# 📚当前世界上最好的通用单词嵌入和句子嵌入

这篇文章的中文版本可以在这里找到，多亏了Jakukyo。

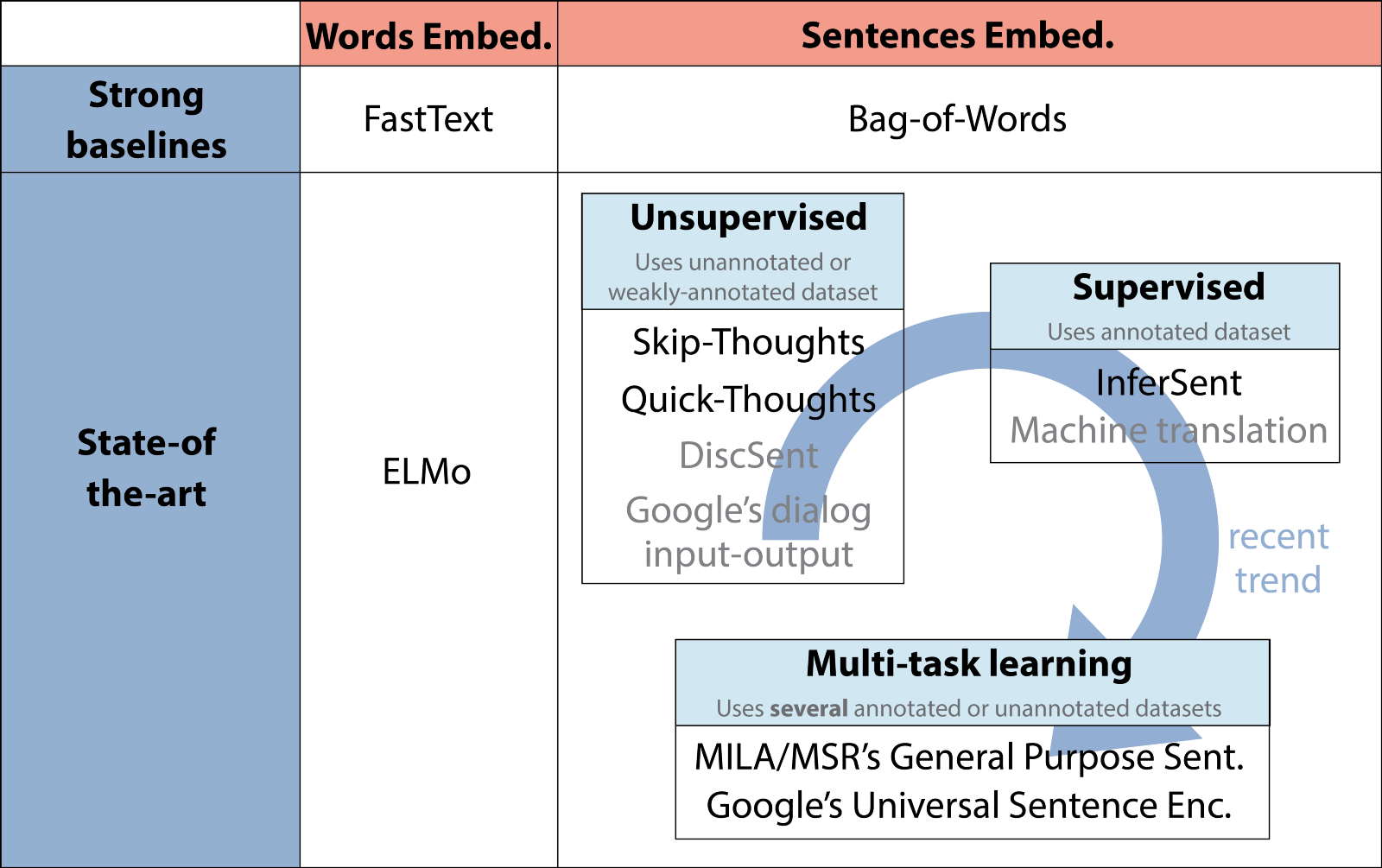
单词和句子嵌入已经成为任何基于深度学习的自然语言处理系统的重要组成部分。

他们用固定长度的密集向量对单词和句子进行编码，以大大改进文本数据的处理。

一个巨大的趋势是寻求通用的嵌入：嵌入是在一个大的语料库上预先训练的，可以插入到各种下游的任务模型（情感分析、分类、翻译……）中，通过合并在更大的数据集上学习到的一些通用单词/句子表示来自动提高它们的性能。

这是转移学习的一种形式。近年来，转移学习被证明能显著提高NLP模型在文本分类等重要任务上的性能。去检查（ULMFiT）看看它的运行情况。

虽然在相当长的一段时间里，句子的无监督表征学习一直是一种常态，但在过去的几个月里，我们看到了向有监督和多任务学习模式的转变，并在2017年底/2018年初提出了一些非常有趣的建议。



因此，这篇文章是一篇简单的入门文章，介绍了通用单词和句子嵌入的最新技术，并详细介绍了一些

* 强/快基线：FastText、Bag of Words
* 最先进的模型：ELMo，Skip-thinks，Quick-thinks，InferSent，MILA/MSR的通用句子表示和Google的通用句子编码器。

如果你想了解2017年之前发生的事情，我推荐塞巴斯蒂安去年写的。

让我们从单词嵌入开始。

### 词嵌入的最新进展

在过去的五年里，人们提出了许多可能的嵌入单词的方法最常用的模型是和，它们都是基于的无监督方法（出现在相同上下文中的单词往往具有相似的含义）。

虽然通过整合语义或句法知识的监督来增强这些无监督方法，但在2017-2018年间，纯无监督方法出现了有趣的发展，最引人注目的是FastText（word2vec的扩展）和ELMo（最先进的上下文词向量）。

由Tomas Mikolov团队开发，他在2013年提出了word2vec框架，引发了对通用单词嵌入的研究热潮。

与原始word2vec向量相比，FastText的主要改进是包含了字符n-grams，它允许计算未出现在训练数据中的单词的单词表示（“词汇表外”单词）。

训练速度非常快，有157种语言可在维基百科和Crawl上学习。他们是一个伟大的底线。

他（ELMo）最近将单词嵌入的技术水平提高了一个显著的水平。它们由艾伦人工智能研究所（Allen institute for AI）开发，将于2018年6月初面世。



在ELMo中，每个词都被赋予了一个表示，它是它们所属的整个语料库句子的函数。嵌入是从两层双向语言模型（LM）的内部状态计算出来的，因此命名为“ELMo”：来自语言模型的嵌入。

ELMo的特殊性：

* ELMo的输入是字符而不是单词。因此，他们可以利用子词单位来计算有意义的表示，即使是词汇表外的词（如FastText）。
* ELMo是biLMs的几层激活的串联，语言模型的不同层编码一个词的不同类型的信息（例如，biLSTM的低层很好地预测了词性标注，而词义消歧则更好地编码在高层）。连接所有层允许自由组合各种单词表示，以便在下游任务中获得更好的性能。

现在，让我们转到通用语句嵌入。

### 普遍句嵌入的兴起



目前有许多学习句子嵌入的竞争方案。虽然简单的基线（如平均单词嵌入）始终会产生很强的效果，但在2017年底至2018年初，出现了一些新的无监督和监督方法以及多任务学习方案，并带来了有趣的改进。

让我们快速浏览一下目前研究的四种方法：从简单的词向量平均基线到无监督/有监督的方法和多任务学习方案（如上图所示）。

在这个领域中，直接提取句子的词向量的简单方法（所谓的词包方法）为许多下游任务提供了一个强大的基线。

一个计算这样一个基线的好算法在去年发表在ICLR上的Arora等人的工作中有详细介绍，这是一个简单但难以超越的句子嵌入基线：使用你选择的流行的单词嵌入，将一个句子编码成一个线性加权组合词向量并执行一个公共成分去除（去除向量在其第一主成分上的投影）。这种普遍的方法有着更深刻和强大的理论动机，它依赖于一个生成模型，该模型使用在话语向量上的随机游走来生成文本（这里我们不讨论理论细节）。

最近一个强大的单词基线包（甚至比Arora的强）的实现是来自Darmstadt大学的串联p-均值嵌入，您会发现（感谢指出这一点）。



除了简单的平均数，第一个主要的提议是使用无监督的培训目标，从杰米·基罗斯和同事在2015年提出的提议开始。

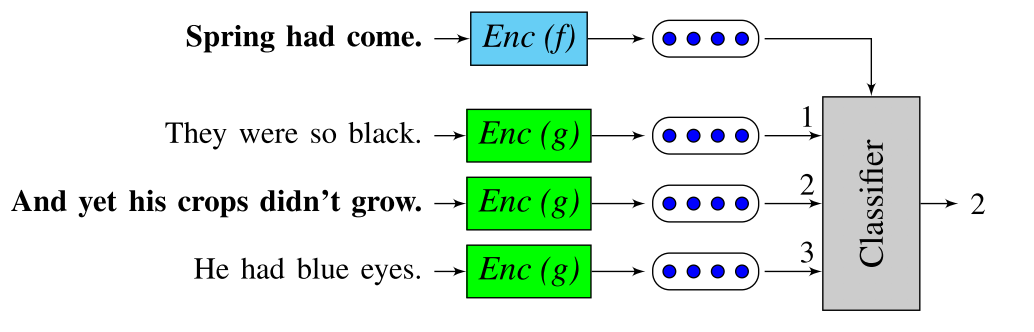
无监督方案学习句子嵌入，作为学习预测句子中连贯的句子序列或连贯的从句序列的副产品。这些方法（理论上）可以利用任何文本数据集，只要它包含以连贯方式并列的句子/从句。

是学习无监督句子嵌入的典型例子。它可以被看作是为单词嵌入而开发的skip-gram模型中句子的等价物：我们试图预测给定句子的周围句子，而不是预测单词周围的单词。该模型由一个基于RNN的编码器-解码器组成，该解码器通过训练从当前句子中重建出周围的句子。

Skip思想论文中一个有趣的见解是词汇扩展方案：Kiros等人通过学习RNN单词嵌入空间和更大的单词嵌入（如word2vec）之间的线性转换来处理在训练期间看不到的单词。

\*\*\*\*是跳跃思维向量的最新发展，今年在ICLR上发表。在这项工作中，预测下一个句子的任务被重新定义为一个分类任务：解码器被一个分类器取代，这个分类器必须在一组候选者中选择下一个句子。它可以被解释为对生成问题的判别式近似。

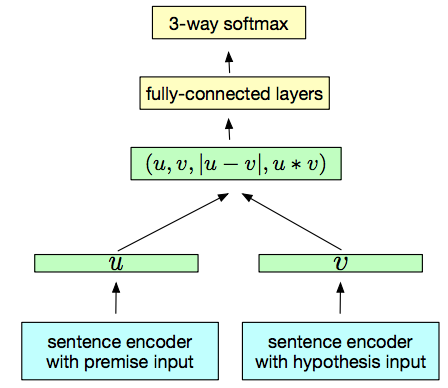
该模型的一个优点是它的训练速度（与跳过思想模型相比是一个数量级），使得它成为开发大量数据集的竞争解决方案。



或者很长一段时间以来，有监督的句子嵌入学习被认为比无监督的方法具有更低的嵌入质量，但这一假设最近被推翻，部分是在结果公布之后。

与之前详细介绍的无监督学习方法不同，有监督学习需要为某些任务（如自然语言推理（例如，包含一对句子）或机器翻译（包含一对翻译句子）添加注释的标记数据集，这些任务提出了要选择的特定任务和相关的大小问题高质量嵌入所需的数据集。我们将在下一节和最后一节中讨论更多关于多任务学习的问题，但在此之前，让我们看看2017年出版的《推理突破》背后的原因。

\*\*\*\*是一个有趣的方法，因为它的架构简单。它使用斯坦福自然语言推理（SNLI）语料库（一组570k对句子，用中性、矛盾和蕴涵三个类别标记）在句子编码器上训练分类器。两个句子使用相同的编码器进行编码，而分类器则使用由两个句子嵌入构成的成对表示进行训练。Conneau等人采用一个双向LSTM，并使用max pooling操作符作为语句编码器。



除了选择最佳神经网络模型的通常探索之外，推理线索的成功还提出了以下问题：

哪一个有监督的训练任务可以学习更好地概括下游任务的句子嵌入？

多任务学习可以看作是跳跃思维、推理和相关的无监督/有监督学习方案的概括，这些方案通过尝试在一个训练方案中结合多个训练目标来回答这个问题。

最近的多项多任务学习建议已于2018年初公布。让我们快速浏览一下MILA/MSR的通用语句表示和Google的通用语句编码器。

Subramanian等人在描述MILA和Microsoft Montreal的工作并在ICLR 2018（）上发表的论文中观察到，为了能够在广泛的不同任务中进行概括，有必要对同一句子的多个方面进行编码。

因此，作者利用一对多的多任务学习框架，通过在多个任务之间切换来学习通用的句子嵌入。所选择的6个任务（跳过下一个/上一个句子的思想预测、神经机器翻译、选区解析和自然语言推理）共享双向GRU得到的同一个句子嵌入。实验表明，加入多语言神经机器翻译任务可以更好地学习句法性质，通过句法分析任务学习长度和词序，训练自然语言推理对句法信息进行编码。

谷歌的通用句子编码器也采用了同样的方法。他们的编码器使用一个变压器网络，该网络在各种数据源和各种任务上进行训练，目的是动态适应各种自然语言理解任务。A可用于TensorFlow。

本文总结了通用词语和句子嵌入的研究现状。

在过去的几个月里，这个领域看到了许多有趣的发展，同时我们在评估和探索这些嵌入的性能以及它们固有的偏见/公平性方面也取得了巨大的进步（当您谈到通用嵌入时，这是一个真正的问题）。我们没有时间讨论这些最新的话题，但你可以在参考资料中找到一些链接。

我希望你喜欢这份简报！

如果你喜欢的话，可以拍几下，让我们贴更多！

一些参考文献

* 最近，C.Perone和他的同事们发表了一篇关于ELMo、InferSent、Google Universal Sentence Encoder、p-mean、Skip Think等的详尽对比文章，以下是文章的链接：
* 一个很好的关于传统单词嵌入的资源，如word2vec，GloVe和它们的监督学习增强是。最近的发展是和。
* 句子嵌入纸：，，&。
* 如果你对我们评估句子嵌入的方式感兴趣，你一定要查看Facebook最近的工作以及纽约大学、华盛顿大学和DeepMind研究人员最近发表的文章。