古学家们挖出来，但这些古文字究竟啥意思，现代的人们看不懂，需要语言学家们耗尽青春来推测。

现在，MIT CSAIL 和谷歌大脑的研究者出手了，他们用机器学习破译了乌加里特文和线性文字 B。

▵ 乌加里特王宫

乌加里特文，Ugaritic，是一种楔形文字，属于闪米特语族。从字面上来看，就知道它是一个叫做乌加里特（Ugarit）的文明使用的语言，这个文明位于当今地中海沿岸的叙利亚，在公元前 6000 年前后就初现踪迹，在公元前 1190 年前后灭亡。

▵ 乌加里特文

线性文字 B，Linear B，由一种人类还没有破译出来的线性文字 A 演化而来，主要存活于公元前 1500 年到公元前 1200 年的克里特岛和希腊南部 ,是希腊语的一种古代书写形式。

△ 线性文字 B

研究者们利用同一语族内不同语言之间的联系，用机器学习的方法来破译这两种失传的语言，这是破译古代语言的新方法，也将对罗曼语族的语言学研究有巨大的影响和提升。

这个方法让许多人惊叹：

简直是现代版的罗塞塔石碑！

PS，罗塞塔石碑是一块用 3 种语言写了同一个内容的石碑，帮助语言学家们读懂古文字。

希望能先把动物和植物的语言破译了，可以发现打开新世界的大门。

人类语言总相通

这项研究的核心方法，是借助人类语言的相似性。

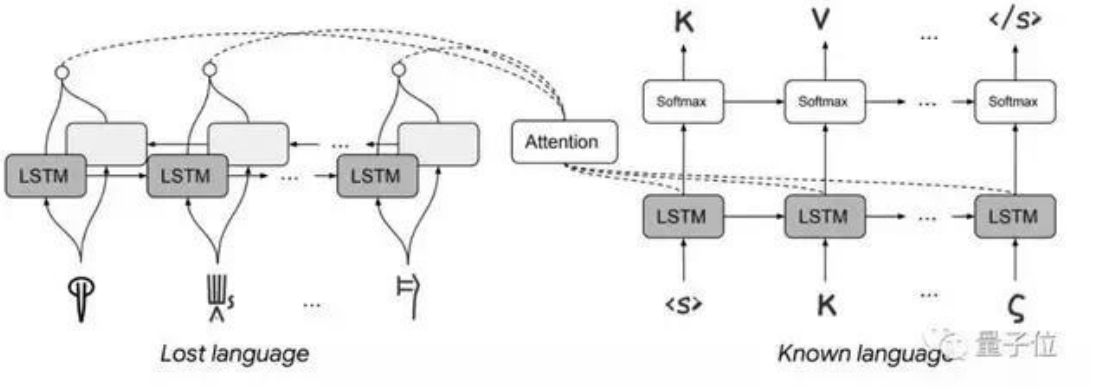
比如，知乎用户@拉队短 在介绍欧洲语言相似性的时候，举了这么个栗子：

句子“那是六月末潮湿阴沉的一个夏日。”
英语：It was a humid, grey summer day at the end of June.
丹麦语：Det var en fugtig, grå sommerdag i slutningen af juni.
瑞典语：Det var en fuktig, grå sommardag i slutet av juni.
挪威语：Det var en fuktig, grå sommerdag i slutten av juni.
冰岛语：Það var rakur, grár sumardagur í lok júní.

看，长得差不多嘛，毕竟同属印欧语系日耳曼语族，单词的分布位置、句子的结构都很相似，如果你能看懂一种语言，就能大致猜测和它“血缘”关系近的另一种语言。

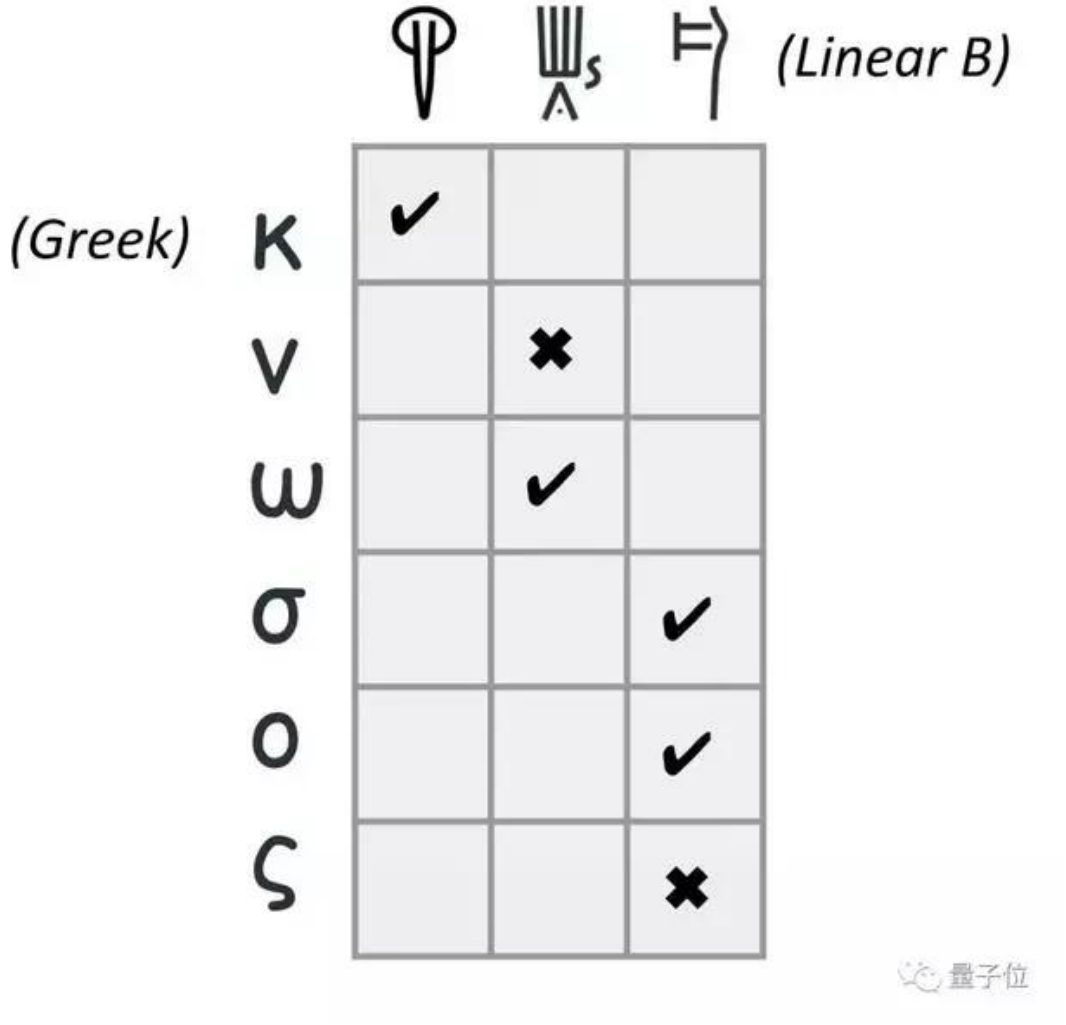
模型训练

为了破解这两种文字，研究者们提出了一个基于字符的 seq2seq 模型。



模型主要包含通用字符嵌入、剩余连接、单调排列正则化几个部分。

其中，线性文字 B 的字母和希腊文需要进行对应。



之后，借助神经解密算法，在具有不同语言特征的多种语言中提供强大的性能。

Algorithm 1 Iterative training

Require:

$\mathcal{X}, \mathcal{Y}$: vocabularies,

T: number of iterations,

N: number of cognate pairs to identify.

$f_{i,j}^{(0)} \leftarrow \frac{N}{|\mathcal{X}| \cdot |\mathcal{Y}|}$ ▷ Initialize

for $\tau \leftarrow 1$ to T **do**

  $\theta^{(\tau)} \leftarrow \text{MLE-TRAIN}(f_{i,j}^{(\tau-1)})$

  $\bar{d}_{i,j}^{(\tau)} \leftarrow \text{EDIT-DIST}(x_i, y_j, \theta^{(\tau)})$

  $\tilde{f}_{i,j}^{(\tau)} \leftarrow \text{MIN-COST-FLOW}(\bar{d}_{i,j}^{(\tau)})$

  $f_{i,j}^{(\tau)} \leftarrow \gamma \cdot f_{i,j}^{(\tau-1)} + (1-\gamma) \cdot \tilde{f}_{i,j}^{(\tau)}$

  $\text{RESET}(\theta^{(\tau)})$

return $f_{i,j}^{(T)}$

function MLE-TRAIN($f_{i,j}^{(\tau)}$)

  $\theta^{(\tau)} \leftarrow \arg\max_\theta \prod_{y_j \in \mathcal{Y}} \Pr_\theta(y_j | \mathcal{X}, \mathcal{F})$

  return $\theta^{(\tau)}$

你懂的语言，和你不懂的语言

在算法模型的基础之下，需要的语料库除了待破解的乌加里特文和线性文字 B，还需要一些现在的人类能看懂的语言。

研究团队选择了罗曼语族的数据库，包含意大利语、西班牙语和葡萄牙语三种语言的同源语音转录，需要对它们进行同源检测。

Dataset	#Cognates	#Tokens (lost/known)	#Symbols (lost/known)
UGARITIC	2214	7353/41263	30/23
Linear B	919	919/919	70/28
Linear B/names	455	919/455	70/28
ROMANCE	583	583/583	25/31/28 (E[illegible]/量子位

因此，数据集就用到上面这些，Symbols 指的是语言中的字符，Token 则是语言学中类似于单词的存在。

准确率

运行成果还不错，乌加里特文在无噪声条件下优于现有方法 3.1%，在有噪声条件下优于现在的贝叶斯方法 5.5%。

System	Noiseless	Noisy
Matcher	90.4	-
Bayesian	-	60.4
NeuroCipher	**93.5**	**65.9**

量子位

而线性文字 B，在无噪声条件下准确率高达 84.7%，在更具挑战性的 LinearB 名称数据集中达到 67.3%的准确度。

System	Linear B	Linear B/names
NeuroCipher	84.7	67.3

量子位

在罗曼语族同源识别任务中，西班牙语准确度提升 3.4%，葡萄牙语提升 1.6%。

System	EsIt	EsPt	ItPt	Avg
Matcher	88.9	**95.6**	85.7	90.1
NeuroCipher	**92.3**	95.0	**87.3**	**91.6**

线性文字 B 的祖先，线性文字 A 还没有被人类破译，它被誉为考古界圣杯。

未来，在这项研究起作用的情况下，或许可以像借助罗曼语族三种语言的数据库一样，直接用机器借助其他已知的人类语言，实现暴力破解。

想破脑壳的语言学家们，可以把工作重心放到别的事情上了。

作者介绍

这项研究的一作 Jiaming Luo，正在 MIT CSAIL 读博，专注 NLP 研究，此前他也曾在北大从事情绪分析方面的研究。

Luo 同学的导师，也是这项研究的第三位作者 Regina Barzilay，她是 MIT CSAIL 的教授，2017 曾因 NLP 方面的研究获得麦克阿瑟奖金，除了 NLP 之外，她还研究深度学习在化学和肿瘤学方面的应用。

传送门

论文：

Neural Decipherment via Minimum-Cost Flow: from Ugaritic to Linear B

Jiaming Luo, Yuan Cao, Regina Barzilay

https://arxiv.org/abs/1906.06718

代码及数据集：

https://github.com/j-luo93/NeuroDecipher

- END -