实验报告　金融新实体发现

I　问题定义

随着互联网技术的不断发展，如何对飞速增长的互联网文本中包含的知识进行高效可靠的挖掘并进行组织，已成为自然语言处理 (NLP) 和信息抽取研究中的重要目标。互联网中各种命名实体及之间的关系纷繁复杂，单纯地以人工和经验的方式进行相关知识的获取及组织已经远远不能满足人们的使用要求。为此，知识的自动获取逐渐地成为文本处理的重要课题。针对这一问题，金融实体识别方案的建立将极大提高金融信息获取效率，从而更好的为金融领域相关机构和个人提供信息支撑。

命名实体识别（Named Entity Recognition，简称 NER），是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。图1 展示了从英文文本中进行实体识别的一个例子。

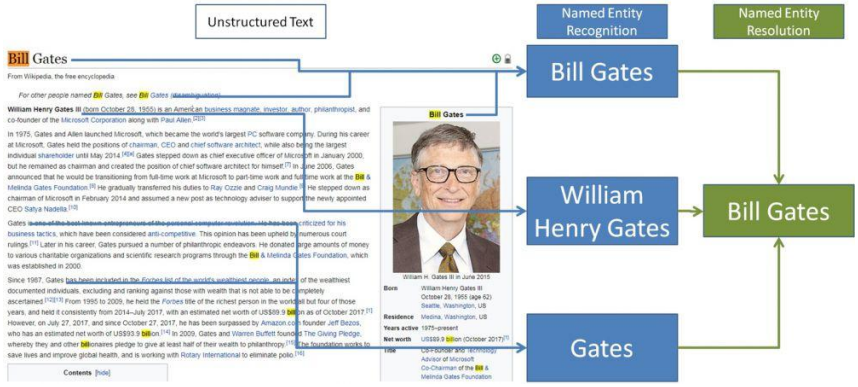


图 1　命名实体识别实例

NER 通常包括两部分：（1）实体边界识别；（2）确定实体类别（人名、地名、机构名或其他）。实体边界识别指从一个句子中分辨出一个个的词，确定实体类别则是指将分辨出来的词归类到不同的类别中。中文命名实体识别主要有如下几种方法，适用于不同的场景：

1. 基于规则和词典的方法。基于规则的方法多采用构造规则模板，以模式和字符串相匹配为主要手段。显而易见，对于新出现的实体，想要设计通用的字典进行匹配几乎是不可能的。

2. 基于统计的方法。传统统计机器学习的方法有很多，常见的包括 HMM, SVM, CRF 等。这些方法比较依赖于语料库。

3. 基于深度学习的方法。深度学习已经在 NLP 的很多领域取得了巨大的成功。尽管算法面临着诸多挑战，如计算资源开销大、依赖数据量及其标注，但模型的性能较之过去有了很大的提升。

本次实验基于 BDCI 2019 的相关赛题，借鉴了 Google Bert 模型。实验报告将围绕数据集分析、算法实现以及实验结果等这几部分展开。

II 数据集分析与处理

2.1 数据集简介

实验处理的是互联网金融新实体数据来自 BDCI 官方提供的赛题数据，数据类型是中文自然语言处理文本，主要是自动爬取的金融网络文本，包括标题和内容，并进行了简单的标注和转换。数据规模上，训练集数据量和测试集数据量都是一万条文本数据。训练数据全部以 csv 格式给出。其文件是一个字符序列。每条数据包括标识号（id）、文本标题（title）、文本内容（text）、未知实体列表（unknownEntities）。每条数据的内容如表1 所示。

字段类型	类型	描述	数据集实例
id	String	数据 ID	2e6a5020
title	String	文本标题	人人都在用手机赚钱
text	String	文本内容	比如做微商，做代理……
UnknownEntities	String	未知实体列表	趣投美元

表 1 数据格式说明

2.2 数据集缺陷

由于数据是直接从网络文段中自动爬取的，只是经过了简单的标注和处理，所以会存在一些问题。与所有传统的数据挖掘 (Data Mining) 任务相似，该数据集也并非尽善美。经过初步的预分析，结果表明数据在如下方面有进一步完善的可能性：

(i) 存在 html 标记语言、IMG 转换标记、非文字字符与混乱字符以及大量重复的字符。如图2(a) 所示。

(ii) 部分数据内容相似性度高，甚至完全重复。很多数据标题和内容也是高度一致的。如图2(b) 所示。

(iii) 标签混淆文本，缺乏实质性意义。如图2(c) 所示。

(iv) 人工标注存在问题。如一些知名企业仍标注为未知实体。如图2(d) 所示。

2.3 数据处理

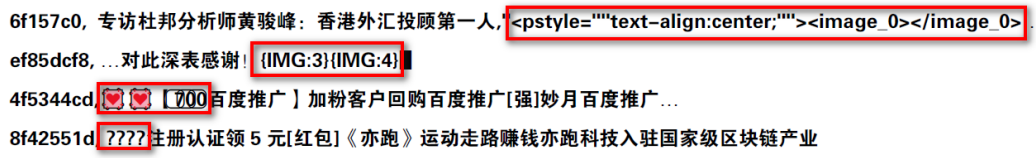
针对上述存在的问题，分别对应进行了如下处理：

(i) 用正则表达式进行模式匹配，删除脚本标记语言、特殊符号、重复符号等。

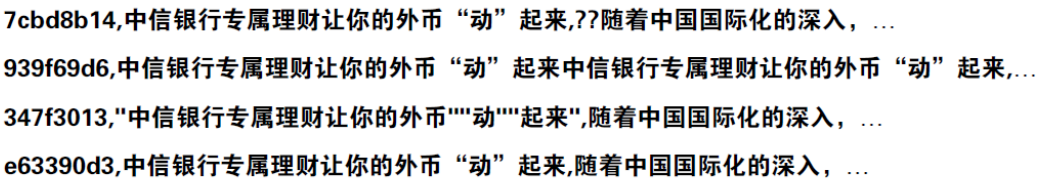
(ii) 通过简单的文本相似性度量方法，剔除重复数据。对于标题和文本内容重叠的，舍弃标题不做处理。

(iii) 找出标签所在上下文模式，剔除标签对文本带来的干扰。

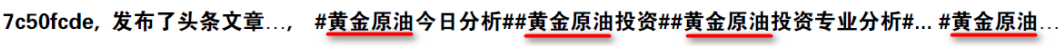
(iv) 对实验中观察到的有问题的数据，进行一部分重新标注。



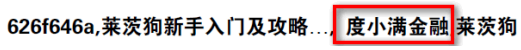
(a) 符号问题



(b) 重复文本数据



(c) 标签混入文本



(d) 标注问题

图 2 数据集部分问题图示

在上述清洗数据的基础上，还会对文本进行多余空格的剔除。考虑到清洗过的数据以及一部分原始数据是没有标题的，对标题和文本内容进行合并处理。实验中因为用到了 Bert 模型，所以用额外的脚本将处理过的数据改写成相应的格式。

III 算法与模型

3.1 Bert 模型框架

在参考文献 [1][2] 的基础上，我们选用 Bert 作为预训练模型。与之前的方法相比，该模型性能优越。报告 [1] 中提出，Bert 在机器阅读理解顶级水平测试 SQuAD1.1 中表现出惊人的成绩：全部两个衡量指标上全面超越人类，并且还在 11 种不同 NLP 测试中创出最佳成绩，包括将 GLUE 基准推至 80.4%（绝对改进 7.6%），MultiNLI 准确度达到 86.7% （绝对改进率 5.6%）等。BERT 模型是 NLP 领域发展的重大突破。

BERT 全称 Bidirectional Encoder Representations from Transformers。BERT 的新语言表示模型，它代表 Transformer 的双向编码器表示。与最近的其他语言表示模型不同，BERT 旨在通过联合调节所有层中的上下文来预先训练深度双向表示。因此，预训练的 BERT 表示可以通过一个额外的输出层进行微调，适用于广泛任务的最先进模型的构建，比如问答任务和语言推理，无需针对具体任务做大幅架构修改。图3 展示了文献 [1] 的结构。

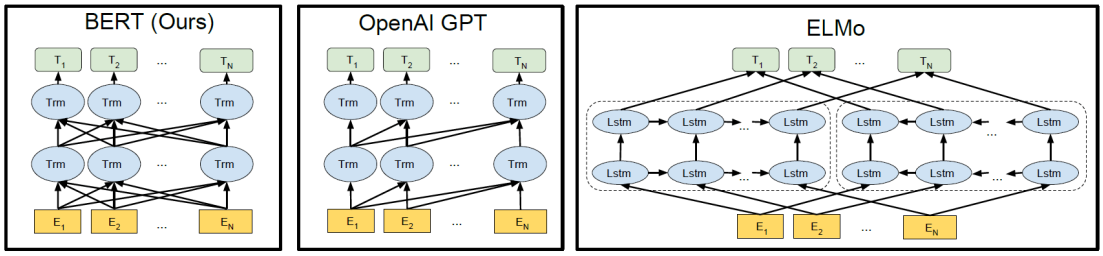


图 3 Bert 与其他模型的比较

BERT 的输入表示（input representation）能够在一个 token 序列中明确地表示单个文本句子或一对文本句子（例如，[Question, Answer]）。对于给定 token，A 其输入表示通过对相应的 token、segment 和 position embeddings 进行求和来构造。图4展示了这种输入表示。

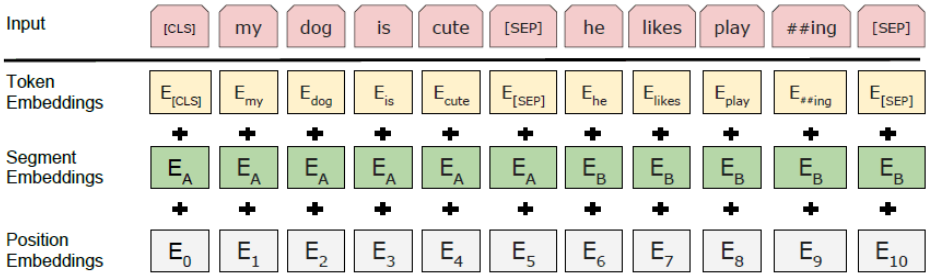


图 4 Bert 输入表示

3.2 环境与配置

互联网金融新实体发现是一个典型的 NER 问题。Google 已经提供了关于中文语料的 BERT 预训练模型，所以我们选择了 BERT 模型来进行微调训练来实现我们的结果。首先我们下载了 BERT 源码与关于中文的 bert 预训练模型（从 BERT−Base Chinese 下载模型）。进行相关准备后，我们编写自定义的 BERT_NER.py 文件，参考 bert 源码中的 run_classifier.py。这一过程中，主要是定义 DataProcessor 和 NerProcessor 类，还有修改其他一些地方的引用关系，这是因为 bert 源码里面没有我们需要的专门的 NER 识别的操纵代码。编写时也参考了网上教程的改写方法。

在预训练的基础上，我们用处理后的数据集进行 fine-tuning，训练 100 个 epoch。代码已经上传到 Github 存储库中，参数设置在 json 配置文件中，模型则上传到 OneDrive 上。

IV 结果分析

4.1 模型评估

由于测试集没有标注数据，线上结果也未来得及及时更新，我们在报告中最终进行离线评估。具体的做法是将原训练数据分出一部分作为测试集并进行重新训练。按照官

方给定的指标，分别计算准确率，召回率与相应的 F Score。计算公式为

$$
\begin{aligned}
Precision &= \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i}, \\
Recall &= \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i}, \\
F\ Score &= \frac{1}{1/Precision + 1/Recall}.
\end{aligned}
$$

其中，n 为样本总数，TP_i 是第 ii 条文本正确识别未知实体的数量，FP_i 表示第 i 条文本中误识别为实体的数量，FN_i 表示第 i 条文本中未识别出的未知实体数量。做评估时保留 1k 条数据，结果如表2 所示。

Samples	Precision	Recall	F Score
1000	0.2634	0.2252	0.2409

表 2 评估结果

可以看到，最终的得分结果从指标上来说不尽人意。即便是得分比较高的队伍，线上得分也只在 0.3 至 0.4 之间。我们将一部分结果可视化，以作分析。结果如图5 所示。

图5(a) 展示了一个典型的相似实体误判的结果。“河南多家公司”与“河南高迈公司”比较相近，模型未能很好地作出区分。这导致了 Precision 指标的下降，而要用后处理来处理这样的结果，仍面临一定的困难。

图5(b) 则是识别不精准的一个实例。与上一种情形相比，这种问题出现的频次相对少一些。在识别出的命名实体中，该例遗漏了“公司”两字，猜测可能是命名实体较长所致。这种情况下，在实际应用时不会带来很大影响，但 Precision 和 Recall 指标则会变成 0。

图5(c) 展示了误识别的一个例子。该例中的实体实际上是一个金融概念，但模型误识别为金融实体。这样的结果也很难作修正，因为在中文语境下，很多概念与实体是可以混淆的。

图5(d) 是典型的漏识别例子。注意到模型单独将标题断句为“文本”加以识别。这可能是漏识别的原因所在。在模型预测阶段，除了上述的短句分句外，还有很大一部分只处理了文本的前 128 个字符。目前 NLP 的深度学习模型对长句的处理性能较差，所以对句长会有限制。由此带来的的漏判问题，我们还没有做很好的处理。

4.2 模型展望

文献 [2] 提出的 Albert 模型通过 Embedding 的处理和参数共享策略，降低了模型的规模，缩短了训练时间。如果改用 Albert 模型，有希望在同样的时间内进行更多的尝试，有潜力进一步提升性能。此外，由之前的结果分析，如果能通过一个混合模型对结果进行后处理，毫无疑问可以将指标提升很多。但是，这一模型本身也需要进行精心设计，目前没有找到比较有力的解决方案。在数据层面，可以通过进一步预处理，来解决 Bert 模型对长句的制约。

ID: 8bccb7bc,
Entities: 河南高迈运输公司
Results
????河南高迈运输公司骗人套路，河南多家运输公司以保证货源，保证司机月收入一万五为由骗人！骗子公司有保护伞，政府部门个别人在做他们的保护伞！,????河南高迈运输公司骗人套路，河南多家运输公司以保证货源，保证司机月收入一万五为由骗人！骗子公司有保护伞，政府部门个别人在做他们的保护伞！

(a) 相似识别

ID: 0f74653b,
Entities: 湖南快线文化传媒有限公司
Results
本人曾是湖南快线文化传媒有限公司长沙市天心区五一广场平和堂商务楼 2608-2609 号的合伙人，因合作过程中发现其主要负责人阳光(实际叫詹*柱）有违规经营行为本人劝其关门未果欲退出要求其签散伙协议遭到其拒绝，本人无奈出此下策，本人电话：130****9158 请跟多受害者与我联系，集体劝其早日关门

(b) 不精准识别

ID: 0f032057,
Entities: None
Results
{IMG:1}SERO-超零币简介 SuperZero（SERO）是一个支持图灵完备智能合约的隐私数字货币，同时也是一个允许开发者自行发布匿名数字资产的隐……喜欢我们的小伙伴请扫码加好友哟{IMG:7}

(c) 错识别

ID: 7cbe1507
Entities: TK 唐卡拆分盘
我们该如何玩转互联网金融常青树 TK 唐卡拆分盘?

(d) 漏识别

图 5 部分可视化结果问题图示

参考文献

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] L. Zhenzhong, C. Mingda, G. Sebastian, G. Kevin, S. Piyush, and o. Radu, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1810.04805*, 2019.

V 附录

Our code is available here.

The model can be downloaded from here.

(Note: You may have to bypass GFW to get access to OneDrive.)

姓名	学号
岳锴	PB16060757
廖云铭	PB16060687
左琦	PB16061247
丁可	PB16061050

表 3 团队成员