打破情感分类准确率 80 分天花板！更加充分的知识图谱结合范式

原创 Severus 夕小瑶的卖萌屋 2021-09-23 12:05

文 | Severus

NLP的研究者们一直都在尝试，怎么样让模型像人类一样，学会“知识”。而最直观的想法莫过于将人类已经总结出来供机器解读的“知识体系”，及其嵌入表示作为额外的特征添加到NLP模型之中。至少，从直觉上看，将知识融入到模型之中，可以**让模型直接“看到”知识体系所带来的“言外之意”，从而与模型本身的统计共现特征形成互补**，以补足训练样本中部分知识过于稀疏的问题。比如某一实体A在训练样本中频次很低，则可以**用与它相似，且频次较高的实体B的特征来补充A**，或者只是样本中的表达比较稀疏，则使用知识体系中的另一种更加常用的表达来补充（例如：OSX vs MacOS，歌神 vs 张学友），从而弥补A的特征过于稀疏的问题；或者可以**使用A所在的归类体系中共享的特征来补充A的特征**。

然而，模型需要什么样的知识，要以什么方式将知识整合到模型之中，一直是存有争议的问题。例如早几年很多工作尝试，使用知识图谱表示，将实体关系融合到模型中，在一些任务上取得了成效，但其最大的限制之一，则是**消歧始终难以做到很高的准确率**，其原因在于，**知识图谱所收录的绝大多数实体，信息都是稀疏的（SPO密度很低）**，它们甚至很难参与到实体链指环节之中，所以很多 KGs+NLP 的工作都是在**有限的知识图谱内**进行的，而难以扩展到**广域的知识图谱**中。

除知识图谱外，则也有将**通用知识**引入到模型之中的工作，例如近两年很多将中文的组词应用到 NER 的工作，将实体类别信息应用于关系抽取的工作等，甚至我们可以开更大的脑洞，**直接利用预训练语言模型从海量语料中学习到的充分的共现知识，用以表示通用知识**，将之应用到基于预训练语言模型的种种方法中。

下面我想要介绍的工作，则是使用大规模知识图谱增强模型，做 aspect-level 的情感识别任务，作者声称，自己的方法相对 baseline 分别有2.5%~4%的提升。

大规模知识图谱增强的 aspect-level 情感识别

论文标题：
Scalable End-to-End Training of Knowledge Graph-Enhanced Aspect Embedding for Aspect Level Sentiment Analysis

论文地址：
https://arxiv.org/abs/2108.11656

Aspect-level 的情感识别，即输入一段文本，询问该文本对某一个文本中提到的片段是什么样的情感倾向。例如句子：However, I can refute that OSX is "FAST". 中，询问句子中对 OSX 表达了什么样的情感。之前的工作很少将这个任务的分数刷到80分以上，本文作者则一鼓作气，将3个数据集的最终指标都刷到了80+。

Aspect-level 情感分类的难点在于，**aspect 有可能是稀疏的，从而导致模型在“观察”文本的时候找不到重点**，例如上面的例句，OSX 在对应的训练样本中仅仅出现了7次，非常的稀疏，而与之相似的 Microsoft Windows 则出现了37次。而使用训练样本中相对高频的 aspect 去补充相对低频的，又恰恰是知识增强的动机之一，所以利用知识图谱来增强这个任务，看上去相当的合适。

但是知识图谱增强又存在两个挑战：

1. **大规模知识图谱难以完全利用起来**，例如 DBPedia 有2200万节点，1.7亿条边，计算其中所有实体的表示显然也是不现实的
2. **知识图谱实体消歧错误传递**，这点在前文也有提到。

针对这两点挑战，本文都给出了相应的解决方案。

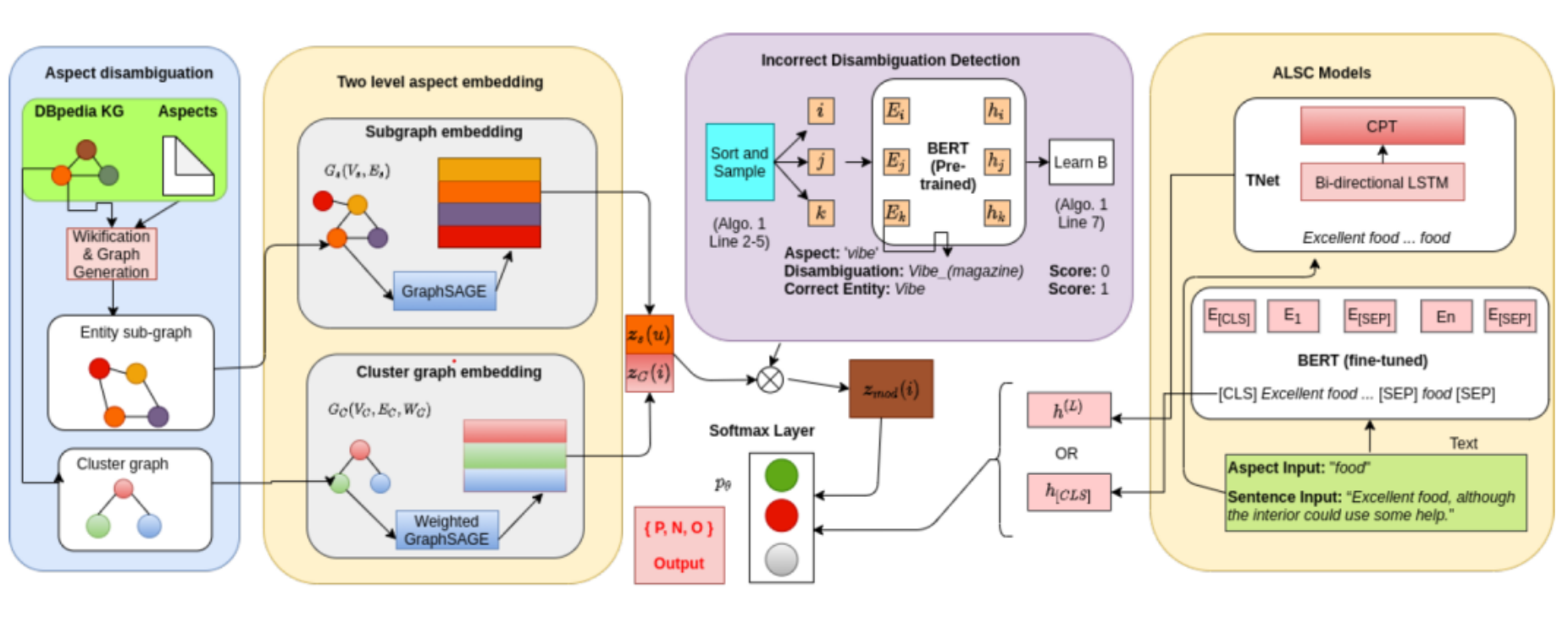


Figure 1: Architecture of our framework for end-to-end training of ALSC model while incorporating aspect relations from large KGs. Yellow backgound denotes modules trained end-to-end, green backgound denotes input.

▲系统总体结构

图谱表示

本文使用了两种方式计算图谱表示，分别为**子图表示**和**连通分量表示**。其中，连通分量表示则是将整个知识图谱中划分为若干个连通分量，每个连通分量看作是一个节点，从而将大规模图缩放成一个相对较小的图，例如本文将 DBpedia 的2200万个节点划分为606个联通分量来计算表示。计算方法使用的都是 GraphSAGE，简单来讲就是用某一个节点随机游走的N跳邻居层层聚合，得到当前节点的表示。

连通分量表示的方式则使用一种比较朴素的方式解决了大规模图谱表示的问题，实则使用的还是子图表示的计算方法。其好处则在于**某一个节点可以得到的“言外之意”变得更多**，更加看上去有关的信息被利用了。

其中，图表示的训练方式也分为静态训练和端到端训练两种，静态图表示是先训练好图表示，再叠加到任务中，端到端训练则是在任务训练的同时也训练图表示。

去掉歧义噪音

针对歧义噪音问题，作者则是使用 BERT 所学习到的统计共现知识去解决。首先我们可以认为，BERT 所学到的文本表示，聚合了很多的信息，而对于一个 aspect ，它的表示则聚合了其描述信息、分布信息等，那么，图谱嵌入空间上相近的实体，则在 BERT 学到的表示空间里面也应该有较高的相似性，但BERT聚合到的信息又太多了，所以需要将所需要的信息相办法抽取出来。所以，定义两个实体i和j的相似函数为：

$$S_B(h_i, h_j) = \sigma((B \cdot h_i)^T (B \cdot h_j))$$

其中B是可训练的参数，h_i和h_j分别是实体i和实体j的BERT表示的 `[CLS]` 向量。然后分别采样在图谱表示空间里距离近的实体和距离远的实体作为正例和负例，训练参数B，loss为：

$$\sum_{(i,j,l)\in\tau} (S_B(h_i, h_k) - S_B(h_i, h_j))$$

其中，i和j是相近实体，i和k则是不相近的实体，该目标是尽可能让BERT学到的表示和乘上参数矩阵B之后，与图嵌入空间里面的距离更加相关。

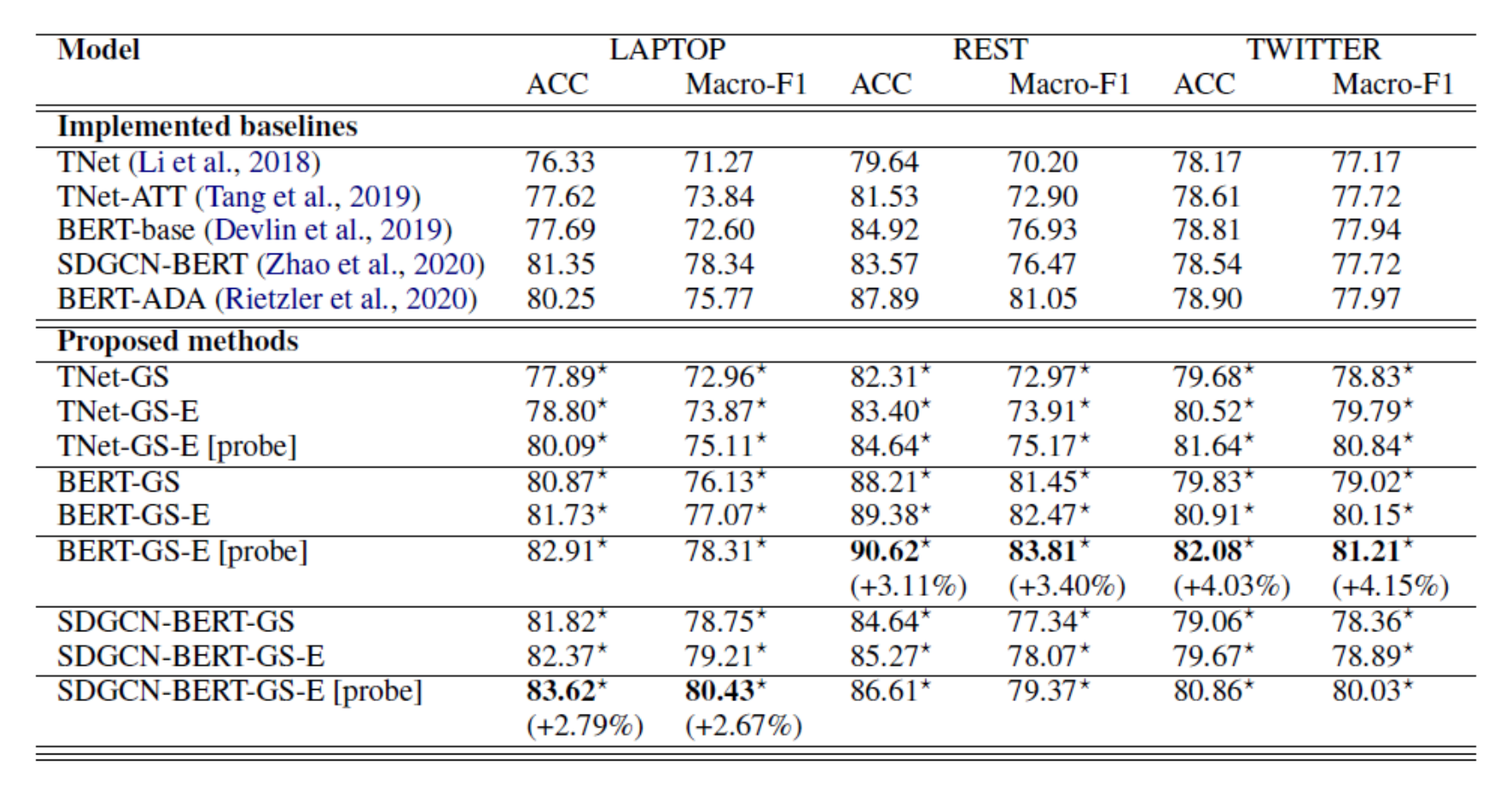
而最终使用的实体i的表示则为：

$$z_{mod}(i) = \begin{cases} \{\vec{0}^{dim_h}\} & \text{, if } S_B(h_i, h_j) - S_B(h_i, h_k) \geq 0 \\ z(i) & \text{, otherwise} \end{cases}$$

也就是说，**如果实体i的BERT表示的相似度和图谱嵌入空间内的相似度出现了矛盾，则屏蔽掉它的图谱表示**，作者认为这样可以屏蔽掉很多消歧算法带来的噪音。

实验结果

Table 2: Experiment results on various datasets(%). The marker * refers to p-value <0.01 when comparing with respective baselines. % in bracket of best performing models implies overall gain wrt. its' baselines.



Model	LAPTOP		REST		TWITTER	
	ACC	Macro-F1	ACC	Macro-F1	ACC	Macro-F1
Implemented baselines						
TNet (Li et al., 2018)	76.33	71.27	79.64	70.20	78.17	77.17
TNet-ATT (Tang et al., 2019)	77.62	73.84	81.53	72.90	78.61	77.72
BERT-base (Devlin et al., 2019)	77.69	72.60	84.92	76.93	78.81	77.94
SDGCN-BERT (Zhao et al., 2020)	81.35	78.34	83.57	76.47	78.54	77.72
BERT-ADA (Rietzler et al., 2020)	80.25	75.77	87.89	81.05	78.90	77.97
Proposed methods						
TNet-GS	77.89*	72.96*	82.31*	72.97*	79.68*	78.83*
TNet-GS-E	78.80*	73.87*	83.40*	73.91*	80.52*	79.79*
TNet-GS-E [probe]	80.09*	75.11*	84.64*	75.17*	81.64*	80.84*
BERT-GS	80.87*	76.13*	88.21*	81.45*	79.83*	79.02*
BERT-GS-E	81.73*	77.07*	89.38*	82.47*	80.91*	80.15*
BERT-GS-E [probe]	82.91*	78.31*	**90.62*** (+3.11%)	**83.81*** (+3.40%)	**82.08*** (+4.03%)	**81.21*** (+4.15%)
SDGCN-BERT-GS	81.82*	78.75*	84.64*	77.34*	79.06*	78.36*
SDGCN-BERT-GS-E	82.37*	79.21*	85.27*	78.07*	79.67*	78.89*
SDGCN-BERT-GS-E [probe]	**83.62*** (+2.79%)	**80.43*** (+2.67%)	86.61*	79.37*	80.86*	80.03*

▲实验结果

上表中，`GS` 后缀是使用了静态训练得到的表示增强的方法，`GS-E` 后缀则是在原有基础上使用了端到端训练得到的表示增强的方法，`[probe]` 后缀则是在原有方法基础上使用了去掉歧义噪音策略的方法。我们可以看到，在3个数据集上，文本所提出的方法都各有不算小的提升，而尤其去掉歧义噪音之后，分别都得到了SOTA的结果，可见作者的方法还是有一定增益的。

> 顺便一提，这个结果里面作者玩儿了个文字游戏，比如 `SDGCN-BERT-GS-E[probe]` 的结果提升了2.79%，这个结果是这么是计算出来的：(83.62-81.35)/81.35*100%=2.79%，同理其他的提升也是这么算出来的，并不是绝对分数的提升，而因为分母不是100，所以提升数值都需要相对减少一些。

小结

我认为，本文还欠缺了一个分析实验，即**连通分量表示是否是有用的**。直观上来看，将2200万个节点硬性划分成606个连通分量，去计算整个图的表示，总是感觉过于朴素和粗暴了。毕竟作者没有讲他是以什么样的标准去划分，我们也没有办法去评析这种划分方式是否合理，同时我们也没法知道，**连通分量表示在这篇工作中到底起到了什么样的作用**，是否仅仅需要子图表示加上去除歧义噪音的策略，就足以得到这么好的效果了呢？甚至**极端情况下，如果数据集里面的 bias 比较大，按照这种划分方式，是否会将绝大多数 aspect 都分配到同一个连通分量里面，从而导致这个特征变成了一个废特征呢**？

并且，感觉上连通分量表示则是**为了大规模图谱而大规模图谱**。不可否认，这篇文章使用图谱增强任务，得到了一定的提升，但是这种提升，我认为更多还是**在于利用有限、固定的图谱的信息，加上噪音消除策略**而达成的，真正到了广域数据，需要大规模图谱的场景下，所要面对的问题绝不仅仅是计算瓶颈那么简单。

例如，在开头我就提到的绝大多数实体过于稀疏的问题，与之相伴的还有收录的问题。**世界上不可能存在一个图谱，能够收录尽世界上所有的事实知识**，莫说图谱，牛津英文词典的收录情况就已经回答了这个问题。哪怕相关研究者们不断地更新、迭代图谱自动收录算法，和图谱自动补全算法，但是也难以赶上新知识的产生速度，同时图谱要保证事实准确、高质量，那么其准入门槛也不可能允许超高速的收录。就说相对还比较固定的专业领域知识，也面临着语言不全等问题，例如生物名录数据库，英文数据库中也存在很多中文数据库里面没有的条目。所以，我们没法指望知识图谱能够枚举世界上所有的事实。

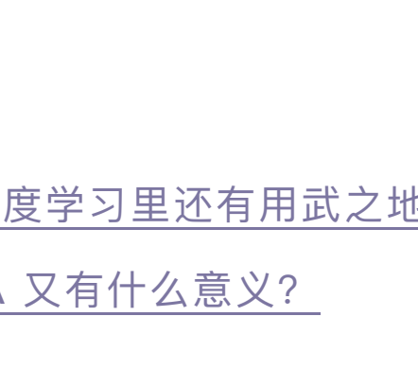
收录问题也不是最关键的因素，毕竟，没有收录的知识，我们可以在任务里面不去使用它，那无非它的效果退化到原始模型的效果而已。信息稀疏所引发的消歧问题，也可以通过置信度阈值去控制它，保证实体链指的准确率，避免错误传递。但是抛开这两个问题，最关键的还是**统计模型与知识图谱的特性**。

如果使用统计模型去将知识图谱嵌入到连续空间中，则必然要面对**统计模型的泛化能力**，但是，**事实知识是不可泛化的**（例如当我们询问GPT-3/ERNIE3.0，太阳有几只眼睛/我的脚有几只眼睛的时候，这个问题事实上是不成立的，但是统计模型总是会泛化出一个结果）。到了大规模稠密知识图谱上，这个问题则会更加严重，例如图嵌入多是使用随机游走采样计算节点相似性，但是**知识图谱上绝大部分的多跳路径是不成立的，其边是不可传递的**。例如：刘德华的搭档是刘伟强，刘德华的老婆是朱丽倩，那么刘伟强和朱丽倩之间有什么关系呢？如果采样过程不受控制，这三者在统计空间里面可能会非常接近，哪怕受了控制，保不齐通过其他采样的泛化，还是会挂上关系。

而在统计模型里面，想要定死了这些规则，恐怕要通过无数的样本去拉近、推远一些表示，才有可能完成在搜索空间里面构建出来一套完整的规则，相比于直接用符号推理来讲，我认为得不偿失。

> 实际上哪怕知识图谱补全任务里面，也有很多数据是不可推理数据，那些数据很有可能就是用统计特征算出来，扔到数据集里面的，例如一个人是美国人，那他的信仰是天主教之类的，当年我做图谱表示的同事分析结论，一些分数很难刷，“很难”的数据集里面，这类数据似乎也占据了相当大的比重。

但是，在benchmark上，图谱增强又往往能带来一些看上去不错的增益，我认为，其主要在于这类任务面对的都是有限、固定的子集，例如本文中，使用到的子图规模100到1000不等，这种规模之下，则绝大多数情况下可以规避掉我上面提到的两个问题。也就是说，**在固定垂直领域下，信息密度大，没有消歧压力、不可控泛化的压力，图谱增强是有用的，但是一旦到了开放领域，则不得不面对上面的问题。**



萌屋作者：Severus

Severus，在某厂工作的老程序员，主要从事自然语言理解方向，资深死宅，日常愤青，对个人觉得难以理解的同行工作都采取直接吐槽的态度。笔名取自哈利波特系列的斯内普教授，觉得自己也像他那么自闭、刻薄、阴阳怪气，也向往他为爱而伟大。

作品推荐

1. 深度学习，路在何方？
2. 数据还是模型？人类知识在深度学习里还有用武之地吗？
3. 在错误的数据上，刷到 SOTA 又有什么意义？

后台回复关键词【入群】
加入卖萌屋NLP/IR/Rec与求职讨论群
后台回复关键词【顶会】
获取ACL、CIKM等各大顶会论文集！