NLP.TM[16] | SIGIR2019: 深度NLP在搜索系统中的应用

原创 机智的叉烧 CS的陋室 2019-07-28

Forever Right Now

Conor Matthews - Forever Right Now

【NLP.TM】

本人有关自然语言处理和文本挖掘方面的学习和笔记，欢迎大家关注。

往期回顾：

- NLP.TM[15] | 短文本相似度-CNN_SIM
- NLP.TM | 命名实体识别基线 BiLSTM+CRF（上）
- NLP.TM | tensorflow做基础的文本分类
- NLP.TM | 再看word2vector
- NLP.TM | 我的NLP学习之路

由于后续工作原因，所以一直在关注有关搜索系统方面的知识，恰巧看到这个linkedin团队在SIGIR2019上做的有关NLP方面的报告，虽然没有看到视频，但是一份PPT已经非常完善，通过对这份材料的学习，我的收获也是不小，所以此处我也简单整理了一下，希望对大家有帮助。

这篇文章优先收录在NLP.TM下，同时由于与搜索系统有关，所以在R&S下也会有这篇文章。

至于PPT，我也给大家准备了，关注我的公众号**"CS的陋室"**，回复"**Linkedin-NLP**"(直接复制不容易出错哈)，即可拿到这份PPT材料，感谢"专知"的分享。

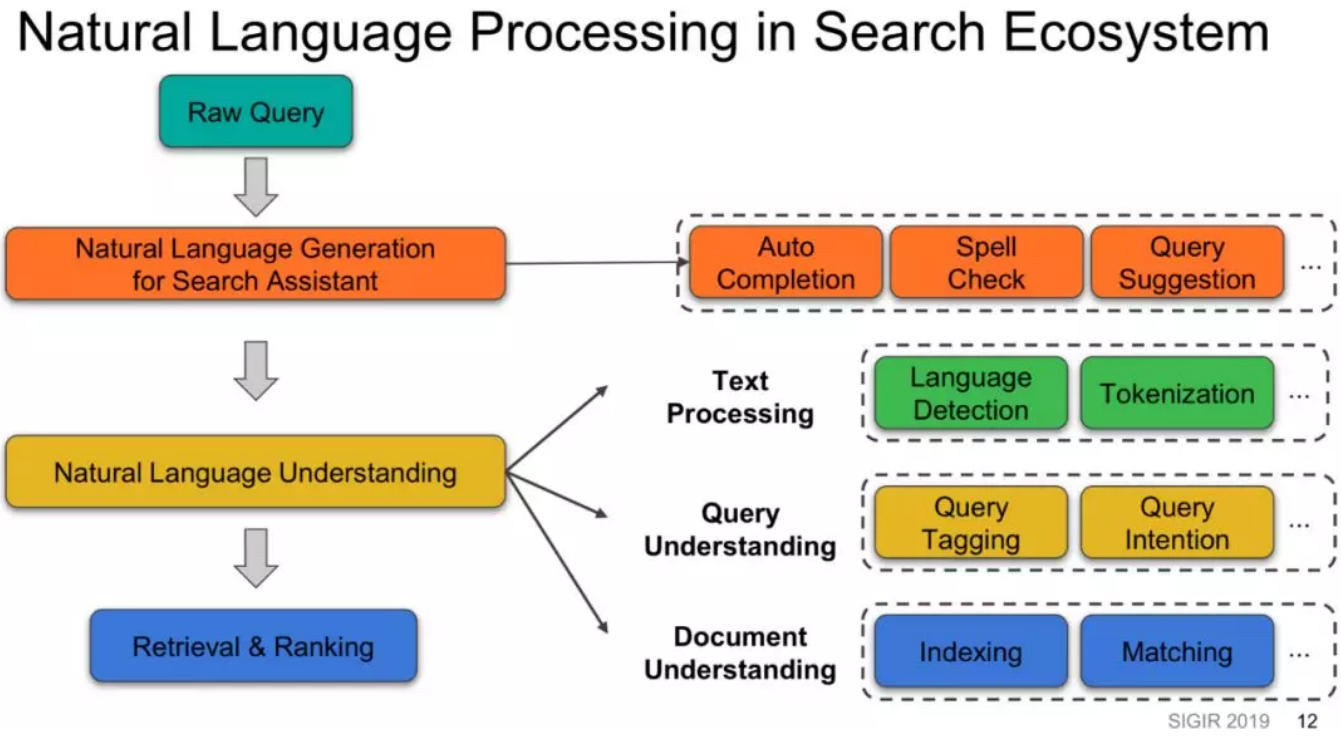
引言

在引言中，作者重点介绍了搜索系统的现状以及NLP在搜索系统中的地位与作用，其实目前，NLP在搜索系统中有非常重要的作用，和推荐系统不同的是，搜索系统需要对用户输入的query进行处理，这是NLP派上用场的重要位置，说的简单，但是实际上NLP在这里面的应用其实很复杂，作者主要分为这几个方面：

- 理解用户意图：既要基于用户本身输入的query，有时还要借助一些用户特征。
- 理解搜索文档：即理解可供搜到的内容(专业术语用document)，标题、评论、标签等。

- 匹配：匹配用户意图和搜索文档，甚至涉及到召回和排序(这里就和推荐系统有点类似)。
- 搜索辅助：例如预想、修正、补全等。

那么在整个搜索生态中，NLP的作用就可以用这张图来表示了。



左边是基于整个搜索系统的过程流程与主要功能，右边则是在NLP角度表述的涉及到的任务。

- 在辅助搜索阶段，主要是一个类似文本生成的过程(个人理解这个和常说的文本生成有些许区别)，主要涉及自动补全，拼写检查(中文则有错别字等)以及搜索建议。
- 在检索词理解阶段，主要包括了文本处理、搜索句理解、与文档理解等部分，里面对应涉及语言检测、切词、打标签、意图识别等过程。
- 另外还有召回和排序。

而近年来，NLP的发展速度加快，很大一个原因就是和深度学习的发展有关，深度学习的发展促成很多NLP任务有了新的突破，主要原因有如下原因：

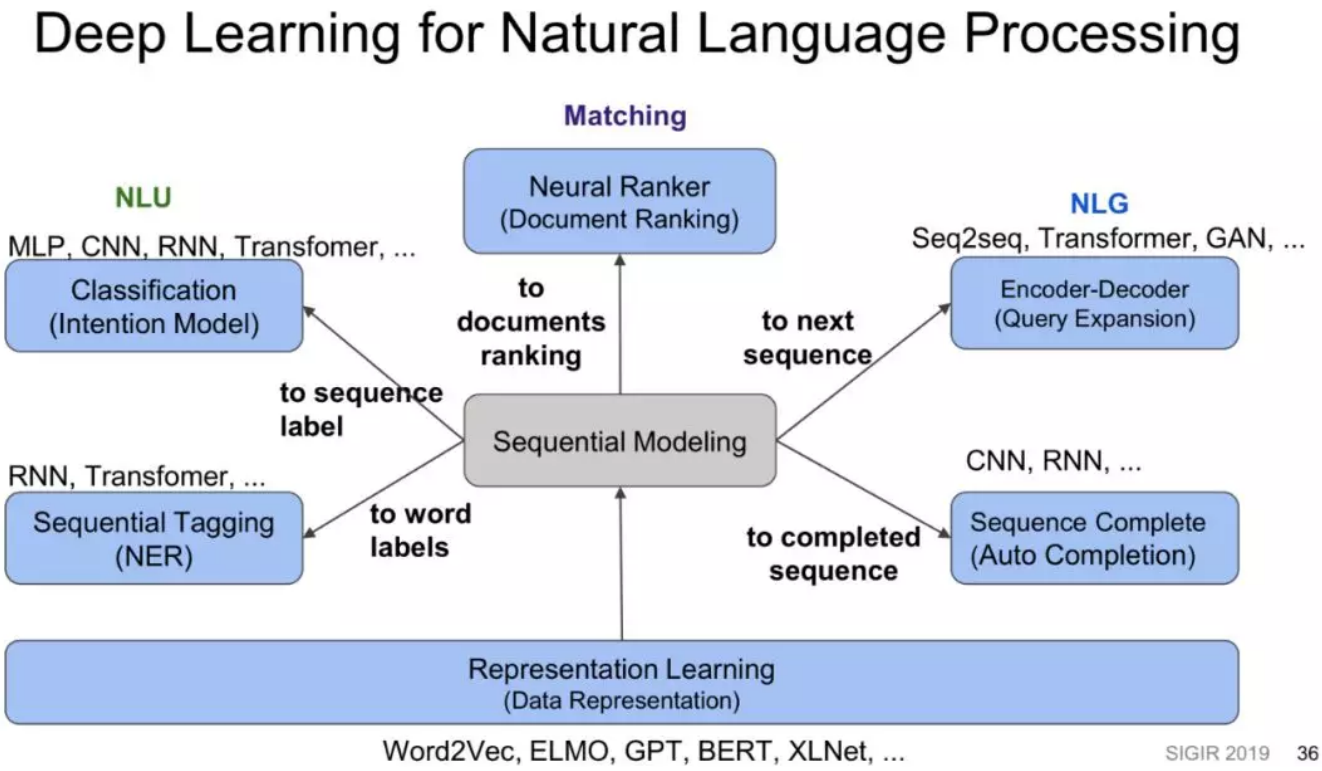
- 深度处理高维稀疏信息
- 简便的特征工程
- 模型的灵活性
- 多层次特征表达

这些原因可以参考这篇论文：

[1] Recent Trends in Deep Learning Based Natural Language Processing, arXiv preprint arXiv:1708.02709v8, 2018

深度学习在NLP中的应用

这块应用作者用了一张图来解释，我感觉总结的还是比较全面的，上面简单提到一些模型，虽然有些会重复，但是实际上这些模型在很多方面效果确实比较好，例如RNN在很多任务中都多多少少有出现(或者其变式)。

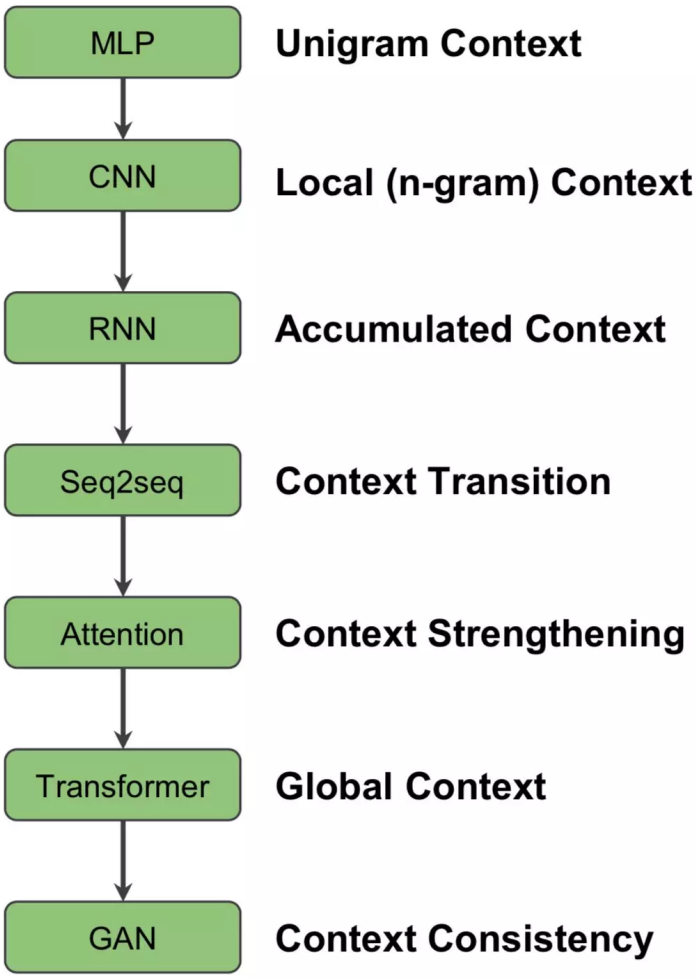


整块核心有4，NLU(自然语言理解)、NLG(自然语言生成)、匹配、表征学习。在报告中，主要针对两个任务来进行阐述，分别是情感上下文问题中的序列模型以及数据处理阶段的表征学习。

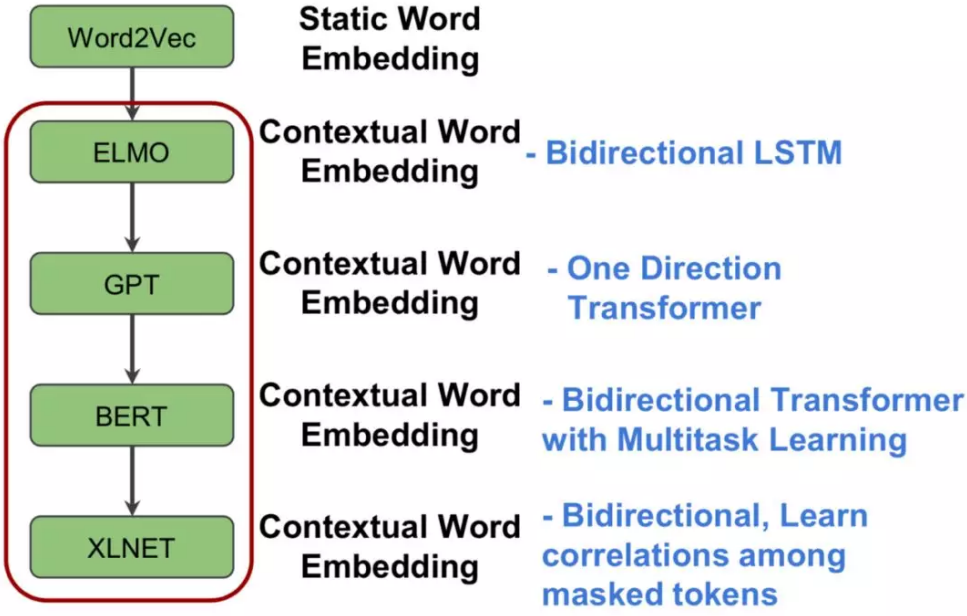
有关序列模型方面，我曾经写过一篇有关基于BiLSTM-CRF的命名实识别模型文章，在这里：

- NLP.TM | 命名实体识别基线 BiLSTM+CRF（上）
- NLP.TM | 命名实体识别基线 BiLSTM+CRF（下）

作者给出了一条比较长的路线，由简单到复杂，无论是从模型使用本身(左列)，还是从模型应用角度(右列)，这个给大家一个十分理想的模型思考路线，根据自己问题的复杂性来判断自己要选用那种模型。(具体这些模型的分别含义，建议大家去查一下吧，前三个应该是非常基础的模型，后面4个逐渐到达前沿水平)。



在表征学习方面，说白了就是特征提取一块，有我们熟悉的word2vec、GloVe等，也有逐步到达前沿的ELMo、GPT、BERT以及XLNet，至于哪个好那个不好，这个我建议还是根据自己了解尝试，一不看模型是否高端，二不看媒体是否吹，而应该理解模型本身，根据自己的问题难易程度来处理。



深度NLP在搜索系统中的应用

这块内容非常多，三分之二甚至更多的内容都在围绕这块讲解。

语言理解

这块的任务都是以分类问题为准，所谓的理解，只是把语言抽象到某一个理解空间，将其进行标准化，以便进行批量化处理。这里作者提到了3块，分别是实体标注(即命名实体识别)、实体消歧或指代消歧(基于知识的预测)、意图识别(句子级别预测，甚至就是简单的文本分类)。

实体标注方面，传统的统计方法是HMM(隐马尔可夫)、MEMM(最大熵马尔可夫)、CRF(条件随机场)，基本就和命名实体识别类似了，而在深度学习引入后，形成了输入层、编码层、解码层的主要架构，同过预训练表征模型(如w2v)、深度学习结构(CNN、RNN等)、以及输出层(CRF、softmax)等结构链接，完成最基本的结构。

实体消歧或指代消歧主要是解决在用户搜索的语句中出现的问题，例如"苹果"到底是水果还是手机等等，这个是依赖上下文信息和知识库合力完成的，例如一句话"我爱吃苹果"，这个"吃"其实就是一个上下文的信息，另一方面我们要通过这个"吃"推断出这个水果的含义，我们就需要借助知识库。

意图识别方面，应该是这几块里面最简单的，就是一套深度学习，框架即可完成，PPT中提到的是fasttext、CNN以及BiRNN-Attention，我这给大家提供两篇我写的文章，讲了我的实现方式。

文档召回和排序

这个思路和推荐系统类似，我们先把有关的全都拿出来，然后再用更为精细的方法排好序展示给用户，此处就有两个大步，召回和排序。

召回方面，要求更全，此处又有句法召回和语义召回。句法召回说白了就是匹配，但这里面的学问可是非常多的，字符串匹配、倒排索引(搜索系统中非常关键的基础知识)、多路召回(多领域)，语义召回则是通过词向量近邻等方式扩大召回的内容。

排序方面，LTR(learning to rank)其实是一个隐含在暗线但实际上已经非常经典的方向，就是为了研究排序的。

文本生成辅助搜索

这块主要用于辅助用户进行搜索，主要体现在下面三块功能上：

- 自动补全。很好理解，大家在很多搜索引擎中都会看到，在百度下输入"自然语言"，他能给你预测出你可能要搜"自然语言处理"。

- 搜索重构。举个例子吧，你输入的是吃鸡，实际上文档库的标题是"和平精英"，那要映射过去，其实就是一种同义词重构。
- 拼写修正。英文有拼写问题，中文有错别字问题， 不能保证用户100%输入正确，平时打字都可能手滑，为了更准确理解语义，我们必须在进行语义分析前修正这些错误。

最后还有一章讲解案例的，其实是对前面内容的重述，此处不再展开啦，上面提到的有关问题和方法如果能解决，其实这块就已经能够基本掌握。

我是叉烧，欢迎关注我！

叉烧，机器学习算法实习生，北京科技大学数理学院统计学研二硕士毕业，本科北京科技大学信息与计算科学、金融工程双学位毕业，硕士期间发表论文6篇，学生一作3篇，1项国家自然科学基金面上项目学生第2参与人，参与国家级及以上学术会议4次，其中，1次优秀论文，国家奖学金，北京市优秀毕业生。曾任去哪儿网大住宿事业部产品数据，美团点评出行事业部算法工程师。

微信 zgr950123
邮箱 chashaozgr@163.com
知乎 机智的叉烧

微信个人公众号
CS的陋室