自然语言处理系列（10）：自然语言处理的发展与趋势

译自Deep Learning for Natural Language Processing (NLP): Advancements & Trends

在过去的几年里，深度学习(DL)架构和算法在图像识别和语音处理等领域取得了令人瞩目的进展。

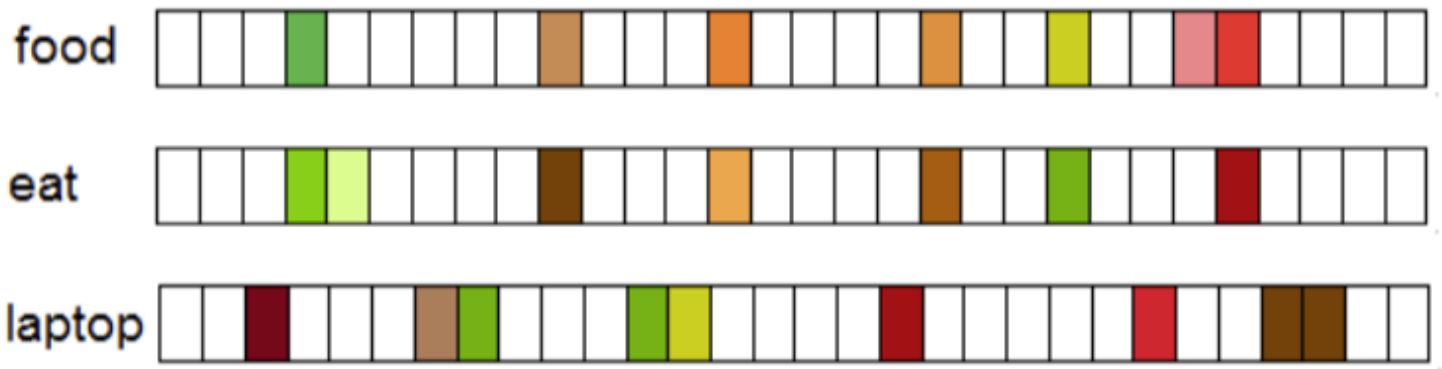
起初，他们在自然语言处理(NLP)上的应用起初并没有那么令人印象深刻，但现在已经被证明能够为一些常见的NLP任务提供最先进的结果。在命名实体识别(NER)，语音(POS)标记或情绪分析等领域，神经网络模型表现优于传统方法。机器翻译的进步也许是最显著的。

在本文中，我将介绍一些依赖于深度学习技术的自然语言处理领域的最新进展。我并不是在装模作样:考虑到有大量的科学论文、框架和工具，这是不可能的。我只想和大家分享一些我最喜欢的作品。我认为过去的几个月对我们的领域非常有利。深度学习在NLP中的使用不断扩大，在某些情况下产生惊人的结果，所有迹象都表明这一趋势不会停止。

From training word2vec to using pre-trained models

可以说，词嵌入是最广为人知的与深度学习相关的技术。他们遵循了Harris (1954)的分布假设，根据这一理论，具有相似意义的词通常出现在类似的语境中。我建议你读一下Gabriel Mordecki的这篇文章。

- Example of distributional vectors of words.



像word2vec(Mikolov et al .,2013)和GloVe(Pennington et al .,2014)已经是这个领域的先驱，尽管他们不能被视为深度学习(word2vec中的神经网络是浅层的而GloVe实现了基于计数的方法)，作为模型训练得到的中间产物，他们在很多的深度学习 NLP的方法中作为输入数据。在我们的领域

使用词嵌入现在普遍认为是一种不错的实践。

一开始，对于一个给定的NLP问题，需要词嵌入，我们倾向于从一个与领域相关的大型语料库中训练我们自己的模型。当然，这并不是普遍使用词嵌入的最好方法，所以预先训练的模型开始慢慢地出现。在维基百科、推特、谷歌新闻、网络爬虫等数据集上进行了训练，这些模型允许你很容易地将词嵌入与你的深度学习算法结合起来。

最新的发展证实，预先训练的词嵌入模型仍然是NLP的一个关键问题。例如，来自Facebook人工智能研究(FAIR)实验室的fastText发布了294种语言的预先训练的向量，它对我们领域做出了巨大贡献。除了大量的不同语言之外，因为fastText使用字符级别的n-grams作为特征，这是非常有用的。这使得fastText可以避免OOV(脱离词汇)问题，因为即使是非常罕见的单词(例如特定的领域术语)，也可能会与更常见的单词共享某些字符n-grams。从这个意义上说，fastText的性能优于word2vec和GloVe，并优于小型数据集。

然而尽管我们可以看到一些进展，但在这方面还有很多工作要做。例如，很棒的 NLP框架spaCy将词嵌入和DL模型集成到诸如NER和依赖性解析等任务中，允许用户更新模型或使用他们自己的模型。

我认为这是一种会不断发展的方法。将来，在NLP框架中使用易于使用的特定领域(如生物学、文学、经济等)的预先训练的模型将是非常棒的。锦上添花的是，用最简单的方法，对我们的用例进行微调。同时，词嵌入的方法也开始出现。

Adapting generic embeddings to specific use cases

也许使用预先训练的词嵌入的主要缺点是，训练数据与实际问题的数据之间存在分布的差异。假设你有一份生物学论文，食物食谱或者经济学研究论文。更有可能的是，通用的词嵌入会帮助你提高结果，因为你可能没有足够大的语料库来训练好的嵌入。但如果您可以将通用的词嵌入应用到特定的用例中呢?

这些类型的适应通常被称为NLP中的跨域或域适应技术，并且非常接近转移学习。Yang等人提出了一项非常有趣的研究。考虑到源域的词嵌入，他们提供了一个正则化的skip-gram模型，用于学习针对目标领域的词嵌入。

关键思想简单而有效。假设我们知道在源域中的单词$word_w$的词嵌入为w_s。为了计算w_t(目标域)的词嵌入，作者在两个域之间添加了一定的传输量。基本上，如果两个域中的单词都是频繁的，那就意味着它的语义不依赖于域。在这种情况下，传输量是很高的，因此在这两个域中所产生的嵌入往往是相似的。但是由于特定领域的单词在一个域中比另一个更频繁，所以传输量很小。基本上，如果两个域中的单词都是频繁的，那就意味着它的语义不依赖于域。在这种情况下，传输量是很高的，因此在这两个域中所产生的词嵌入往往是相似的。但是如果由于特定领域

的单词在一个域中比另一个更频繁，则传输量很小。

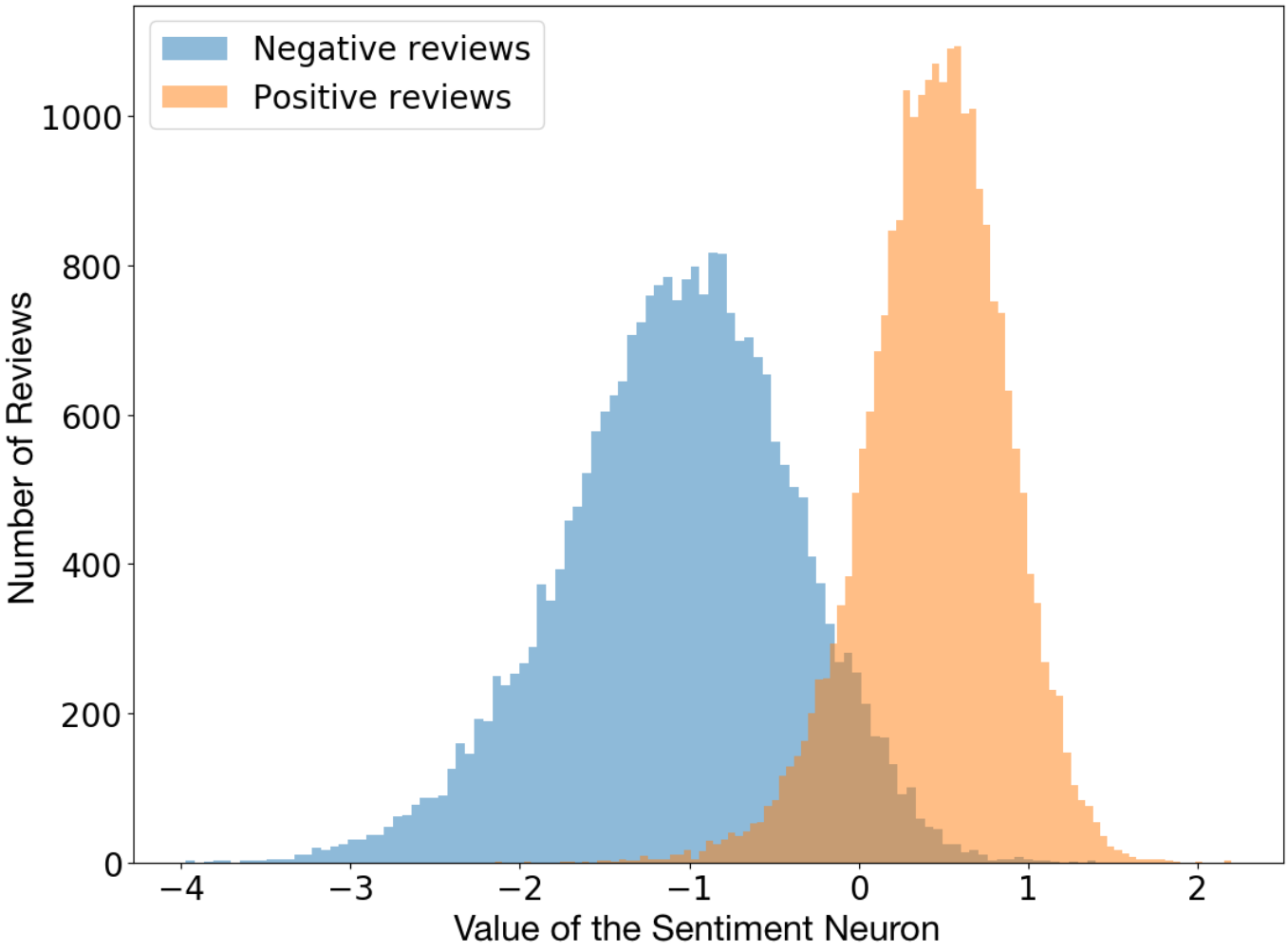
这是一个关于词嵌入的研究课题，并没有得到广泛的研究，我认为它将在不久的将来得到更多的关注。

Sentiment analysis as an incredible side effect

盘尼西林，x光，甚至是post-it都是出乎意料的发现。

Radford等人正在研究字节级循环语言模型的特点，目标是预测亚马逊评论文本中的下一个字符，当他们发现经过训练的模型中的单个神经元对情绪价值的预测非常准确。

- Review polarity vs Value of the neuron.

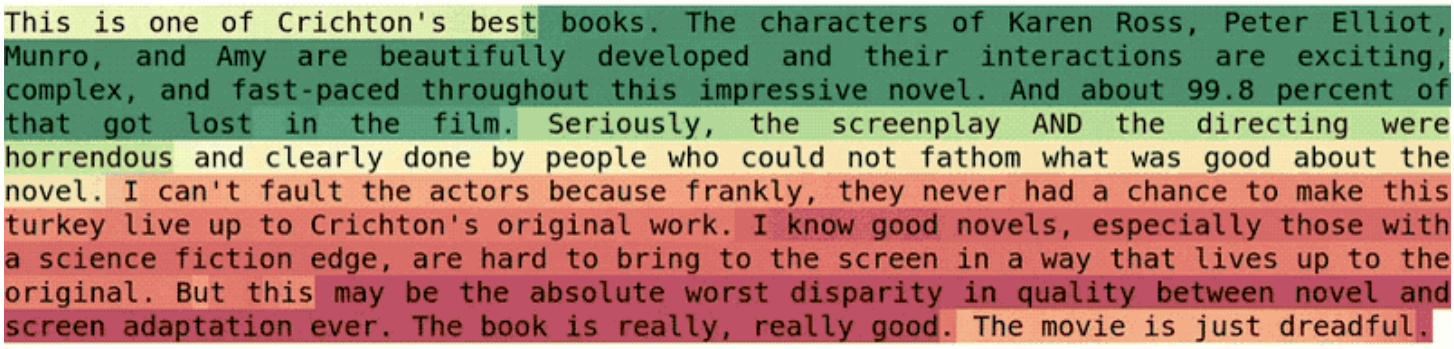


在注意到这一表现之后，作者决定在斯坦福情绪树数据集上对模型进行测试，发现其准确性为91.8%，而之前最好的是90.2%。这意味着，使用明显较少的例子，他们的模型，以无监督的方式训练，至少在一个特定但广泛研究的数据集上达到了最优的情绪分析效果，。

The sentiment neuron at work

由于该模型在作用在字符级之上，所以神经元在文本中改变每个字符的状态，看到它的行为如何运作是非常惊人的。

- Behavior of the sentiment neuron.



例如，在单词best之后，神经元的值变得非常积极。然而，这个效应随着“可怕”这个词的出现而消失了，这是有道理的。

Generating polarity biased text

当然，经过训练的模型仍然是一个有效的生成模型，因此它可以用来生成类似于Amazon评论的文本。但我发现，你可以通过简单地改写情绪神经元的值来选择生成的文本的极性。

- Examples of generated texts (source).

Sentiment fixed to positive	Sentiment fixed to negative
Best hammock ever! Stays in place and holds its shape. Comfy (I love the deep neon pictures on it), and looks so cute.	They didn’t fit either. Straight high sticks at the end. On par with other buds I have. Lesson learned to avoid.
Just what I was looking for. Nice fitted pants, exactly matched seam to color contrast with other pants I own. Highly recommended and also very happy!	The package received was blank and has no barcode. A waste of time and money.

作者选择的NN模型是由Krause等人(2016)提出的一种multiplicative LSTM，主要是因为他们观察到，对于他们所探索的超参数设置，它比普通的LSTMs收敛速度快更快。它拥有4096个单元，并接受了8200万亚马逊评论的训练。

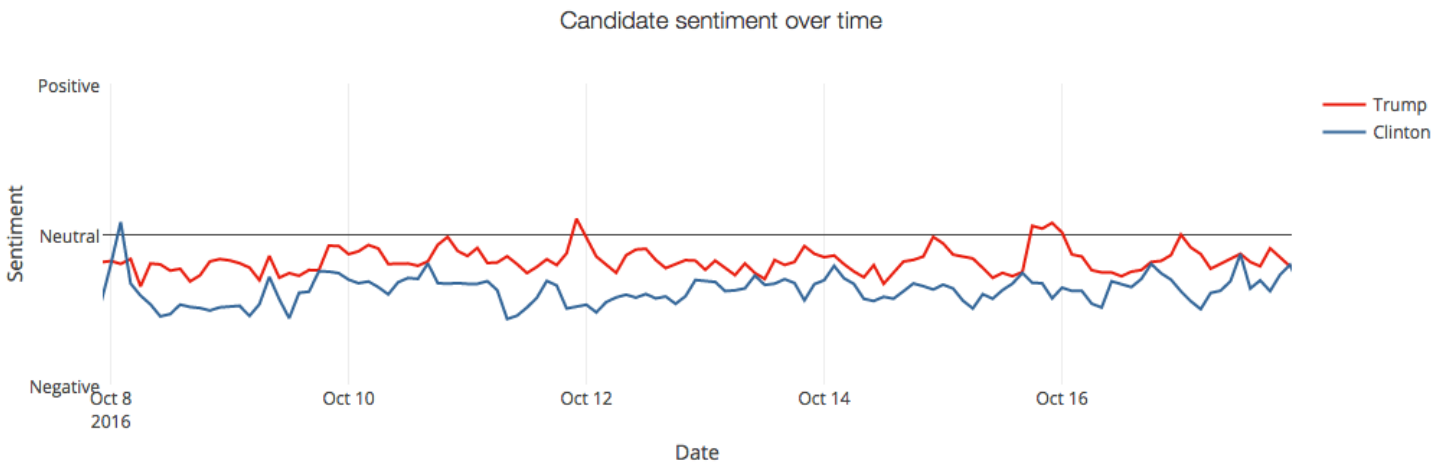
为什么受过训练的模特如此精确地捕捉到情感的概念仍然是一个开放而迷人的问题。同时，你可以试着训练你自己的模型和实验。如果你有时间和gpu，当然这个特定模型通过四个NVIDIA

Pascal gpu用了一个月的时间来进行训练。

Sentiment Analysis in Twitter

无论是去了解人们对你的商业品牌的看法，还是去了解分析营销活动的影响或者在最后一场竞选活动中评估全球对希拉里•克林顿(Hillary Clinton)和唐纳德•特朗普(Donald Trump)的感觉，Twitter的情绪分析是一个非常有力的工具。

- Donald Trump vs Hillary Clinton: sentiment analysis on Twitter.



SemEval 2017

在Twitter上的情绪分析引起了NLP研究人员的广泛关注，同时也引起了政治和社会科学的关注。这就是自2013年以来，SemEval提出了一个具体的任务的原因。

在2017年，共有48个团队参与了评估，这体现了它产生了多大的兴趣。为了让您了解对于Twitter数据SemEval是如何评估的，让我们来看看今年提出的五个子任务。

- 子任务 A：发一条推文，决定它表达的是积极的、消极的还是中性的情绪。
- 子任务 B：给定一条推文和一个主题，将对这个主题的情绪分成两部分:积极的和消极的。
- 子任务 C：给定一条推文和一个主题，将推文中传达的情绪分类为5个点:强阳性、弱阳性、中性、弱阴性和强阴性。
- 子任务 D：给定一组关于主题的tweet，估计在正和负类上的tweet的分布。
- 子任务 E：给定一组关于主题的推文，估计5个类的推文的分布:强阳性、弱阳性、中性、弱阴性和强阴性。

如您所见，子任务A是最常见的任务，有38个团队参与其中，但其他任务更具挑战性。组织者注意到，深度学习方法的使用非常突出，并且不断增加，今年有20个团队使用了像卷积神经网络(CNN)和长短期记忆网络(LSTM)这样的模型。此外，因为SVM模型仍然非常流行，一些参与者将它们与神经网络方法或使用的词嵌入特性结合起来。

The BB_twtr system

值得注意的是，一个纯粹的深度学习系统，即_BB_twtr_系统(Cliche，2017)，在英语5个子任务中排名第一。作者结合了10个CNNs和10个biLSTMs的合集，采用不同的超参数和不同的预训练策略。您可以在本文中看到网络结构的详细信息。

为了训练模型,作者用人类标注的tweets数据(子任务A达到了49693的数量级)，并构建一个由1亿条无标注tweets组成的数据集，他提取一个distant dataset，通过将存在积极正面的表情符号如:-) 标记为正类，反之亦然。这些推文中，小写的、标记化的、url和表情符号等等被特定的标记替换，并统一重复字符，例如，“niiice”和“niiiiiice”都变成了“niice”。

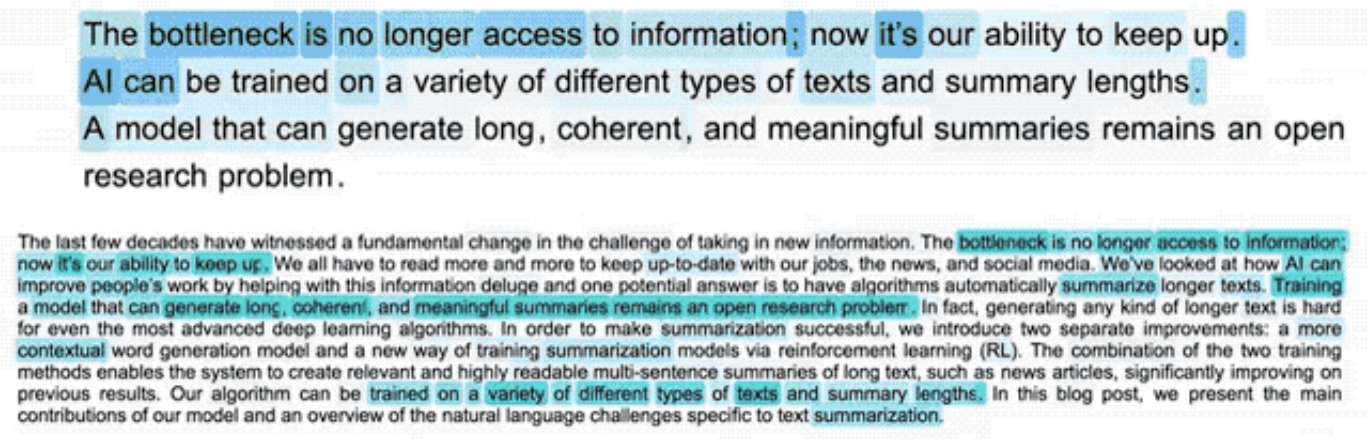
使用之前的SemEval数据集的实验表明，使用Glove会降低性能，而且对于所有的黄金标准数据集，并没有唯一的最佳模型。然后作者将所有的模型与软投票策略结合起来。结果模型比之前的2014年和2016年的最佳历史成绩要好，而且在其他年份非常接近。最后，它在2017年的5个英语子任务中排名第一。

即使组合不是以一种有机的方式进行，而是使用简单的软投票策略，这项工作显示了结合深度学习模型的潜力，同时也说明了端到端方法(输入必须预先处理)可以在Twitter的情绪分析中胜过监督学习的方法。

An exciting abstractive summarization system

自动摘要，自动翻译，是NLP首要任务之一。有两种主要的方法: 基于抽取的方法，它通过从源文本中提取最重要的部分来构建摘要；基于抽象的方法，通过生成文本来构建摘要。从历史上看，基于抽取的方法是最常见的，因为它们的简单性超过了抽象的方法。

在过去的几年里，基于RNN的模型在文本生成方面取得了惊人的成果。它们在短文本的输入和输出中表现得非常好，但对于长文本来说，它们往往是不连贯和重复的。在他们的研究中，Paulus等提出了一种新的神经网络模型来克服这一局限性。结果是令人兴奋的，正如您在下图所看到的。



作者使用一个bi-LSTM编码器读取输入和一个LSTM解码器来生成输出。它们的主要贡献是一种新的内部注意力机制，即对输入和连续生成的输出进行单独的处理，以及一种新的训练方法，它结合了标准的监督词预测和强化学习。

Intra-attention strategy

提出的内部注意力机制的目标是避免输出重复。为了实现这一目标，他们在解码之前对输入文本的前半部分进行解码，然后再决定下一个词的生成。这迫使模型在生成过程中使用不同的输入部分。它们还允许模型从解码器访问以前的隐藏状态。然后将这两个函数组合起来，为输出摘要选择最佳的下一个单词。

Reinforcement learning

为了生成一个摘要，两个不同的人将使用不同的单词和句子顺序，这两个摘要都可能被认为是有效的。因此，一个好的摘要并不一定必须是一组序列单词来尽可能地匹配训练数据集的序列。知道了这一点，作者就避免了标准的teacher forcing 算法，这使得每个解码步骤(即每个生成的单词)的损失最小化，并且他们依赖于一个强化学习策略，这证明是一个很好的选择。

Great results for an almost end-to-end model

该模型在CNN/Daily Mail数据集上进行了测试，并取得了最好的结果。此外，人类评估者的一项具体实验表明，对于人类的可读性和质量也有所提高。此外，基本的预处理的结果令人印象深刻:输入文本被标记，小写，数字被替换为“0”，数据集的某些特定实体被删除。

A first step towards fully unsupervised machine translation?

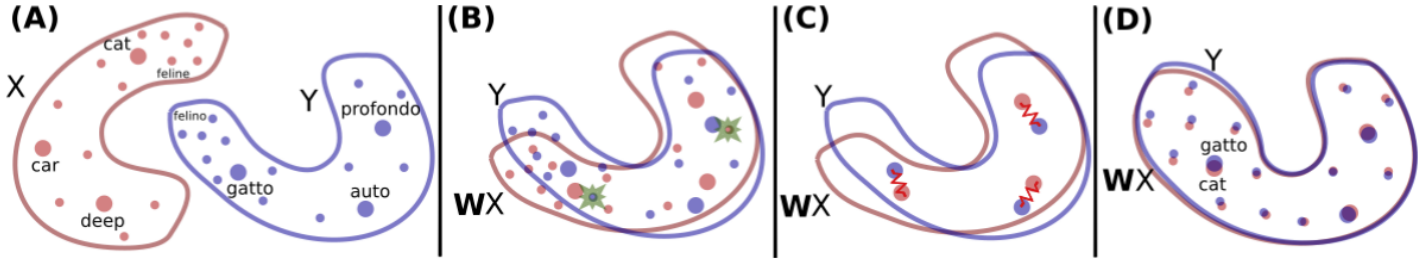
双语词典构建是一项古老的NLP任务，即在两种语言中使用源和目标的单语语料库来识别词翻译匹配对。自动构建的双语词典对其他NLP任务有所帮助，如信息检索和统计机器翻译。然而，这些方法大部分时间都依赖于某种资源，通常是最初的双语词典，它并不总是可用，也不容易构建。

随着词嵌入的成功，跨语言词嵌入的思想出现了，目的是将嵌入空间而不是词汇图标对齐。不幸的是，第一种方法也依赖于双语词汇或平行语料库。在他们的研究中，Conneau等人(2018)提出了一种非常有前途的方法，它不依赖于任何特定的资源，并且对于单词翻译、句子翻译检索和跨语言单词相似性的任务，在几个语言的pairs数据上优于最先进的监督方法。

作者所开发的方法是输入两套在单语数据上单独训练的词嵌入，并在它们之间学习一种映射，这样翻译在公共空间中就很接近了。他们使用未经监督的单词向量，在维基百科上用fastText训

练。下面的图片说明了这个关键的想法。

- Building the mapping between two word embedding spaces.



红色的X分布是英语单词的嵌入，蓝色的Y分布是意大利语词的嵌入。

首先，他们使用对抗学习来学习一个旋转矩阵W，它将执行第一个原始的对齐。他们主要训练一个生成对抗网络(GAN)，遵循Goodfellow等人(2014)的主张。为了直观地了解甘斯的工作方式，我向您推荐巴勃罗·索托的这篇优秀文章。

为了从对抗性学习的角度对问题进行建模，他们定义了discriminator来决定，给定一些从WX和Y中随机取样的元素(见上图中的第二列)，每种语言都属于哪一种语言。然后他们训练W，以防止discriminator做出正确的预测。这在我看来是非常聪明和优雅的，最终的结果是相当不错的。

在此之后，他们再应用两个步骤来完善映射。一是避免在映射计算中引入稀疏的噪声。另一个是建立实际的翻译，主要是利用学习的映射和距离度量。

在某些情况下，结果令人印象深刻的是最先进的。例如，在英语-意大利语单词翻译的例子中，在P@10的案例中，它们的平均精度超过了1.500个源单词的最佳平均精度。

- English-Italian word translation average precisions.

	English to Italian			Italian to English		
	P@1	P@5	P@10	P@1	P@5	P@10
Methods with cross-lingual supervision						
Mikolov et al. (2013b) †	33.8	48.3	53.9	24.9	41.0	47.4
Dinu et al. (2015)†	38.5	56.4	63.9	24.6	45.4	54.1
CCA†	36.1	52.7	58.1	31.0	49.9	57.0
Artetxe et al. (2017)	39.7	54.7	60.5	33.8	52.4	59.1
Smith et al. (2017)†	43.1	60.7	66.4	38.0	58.5	63.6
Procrustes - CSLS	44.9	61.8	66.6	38.5	57.2	63.0
Methods with cross-lingual supervision (Wiki)						
Procrustes - CSLS	63.7	78.6	81.1	56.3	76.2	80.6
Methods without cross-lingual supervision (Wiki)						
Adv - Refine - CSLS	**66.2**	**80.4**	**83.4**	**58.7**	**76.5**	**80.9**

作者声称，他们的方法可以作为无监督机器翻译的第一步。如果真是这样，那就太好了。同时，让我们看看这个新的有希望的方法能走多远。

Specialized frameworks and tools

有很多通用的深度学习框架和工具，其中一些被广泛使用，比如TensorFlow、Keras或PyTorch。然而，特定的开源NLP导向的深度学习框架和工具刚刚出现。对于我们来说，这是一个好年头，因为一些非常有用的开源框架已经向社区开放了。其中三个特别引起了我的注意，你可能也会觉得有趣。

AllenNLP

AllenNLP框架是建立在PyTorch之上的一个平台，它可以在语义NLP任务中轻松使用DL方法。它的目标是让研究人员设计和评估新的模型。它包括用于常见语义NLP任务的模型的参考实现，如语义角色标记、文本蕴涵和相关解析。

ParlAI

ParlAI框架是一个用于对话研究的开源软件平台。它是在Python中实现的，其目标是为对话模型的共享、训练和测试提供一个统一的框架。ParlAI提供了一种与Amazon Mechanical Turk轻松集成的机制。它还提供了该领域的流行数据集，并支持多种模型，包括神经模型，如内存网络、

seq2seq和专注的LSTMs。

ParlAI框架是一个用于对话研究的开源软件平台。它是在Python中实现的，其目标是为对话模型的共享、训练和测试提供一个统一的框架。ParlAI提供了一种与Amazon Mechanical Turk轻松集成的机制。它还提供了该领域的流行数据集，并支持多种模型，包括神经模型，如记忆网络、seq2seq和基于注意力机制的LSTMs。

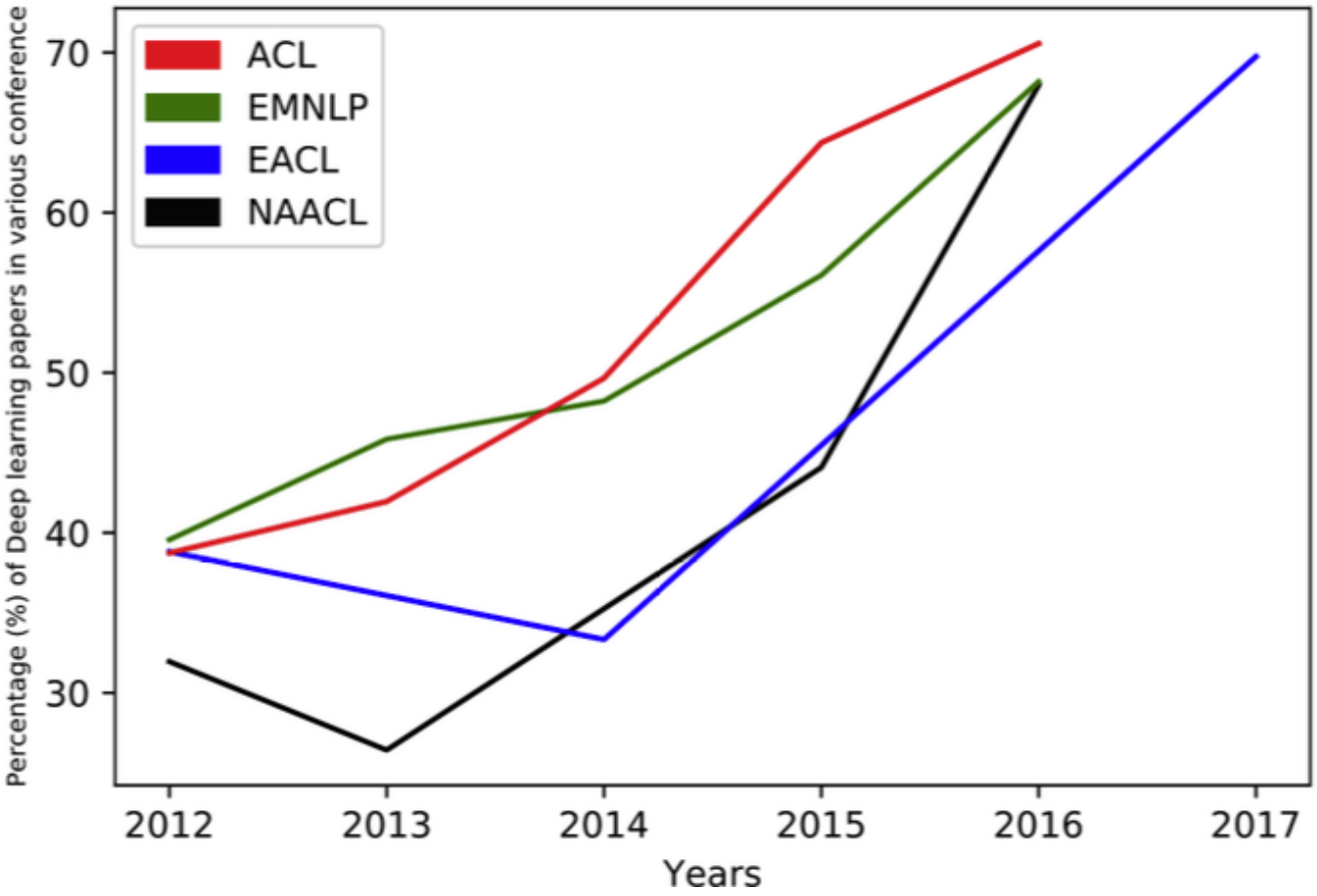
OpenNMT

OpenNMT工具包是一个专门用于序列到序列模型的通用框架。它可以用来执行诸如机器翻译、摘要、图像到文本、语音识别等任务。

Final thoughts

用于NLP问题的DL技术的持续增长是不可否认的。一个很好的指标是，在过去的几年里，ACL、EMNLP、EACL和NAACL等关键NLP会议的深度学习论文百分比的变化。

- Percentage of deep learning papers.



然而，真正的端到端学习才刚刚开始。我们仍在处理一些典型的NLP任务，以准备数据集，例如清理、标记或统一某些实体(例如url、数字、电子邮件地址等)。我们还使用了通用的嵌入，缺点是它们未能捕捉到特定领域术语的重要性，而且它们在多字表达式中表现不佳，这是我在我所研

究的项目中反复发现的一个关键问题。

最新的进展对DL应用于NLP来说非常好。我希望未来能够带来更多的端到端学习的工作，并且特定的开放源码框架得到了更大的发展。请在评论部分与我们分享您对这些作品和框架的看法，以及您今年喜欢的和我在这里没有提到的。

Further reading

关于NLP研究中深度学习方法的更多信息，我强烈推荐你在Young等人的“Recent Trends in Deep Learning Based Natural Language Processing”这篇优秀论文。

另一个有趣的阅读报告是Blunsom等人的“From Characters to Understanding Natural Language (C2NLU): Robust End-to-End Deep Learning for NLP”(2017)，研究人员在NLP，计算语言学、深度学习和通用机器学习所讨论的优势和挑战为深度学习模型使用字符作为输入而不是特定的语言符号。

为了在模型之间有一个比较的视角，我可以给你推荐一个非常有趣的CNN和RNN的比较研究，由Yin等人(2017)进行。

为了直观地了解GANs的工作原理，你可以阅读巴勃罗·索托(Pablo Soto)的这篇出色的文章，它展示了2016年深度学习的重大进展。

我建议你读一下加布里埃尔·莫德基的这篇文章。它以一种说教和娱乐的方式写成，解释了不同的方法，甚至是一些关于词嵌入的传说。

最后，Sebastian Ruder在2017年写了一篇关于词嵌入的很好的文章，你可能会觉得很有用:关于2017年的词汇嵌入: About Word embeddings in 2017: Trends and future directions

Bibliography

- From Characters to Understanding Natural Language (C2NLU): Robust End-to-End Deep Learning for NLP Phil Blunsom, Kyunghyun Cho, Chris Dyer and Hinrich Schütze (2017)
- From Characters to Understanding Natural Language (C2NLU): Robust End-to-End Deep Learning for NLP Phil Blunsom, Kyunghyun Cho, Chris Dyer and Hinrich Schütze (2017)
- BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs Mathieu Cliche (2017)
- Word Translation without Parallel Data Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou (2018)
- Generative adversarial nets Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio (2014)

- Distributional structure Zellig Harris (1954)
- OpenNMT: Open-source toolkit for neural machine translation Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart and Alexander M Rush. (2017)
- Multiplicative lstm for sequence modelling Ben Krause, Liang Lu, Iain Murray and Steve Renals (2016)
- Parlai: A dialog research software platform Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh and Jason Weston (2017)
- Linguistic Regularities in Continuous Space Word Representations Tomas Mikolov, Scott Wen-tau Yih and Geoffrey Zweig (2013)
- Glove: Global vectors for word representation Jeffrey Pennington, Richard Socher and Christopher D. Manning (2014)
- Learning to Generate Reviews and Discovering Sentiment Alec Radford, Rafal Jozefowicz and Ilya Sutskever (2017)
- A Simple Regularization-based Algorithm for Learning Cross-Domain Word Embeddings Wei Yang, Wei Lu, Vincent Zheng (2017)
- Comparative study of CNN and RNN for Natural Language Processing Wenpeng Yin, Katharina Kann, Mo Yu and Hinrich Schütze (2017)
- Recent Trends in Deep Learning Based Natural Language Processing Tom Younga, Devamanyu Hazarikab, Soujanya Poriac and Erik Cambriad (2017)