# 基于Bradely-Terry的数据增强方法[[1]](#footnote-1)

潘\*\*，\*\*\*，\*\*\*

（解放军战略支援部队信息工程大学，河南 郑州 450001）

摘 要：在回归任务中，不同身份、个人对单个样本标注结果存在较大偏差，因而许多机构通过向注释者提供成对比较样本，来缓解标注的偏差。成对样本标注方式出于人工成本考虑不能对任意样本对进行标注，无法做到完全标注。为了能从已有标注信息中获取更多成对样本信息，本文从数据增强出发，提出了一种基于Bradley–Terry方法结果进行采样的数据增强方法，并在原始方法上利用分层和排列的思想，改善了原始方法采样结果容易受随机种子影响和不同句子采样不公平的缺陷。

关键词：样本标注；数据增强；Bradley-Terry

中图分类号：TP393

文献标识码：A

**doi:** 10.11959/j.issn.2096−109x.2019xxx

**Data enhancement method based on Bradley Terry**

\*\*\*, \*\*\*, \*\*\*

People’s Liberation Army Strategic Support Force Information Engineering University, Zhengzhou 450001, China

**Abstract:** In the regression task, different identities and individuals have large deviations from the labeling results of a single sample. Therefore, many institutions alleviate the labeling deviation by providing paired comparison samples to annotators. For the sake of labor cost, the paired sample labeling method cannot label any sample pair and cannot label completely. In order to obtain more paired sample information from the existing annotation information, starting from data enhancement, this paper proposes a data enhancement method based on the results of Bradley Terry method, and uses the idea of layering and arrangement in the original method to improve the defects that the sampling results of the original method are easy to be affected by random seeds and unfair sampling of different sentence.

**Key words**: sample labeling, data enhancement, bradley terry

## 1.引言

情感分析是对人们对产品、服务、组织、个人、问题、事件、话题及其属性的观点、情感、情绪、评价和态度的计算研究。

该领域的开始和快速发展离不开社交媒体的高速发展，论坛、博客、微博等社交平台的快速增长，为此带来了海量的数字形式的数据。

早在2000年，情感分析就成为NLP中最活跃的研究领域之一。它在数据挖掘、Web挖掘、文本挖掘和信息检索方面得到了广泛的研究。现有的许多研究已经被应用到情绪分析的各种任务中产生了，包括有监督和无监督的方法。近年来，将深度学习应用于情绪分析也变得非常流行。

在有监督情况下，想要进行情绪分析，最基本的条件就是一份高质量的标注数据集。标注人员一般能较以较高质量完成一些粗粒度的标注，如一句话表达情绪是否积极，但是当标注人员被要求进行更细粒度的实值分数标注（如积极程度分数），这个任务变得相当困难。当被问及文本的实值情绪分数时，相比于简单地将情绪分为正面或负面，给出具体分数让受访者面临更高的认知负荷。此外，不同注释者很难保持一致，每个人都有自己的评判标准，相同的文本在不同身份的注释者眼中的情绪程度可能有较大差异。

通过向注释者提供成对的文本并询问哪一个更积极，可以克服这些问题，这缓解了标准评分量表中存在的几种注释偏差，如量表区域偏差，[1,2]并提高了注释的一致性[3] 。

然而，现在面临的情况从注释n个实例变为需要注释个实例对，这可能是不现实的。出于人工成本考虑，实际标注的数量要远远少于。事实上也不需要完完全全标注句子对。举一个简单的例子，我们标注了A>B,B>C,显然我们可以知道一个隐藏的句子对关系就是A>C。我们希望能根据标注结果挖掘出更多这样的隐藏信息作为数据集的补充，帮助有监督情况下的情感分析。

本文以数据增强为目的，提出了一种基于Bradley–Terry方法结果进行采样的数据增强方法，并在原始方法上利用分层和排列的思想，改善了原始方法采样结果容易受随机种子影响和不同句子采样不公平的缺陷。

1.根据比较数据集，使用了Bradley–Terry方法的结果来进行采样策略来挖掘更多数据信息。

2.在原始采样方法的基础上，结合分层策略和排列组合思想，缓解了原方法中存在的采样结果受随机性影响较大和采样不公平情况。

## 2.相关工作

从嘈杂的两两比较中对n个项目的集合进行排序的研究具有广泛的应用场景，包括对电影、书籍或其他消费项目进行评级的推荐系统[4,5]；在大规模开放在线课程中排名学生的综合评分[6]；搜索引擎；通过对城市街道景观的两两比较，量化人们对城市的感知[7]；

成对比较法是由心理学家提出的一种数学方法，后来被应用到了统计、经济等领域中得到了更加充分的发展。Bradley-Terry 模型与 Thurstone-Mosteller 模型[8] 是拟合成对比较法的两个主要模型，通过极大似然估计方法得到一个近似排名。

Zermelo[9] 在1929年利用Bradley-Terry方法给参与不完全循环赛的桥牌选手进行实力评估，提出了一个解极大似然估计的迭代算法。之后Bradley[10] 讨论了模型中使用的极大似然估计方法的在渐近线、假设检验、置信空间等多个问题。在模型的进一步推广上，luce、Rao，Kupper、Agresti[11,12,13]等完成了很多优秀的工作。

胡超欧与刘知青[14] 在《比赛评价模型在新闻信息排序中的应用》一文使用了Bradley-Terry 模型与 ELO 算法，通过对两个算法改进将其应用在新闻标题的排序问题上。

陈冬进与刘建平[15] 在《基于Bradley-Terry模型与层次分析法的一种综合评价方法》中提出了一种新的评价方法，先对各指标建立成对数据的Bradley-Terry模型，接着通过层次分析法求出它们的排序权重，最后得出评价的结果。

## 3.方法描述

### 3.1基于比较关系的数据增强

当数据为成对句子对(A,B)的情况下，一个简单的增强方法是根据比较关系来进行数据增强。即已知(A,B)代表A>B,又知道(B,C)代表B>C，我们可以得出一个新的句子对关系(A,C)。如果将单个句子作为点，比较关系(A,B)构建一条从B到A的有向边，由此定义可由数据集构建一个大的有向图，代表了句子之间的比较关系。在该有向图中可能存在一个或多个弱连通子图，子图是一个弱连通图（即至少有一对结点不满足单向连通，但去掉边的方向后从无向图的观点看是连通图），子图之间没有边连接。子图如[1](#fig:map)所示：

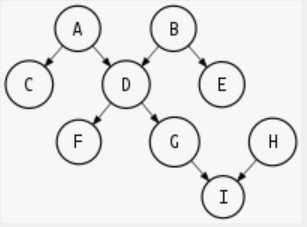


图1: 一个有向无环图

由于是弱连通图，想要根据比较关系进行采样，我们要从每个入度为0的节点出发，如图中A节点，沿着有向边深搜整个图，可以得到数个比较关系链。如[2](#fig:linknode)所示。

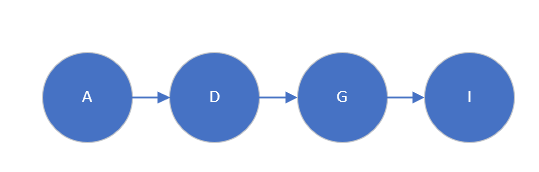


图2: 一条比较关系链

对比较关系链去重后，可以按分数从低到高进行排列组合采样。如D->C->B->A，可以采样得到(C,D)、(B,D)、(A,D)、(B,C)、(A,C)、(A,B)。

### 3.2基于Bradley–Terry结果的比较关系采样

通过向注释者提供成对的文本并询问哪一个具有更高分数的比较方法，能得到具有更高一致性的数据。但是每个文本对仅使用一个注释者注释是不可靠的。为了减少单个注释者可能的标注错误，一般会选择两个及以上的注释者，当注释者之间的标注结果存在冲突时，根据少数服从多数原则，选择多数结果。3.1中的比较关系采样方法，只使用了最终的标注结果，而没利用标注中可能存在的冲突信息，事实上我们可以认为当注释者之间标注存在冲突时，往往意味着文本对之间文本分数比较接近。

Bradley-Terry 模型是一个体育比赛的统计模型，用几个参赛队（或运动员）两两竞技的胜负场次来估计每个参赛队的实力，进而预报任意两支参赛队交手时的胜负概率。如果将句子对之间的比较结果作为胜负次数，我们可以估算出每个句子的“实力”，根据这个分数我们可以得到一个更为合理的比较关系链。用同样排列组合方法我们可以采样得到更多数据。

己知有n个队伍以及其中任意两个队伍间的胜负次数，通过这些数据就可以预测出其中任意两个队伍下在一次比赛时胜负的情况。在已知的比赛胜负信息的情况下，假设每个队伍的强度为，队伍打败的次数为，便有下面的对数最大似然。

关于上式的求解问题有着一个非常常见的迭代算法，令是一个随机选择的初始解，那么就有

用归一化使。Ford[16] 证明了在该模型条件下，极大似然估计的存在性和唯一性。

根据上述方法，可以得到所有句子的分数排行，据此构建一条新的比较关系链。如果要对该链进行采样时要注意，我们的采样结果可能高达条。(方法3.1中存在较大数量的比较关系链条数，即使排列组合结果数也远小于。为了尽可能使得每个句子都被采样到，我们可以对每个句子按比较关系链往后采样k次，数据数量为k\*n。如[3](#fig:sample_pool)所示

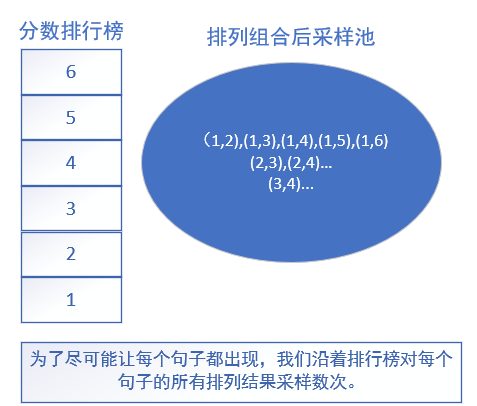


图 3: 简单采样示例

### 3.3改进采样方法

3.2中方法在原有的排序链基础上引入了Bradley–Terry算法，更好利用了比较结果中可能存在的冲突信息。但Bradley–Terry是一个近似方法，与3.1中由标签数据直接构建比较链不同，该方法并不能得到一定正确的比较关系链。相当于做了错误标注为数据带来了噪声，同时显然对于一条比较关系链a->b->c->d->e，将b->c标注错误为c->b的情况要比a->e错标为e->a的情况更有可能出现。简而言之，噪声发生在相邻的比较关系之间概率要远大于在间隔的比较关系之间。同时3.2方法在具体实施中存在两个问题。

首先是采样结果受随机种子影响较大。一条n个文本组成的比较链，对于排名较低的文本进行采样时沿图二中链进行采样，由于完全采样数量太大，每个节点选择k次，作为比较关系中分数较小的文本A，每次随机在A节点后选一个分值更大的文本B，构成（A,B）对。 在对B进行采样时，采样范围为n-rank(A)（其中rank（A）是A的排名）。当采样次数k<<采样范围时，采样结果受采样所使用的随机种子影响较大。选取的随机种子不同得到的结果差异也较大。整体采样质量受随机种子影响非常大。

其次是采样不公平，沿图二中链进行采样，作为比较关系中分数较小的文本A，与一个分值更大的文本B，构成（A,B）对。显然，一个排名高的文本，被采样的次数的期望要高于一个排名低的文本，因为该文本至少被采样k次，同时有可能被排名低的文本采样到（对于链尾的k个文本不进行采样，因为没有足够的文本来进行采样，n为文本数目，一般k<<n，该情况对结果影响较小）。

对于上述问题，我们提出了一种新的采样策略方法，分级采样方法。如[4](#fig:degreeSamplePool)所示

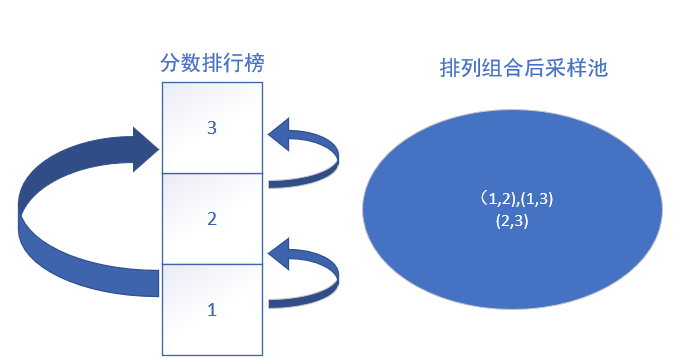


图4: 分级采样图

我们将原来的比较关系划分为数个连续区间，对于区间内的文本互相之间不进行采样，减少上述的可能出现的Bradley-Terry方法带来的噪声影响。采样只在不同区间的文本间进行。例如图中的区间1文本对区间2内文本进行采样，此时每次采样范围缩小到了区间大小，缓解了上述的采样受随机种子影响的情况。其次对于排名低的区间增大采样次数，排名高的区间减少或者不采样，一个简单方法是对总数为m的区间，排名为m的区间（即最低的区间）采样m-1次，对排名为1的区间不进行采样，也能有效缓解采样不公平的问题。

## 4.实验

### 4.1 数据集

2017年底，Civil Comments平台关闭，并选择将其平台上的约200万条公开评论保存在一个持久的开放档案中，以便研究人员在未来几年内理解并改善在线对话中的礼貌。Jigsaw赞助了这项工作，并对这些数据进行了扩展注释，由人类对各种有毒会话属性进行评分。有毒程度定义指的是让人觉得被冒犯，想离开聊天的程度。Jigsaw是来自谷歌的一个团队，目的是探索开放社会面临的威胁，并构建激发可扩展解决方案的技术。在本数据集中，提供了大约1.4万条评论。成对的评论被提交给标注人员，他们根据各自的毒性概念，将两个评论中的一个标记为更有害的。最后生产出一个包含条文本对的训练集和大小的验证集的toxic-dataset。

### 4.2 基线模型和Metrics

在本节中，我们将通过实现一些常用的模型体系结构，对toxic-dataset进行基准测试。这些模型的任务是预测给定评论的攻击性得分。我们对每个模型进行了5折交叉验证。

#### 4.3.1模型

Bert：我们对Bert-base进行了微调[17] 。我们在预先训练的模型中添加了一个包含线性层的回归头。我们使用margin-loss损失作为目标函数，批量大小为32，学习率为2e− 5(其他超参数与[17] 相同).我们使用AdamW优化器和线性学习率调度器，无需预热步骤。该模型经过3个epoch的训练。

HateBERT：HateBERT[18] 是bert经预训练的一个版本，用于英语中的辱骂性语言检测。HateBERT在RAL-E数据上进行了预训练，RAL-E是一个大型的英语Reddit评论数据集，数据内容来自因冒犯或仇恨而被禁止的社交平台或网站。在一些常用的辱骂性语言的数据集(OffensEval,AbusEval,HatEval[19,20,21])的结果上来看，HateBERT的表现优于通用的BERT模型。

LUKE：LUKE[22] 作者提出的模型是在BERT的MLM的基础上使用一种新的预训练任务来训练的。这项将给定文本中的单词和实体视为独立的标记，并要求输出它们的上下文化表示。作者还提出了一种实体感知的自我注意机制，它是transformer中的自我注意机制的扩展，并在计算注意力分数时考虑了token（单词或实体）的类型。LUKE是以Roberta作为基础的预训练模型，在多个命名实体识别数据集上达到了最佳效果。考虑到毒性分值是和某些不文明用词高度相关，所以我们选择了在实体识别上最优的模型LUKE。

#### 4.3.2结构和Metrics

考虑到具体数据集，一个判断一对文本哪个毒性更高的模型并不能直接用到实际中，尤其是toxic-dataset经常相关的内容审查领域上（内容审查的对象往往是单文本形式）。为了能直接对单文本的排序质量进行评估，我们通过[5](#fig:train_struct)所示结构来直接对单文本打分。

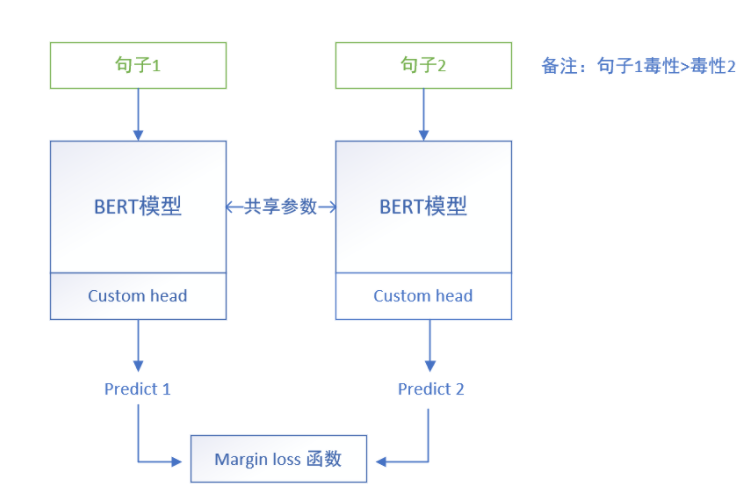


图5: 训练模型结构

模型将文本向量作为输入，输出一个毒性分数。模型接受两次输入分别为两个文本对中的文本，得到两个分数predict1和predict2，通过margin-loss函数计算损失，然后使用反向梯度传播更新参数。 Margin-loss函数是一个经典的Ranking loss函数。不像其他损失函数，比如交叉熵损失和均方差损失函数，这些损失的设计目的就是学习如何去直接地预测标签，或者回归出一个值，又或者是在给定输入的情况下预测出一组值，这是在传统的分类任务和回归任务中常用的。ranking loss的目的是去预测输入样本之间的相对距离。

为了进行排序质量评估，首先提取验证集上所有单个文本，让模型在验证集上得到其中所有单文本的排序关系，比对根据验证集上标注的文本对的大小关系。排序与标签吻合记录为正确，计算总的准确率。

为了评估随机种子对方法3.2与3.3的影响，实验中选取了24、100、200、1024、2022等随机种子重复进行实验，检验随机种子设置对抽样影响。方法3.2采样次数K=4，3.3中划分次数m=8，保证两种方法产生额外数据量一致左右。

### 4.4结果和分析

表1实验结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | 原数据 | 比较 | B-T采样 | 改良方法 | |
| Bert | 0.788 | 0.793 | 0.801~0.808 | 0.805~0.808 |
| HateBert | 0.818 | 0.819 | 0.8.38~0.846 | 0.841 |
| Luke | 0.831 | 0.830 | 0.840~0.862 | 0.855~0.857 |

从实验结果可以看出，三种增强方法对模型在验证集上的表现或多或少都有一定的帮助。其中3.1方法效果最差，可能有两个方面的原因。 1. 数据量，3.1方法获取的额外数据量是固定的，且由于数据中文本对比关系构成的图较为稀疏，该图能分成个互相之间不连接的子图，相应的能找到的链的长度相对较小，最终只提供了数据(法3.2和3.3均能提供数据)。 2. 额外数据信息量，正如我们前面所示，3.1方法不能利用到标注中可能存在的冲突信息。其次对于一个(A,B)、(B,C)得出的新的句子对关系(A,C)，所包含的信息比前两者都要小，模型训练过程中可能并不能获取太多有用信息。

再对比方法3.2和3.3实验结果，显然3.2相比3.3有更好的上限但也同时有着更为糟糕的下限。在luke模型上3.2方法的最次和最优结果准确率之差能达到2%，这一切都只是改变了随机种子的结果。事实上，随机种子确实在很大程度上影响实验结果，我们可以通过反复实验来选取最佳的随机种子，但我们认为这在本质上是对验证集的一种过拟合，对现实中的新的数据没有意义，所以虽然3.3方法的上限相比而言要差，但下限更高，减少了随机种子的影响，具有更好的鲁棒性。

## 5.结论与展望

本文针对现有的一种的成对标注方法进行了利弊分析，优点是该方法使标注一致性提高，缺点是理论的标注次数达到了,实际标注数量远小于该数目。为了在标注中得到更多信息，以直接用于模型训练。本文在简单的比较法基础上提出了一种基于Bradley–Terry方法结果进行采样的数据增强方法，基于分层和排列的思想，使原方法存在的随机种子影响和采样不公情况得到改善。实验结果也证明在使用最新方法的模型上，精度平均值和鲁棒性都要好于原方法，同时原方法和改良方法效果都好于简单的比较法。基于Bradley–Terry的近似方法能有效的一个重要原因是成对的文本标注对比单一标注在标注出现偏差时有更高的容忍程度。事实上除了Bradley-Terry算法仍有许多近似比较方法可以实践，如Thurstone-Mosteller 模型[8] 。使用不同的方法在效果上会带来什么差别，在采样特点上有什么差别等，仍有许多可以探究之处。

参考文献：

[1] SCHUMAN H. PRESSER S. Questions and answers in attitude surveys: Experiments on question form, wording, and context[M]. Sage, 1996.

[2] ASAADI S. MOHAMMAD S. KIRITCHENKO S. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 505-516.

[3] KIRITCHENKO S. MOHAMMAD S M. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation[J]. arXiv Preprint arXiv:1712.01765, 2017.

[4] PIECH C. HUANG J. CHEN Z. et al. Tuned models of peer assessment in MOOCs[J]. arXiv Preprint arXiv:1307.2579, 2013.

[5] AGGARWAL C C. OTHERS. Recommender systems: 卷 1[M]. Springer, 2016.

[6] SHAH N B. BRADLEY J K. PAREKH A. et al. A case for ordinal peer-evaluation in MOOCs[C]//NIPS workshop on data driven education. 2013: 1-8.

[7] SALESSES P. SCHECHTNER K. HIDALGO C A. The collaborative image of the city: mapping the inequality of urban perception[J]. PloS One, 2013, 8(7): e68400.

[8] THURSTONE L L. A law of comparative judgment.[J]. Psychological Review, 1927, 34(4): 273.

[9] ZERMELO E. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung[J]. Mathematische Zeitschrift, 1929, 29(1): 436-460.

[10] BRADLEY R A. Incomplete block rank analysis: On the appropriateness of the model for a method of paired comparisons[J]. Biometrics, 1954, 10(3): 375-390.

[11] LUCE R D. Several unresolved conceptual problems of mathematical psychology[J]. Journal of Mathematical Psychology, 1997, 41(1): 79-87.

[12] RAO P. KUPPER L L. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model[J]. Journal of the American Statistical Association, 1967, 62(317): 194-204.

[13] AGRESTI A. Categorical Data Analysis Wiley: New York[J]. 1990.

[14] 胡超欧. 刘知青. 比赛评价模型在新闻信息排序中的应用[J]. 软件, 2011, 32(10): 24-25.

HU C O, LIU Z Q. Application of competition evaluation model in news information ranking [J]. Computer Engineering & Software. 2011, 32(10): 24-25.

[15] 陈冬进. 刘建平. 基于 Bradley-Terry 模型与层次分析法的一种综合评价方法[J]. 统计与决策, 2013(3): 81-83.

CHENG D J, LIU J P. A comprehensive evaluation method based on Bradley Terry model and analytic hierarchy process [J]. Statistics & Decision. 2013(3): 81-83.

[16] FORD JR L R. Solution of a ranking problem from binary comparisons[J]. The American Mathematical Monthly, 1957, 64(8P2): 28-33.

[17] DEVLIN J. CHANG M W. LEE K. et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv Preprint arXiv:1810.04805, 2018.

[18] CASELLI T. BASILE V. MITROVIĆ J. et al. Hatebert: Retraining bert for abusive language detection in english[J]. arXiv Preprint arXiv:2010.12472, 2020.

[19] ZAMPIERI M. MALMASI S. NAKOV P. et al. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)[J]. arXiv Preprint arXiv:1903.08983, 2019.

[20] CASELLI T. BASILE V. JELENA M. et al. I feel offended, don’t be abusive! implicit/explicit messages in offensive and abusive language[C]//Language Resources and Evaluation Conference. The European Language Resources Associatio, 2020: 6193-6202.

[21] BASILE V. BOSCO C. FERSINI E. et al. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter[C]//13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2019: 54-63.

[22] YAMADA I. ASAI A. SHINDO H. et al. Luke: deep contextualized entity representations with entity-aware self-attention[J]. arXiv Preprint arXiv:2010.01057, 2020.

1. 收稿日期：xxxx−xx−xx；修回日期：xxxx−xx−xx

   通信作者：小5号宋体，段落最小值12磅，

   基金项目：中文，段落最小值12磅，

   Foundation Items: 英文

   线条 [↑](#footnote-ref-1)