面向中文短文本的多因子融合实体链指研究

吕荣荣，王鹏程，陈帅

小米人工智能实验室，北京，中国
{lvrongrong, wangpengcheng, chenshuai3}@xiaomi.com

Abstract. 实体链指，也称实体链接、Entity Linking或EL，即对于给定的一个文本，EL将其中的指称项（Mention）与给定知识库中对应的实体（Entity）进行关联。相比于长文本拥有丰富的上下文信息来辅助实体的消歧消解并完成链指，中文短文本的实体链指存在很大的挑战。本文针对百度发布的面向中文短文本的实体链指任务，设计的多因子融合实体链指模型。首先采用了预训练的BERT来对短文本中的指称项进行类别预测，利用预测的类型构建一个NIL实体，和其他候选构成完备候选实体集，然后对每一个候选实体进行多方位的特征因子抽取，利用一个多层感知机将多个特征因子融合打分，最后根据每一个候选实体和文本的关联分数进行排序，选择分数最高的候选实体作为实体消歧预测结果。在CCKS 2020中文短文本的实体链指这个评测任务上，本文提出的方法在最终的评测数据上F1达到0.89540的分数，取得了第一名的成绩。

Keywords: BERT，实体链接，实体消歧，特征融合

1 引言

近年来，随着深度学习的重燃以及海量大数据的支撑，NLP领域迎来了蓬勃发展。知识图谱作为机器大脑中的知识库也发展迅猛，如百度、阿里、小米等科技公司都拥有数亿实体、千亿事实，具备丰富的知识标注与关联能力的中文知识图谱。实体链指是NLP领域的基础任务之一，也是知识图谱关键技术之一，对许多自然语言处理应用如智能问答、信息检索、内容推荐等任务都能产生积极的助力作用，能让机器更好的理解文本。

传统的实体链指任务主要是针对长文档，主要利用词袋模型计算指称项所在上下文文本与候选实体所在文本之间的文本相似度，进而用文本的相似度来衡量实体间的相似度，长文档拥有丰富上下文信息能辅助实体的歧义消解并完成链指。相比之下，针对中文短文本的实体链指存在很大的挑战主要原因如下：

（1）口语化严重，导致实体歧义消解困难；

（2）短文本上下文语境不丰富，须对上下文语境进行精准理解；

（3）相比英文，中文由于语言自身的特点，在短文本的链指问题上更有挑战。

相比于去年的比赛，百度CCKS2020年的实体链指任务更专注于中文短文本场景下的多歧义实体消歧技术，增加了对NIL实体的上位概念类型判断，增加多模任务场景下的文本源，同时调整了多歧义实体比例。针对这些问题，本文设计了一个多因子融合实体链指模型。首先采用了预训练的BERT来对短文本中的实体进行类别预测，利用预测类型构建一个仅包含类型特征的实体称为NIL_type实体，和知识库中其他可以检索到的实体构成完备候选实体集，确保文本中的给定的指称项都能有一个正确的链接实体。然后对每一个候选实体进行多方位的特征因子抽取，特征因子抽取的抽取包括上下文相关特征的抽取和上下文无关特征的抽取，上下文相关特征包括文本上下文和候选实体描述的相似度计算，多个指称项之间的关联度计算等，上下文无关特征包括实体的流行度、实体的类型等，将这些特征因子利用一个多层感知机模型进行融合打分，预测每一个候选实体和文本的关联分数。最后对这些分数进行排序，选择分数最高的候选实体作为实体消歧预测结果。

本文主要的创新点如下：

（1）利用标记符在文本中直接标记出指称项的位置，从而可以利用通用的BERT计算语义相似度的模型完成实体消歧，大大节省了工作量；

（2）利用指称项类型预测，构建NIL_type实体，解决无链接指代预测问题。

2 相关研究

2.1 预训练语言模型

目前，预训练已经成为了自然语言处理领域最常用的技术之一，高质量的预训练模型能够为下游任务带来显著的提升。BERT[1]预训练模型，以Transformer Encoder为骨架，以屏蔽语言模型（Masked Language Model）[2]和下一句预测（Next Sentence Prediction）这两个无监督预测任务作为预训练任务，用英文Wikipedia和Book Corpus的混合语料进行训练得到。在海量单预料上训练完BERT之后，便可以将其应用到NLP的各个任务中了。

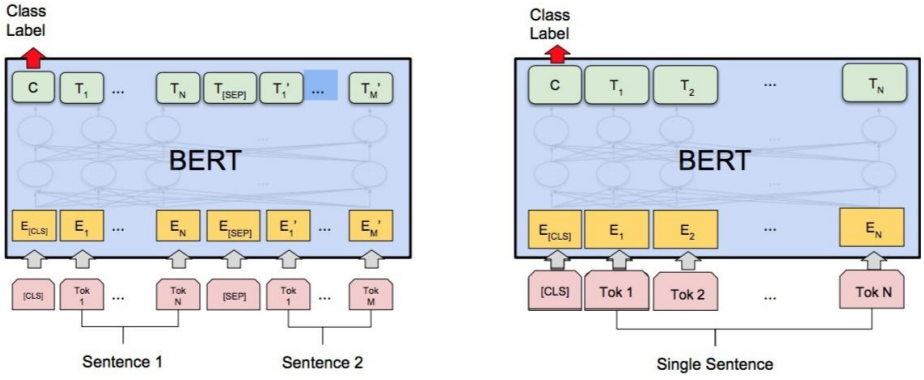

图. 1. BERT应用到下游任务

图1展示了如何将BERT应用到下游任务中，它们只需要在BERT的基础上再添加一个输出层便可以完成对特定任务的微调。图1左侧可以完成基于句子对的分类任务，图1右侧可以完成基于单个句子的分类任务。其条件概率表示如公式(1)，C其中是BERT输出中的[CLS]符号，W是可学习的权值矩阵。

$$P = softmax(CW^T) \tag{1}$$

BERT取得巨大成绩后，几大预训练模型轮番登场。ERNIE[3]是百度提出的语义表示模型，同样基于Transformer Encoder，相较于BERT，其预训练过程利用了更丰富的语义知识和更多的语义任务。ERNIE2[4]又提出了一个持续学习的机制，模仿人不断学习的形式，并且是多种任务不停交叉学习。RoBERTa[5]改变了BERT预训练的方法，利用动态掩码，保证每次输入到序列的掩码都不一样，实验表明动态掩码确实能提高性能。另外RoBERTa还移除了Next Sentence Prediction，用了更大batch size，更多的数据和更长时间的训练。Whole Word Masking[6]主要更改了原BERT预训练阶段的训练样本生成策略。原有基于WordPiece的分词方式会把一个完整的词切分成若干个子词，在生成训练样本时，这些被分开的子词会随机被掩码。在Whole Word Masking中，如果一个完整的词的部分WordPiece子词被掩码，则同属该词的其他部分也会被掩码，即全词掩码，相关的模型包括BERT-wwm-ext，RoBERTa-wwm-ext，RoBERTa-wwm-ext-large等。

2.2 实体链接

实体链指的主要目标是识别上下文中的名称指代哪个现实世界中的实体。具体而言，实体链指是将给定文本中的一个指称项映射到知识库中的相应实体上去，如果知识库尚未收录相应实体，则返回空实体。最近有不少这方面的优秀工作。Ganea O E& Hofmann T.[7]开创性地在EL中引入Entity Embedding 作为信息，利用Attention机制来获得Context的表征，通过实体间的一致性，和Mention到Entity的LinkCount先验概率联合消歧。Le, P., & Titov, I. [8]不仅仅考虑Local/Global的影响，同时将实体的关系也考虑进Embedding中，对Entity，Mention，Relation元组进行Embedding，借用ESIM思想进行对多关系加权处理，并使用网络进行匹配操作。Raiman JR&Raiman OM [9]认为当我们能预测出实体 Mention 的 Type，消歧这个任务就做的差不多了，主要利用Type System、Type Classifier和LinkCount来达到消歧的目的。 Sil et al.[10]不但利用包含Mention的句子和Wiki页面的相似度，还加入了细粒度的相似度计算模型，将几种相似度作为神经网络的输入，避免了句子中不相关单词对Mention消歧的影响。综合来看，实体链接不仅要考虑 Text 的文本信息、KB 的信息、消歧后的一致性，还需要根据具体的业务场景采用不同的方案，需要灵活的运用LinkCount、Context、Attributes、Coherence这四大特征。

3 实验方法

3.1 指称项分类

指称项分类是主要基于BERT模型，输入数据文本，指称项的起始位置。输入文本，经过BERT模型编码，取CLS 位置的特征向量、指称项开始和结束位置对应的特征向量，三个向量拼接，经过全连接层，最后Softmax激活得到指称项的类别概率分布。模型结构如图2，其中优化主要改进的点包括：

（1）二次训练：训练集合中非NIL部分的分类数据与NIL的分布不同，直接与NIL部分的数据一起训练会导致模型整体预测NIL实体的准确率下降，而直接用NIL部分的数据训练则有些训练数据较少的类会训练的不充分。所以我们采用二次训练的方法，第一次的时候使用了训练集中非NIL的部分，训练两个Epoch，然后再在这个基础上去训练NIL部分。

（2）对抗学习：对抗训练是对抗防御的一种，它构造了一些对抗样本加入到原数据集中，希望增强模型对对抗样本的鲁棒性。我们在模型训练的时候加入了对抗学习，所使用的对抗学习方法是Fast Gradient Method（FGM）[11]。

（3）模型融合：本次使用了4个BERT预训练模型。模型融合的方法是使用多折的方法训练了一个基于MLP的分类模型。

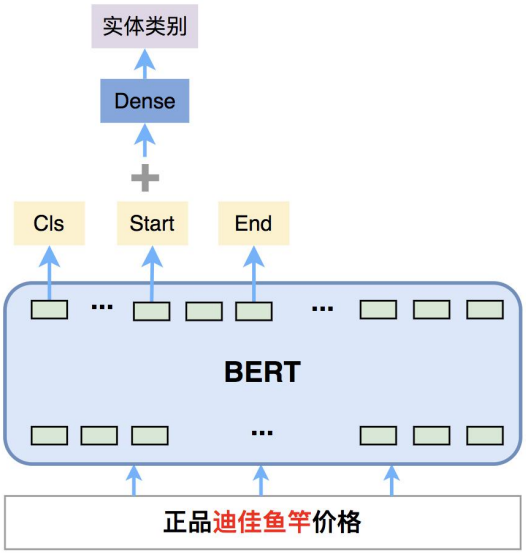

图. 2. 实体分类模型图

3.2 候选实体获取

利用实体的Alias字段生成Mention和实体的映射表，实体的Alias的属性值即为该实体的Mention，包含该Mention的所有实体组成候选实体集合。在候选实体获取时，从Mention和实体的映射表中，取出该Mention的候选实体集合，然后指称项的类别构成的NIL实体组成完备候选实体集。这样组成的完备候选

实体集中，必有一个正确的实体和文本中的指称项关联。训练时，指称项的类别来自标注文本中Kb_id对应的实体类型，预测时，指称项的类别由3.1部分描述的指称项分类模块预测得到。

为了后续使用方便，我们将完备候选实体集中的实体属性进行拼接，处理成实体的描述文本。由于Type字段，义项描述和摘要字段的信息重要且占比较大，描述文本中都按照Type、义项描述、摘要和Data中其他Predicate、Object对的顺序进行拼接。例如文本"永嘉厂房出租"中“出租”对应的候选实体Id和描述文本为[["211585", "类型：其他|简介：动词，收取一定的代价，让别人在约定期限内使用|外文名：rental|拼音：chū zū|解释：交纳租税|中文名：出租|举例：出租图书|日本語：レンタル|标签：非娱乐作品、娱乐作品、小说作品、语言、电影、字词"], ["304417", "类型：车辆|描述：辞源释义|简介：出租车，供人临时雇佣的汽车，多按里程或时间收费，也叫出租车|外文名：Taxi、Cab、Hackies|粤语：的士|台湾名：计程车|拼音：chūzūchē|中文名：出租车|新加坡名：德士|标签：交通工具、社会、生活"], ["NIL_Other", "类型：其他|描述：未知实体"]]，其中“211585”和“304417”为检索到的候选实体集合，NIL_Work为生成的候选实体，一起组成了“出租”在该文本下的完备候选实体集。

3.3 实体消歧

针对实体消歧任务，目前最常用的方法是将其视为二分类问题。对每一个候选实体进行多方位的特征因子抽取，将这些特征因子利用一个多层感知机模型进行融合打分，预测每一个候选实体和指称项的关联分数。最后对这些分数进行排序，由于我们在候选实体获取阶段，构建的是完备候选实体集，那么必有一个正确候选实体，所以在排序后选择Top1即可作为指称项的关联实体。

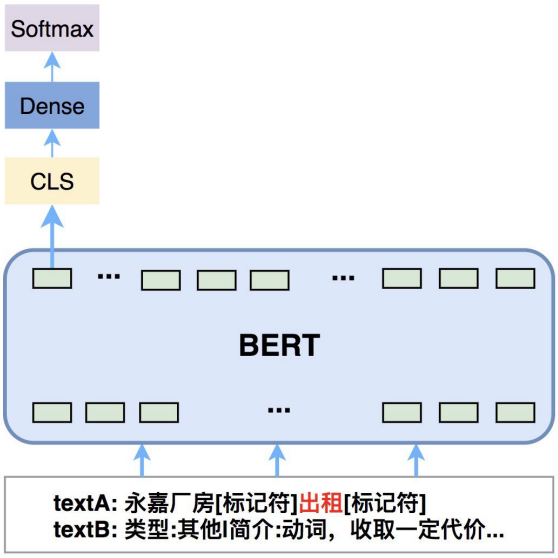

图. 3. 实体链接模型图

特征因子抽取的抽取包括上下文相关特征和上下文无关特征，其中上下文相关特征包括指称项和候选实体的关联概率计算，多个指称项之间的关联概率计算等，上下文无关特征包括实体的流行度、实体的类型等。这里起到关键作用的特征就是指称项和候选实体的关联概率。指称项和候选实体的关联概率和语义相似度计算的区别在于需要指明文本中待消歧的指称项。我们利用标记符在文本中直接标记出指称项的位置，指明待消歧的指称项。输入文本和候选实体描述文本，在文本的指称项开始和结束位置添加标记符，经过BERT模型编码，取CLS 位置的特征向量，经过全连接层，最后Softmax激活得到文本中指称项和候选实体之间的相关性。模型结构如图3所示。

另外我们在实体消歧模块也尝试加入对抗学习来提高模型的鲁棒性，其中对抗学习的方法是FGM。不同的BERT预训练模型抽取的特征不同，为了丰富特征，本模块采用了19个特征因子来从不同方面刻画指称项和候选实体的相关性 。这19个特征如下表1所示，分别为：

特征	备注
popularity	统计的标注数据中指称项映射到实体的关联概率
coherence	文本中其他指称项出现在候选实体描述文本中的概率
coherence2	统计的标注数据中其他指称项出现时指称项和候选实体的关联概率
type	候选实体的类型映射到指称项类别的预测概率
nil	候选实体是否为NIL实体
bert-base	bert-base预测的指称项和候选实体的关联概率
bert-base-rank	bert-base预测的指称项和候选实体的关联排名
bert-wwm-ext	bert-wwm-ext预测的指称项和候选实体的关联概率
bert-wwm-ext-rank	bert-wwm-ext预测的指称项和候选实体的关联排名
UER-base	UER-base预测的指称项和候选实体的关联概率
UER-base-rank	UER-base预测的指称项和候选实体的关联排名
bert-base-adv	加入对抗学习后bert-base预测的指称项和候选实体的关联概率
bert-base-adv-rank	加入对抗学习后bert-base预测的指称项和候选实体的关联排名
bert-wwm-ext-adv	加入对抗学习后bert-wwm-ext预测的指称项和候选实体的关联概率
bert-wwm-ext-adv-rank	加入对抗学习后bert-wwm-ext预测的指称项和候选实体的关联排名
UER-base-adv	加入对抗学习后UER-base预测的指称项和候选实体的关联概率
UER-base-adv-rank	加入对抗学习后UER-base预测的指称项和候选实体的关联排名
roberta_wwm_large_ext	roberta_wwm_large_ext预测的指称项和候选实体的关联概率
roberta_wwm_large_ext-rank	roberta_wwm_large_ext预测的指称项和候选实体的关联排名

表. 1. 实体消歧特征

特征因子融合的方法是使用多折的方法训练了一个MLP的模型。将所有数据集分成n份，不重复地每次取其中一份做测试集，用其他四份做训练集训练模型，训练得到n个模型。预测时，取n个模型的预测结果的平均值，作为预测结果。

4 实验结果

4.1 实验数据

CCKS 2020 中文短文本的实体链指比赛，限定在给定的标注数据和知识库中。标注数据集由训练集、验证集和测试集组成，给定的训练数据共70000条，

给定的验证数据共10000条，给定的第一阶段测试数据共10000条，给定的第二阶段测试数据供25000条，标注数据均通过百度众包标注生成，准确率95%以上。标注数据集主要来自于：真实的互联网网页标题数据、视频标题数据、用户搜索 Query。每条标注数据包含Text，Text_id和Mention_data字段，Mention_data里面包含连接的Mention，Offset以及Kb_id字段。知识库包含来自百度百科知识库的约39万个实体。知识库中的每个实体都包含一个 Subject_id (知识库 Id)，一个 Subject 名称，实体的别名，对应的概念类型，以及与此实体相关的一系列二元组< Predicate，Object>（<属性，属性值>）信息形式。知识库中每行代表知识库的一条记录（一个实体信息），每条记录为 Json 数据格式。

统计可得，给定的所有标注数据的Text长度不超过50个字符占99.99%，知识库中的实体包含文字不超过256个字符的占99.99%，其中每个实体均有Type字段和Data字段，Type字段值全部包含在给定的24个类别中，Data字段中包含约有6.8个< Predicate，Object>信息对，包含"摘要"信息对的实体占90.89%，包含“义项描述”信息对的实体占89.42%。

4.2 指称项分类

本次使用了4个不同的BERT预训练模型，分别为bert-wwm-ext、roberta_wwm_ext、ernie和roberta_wwm_large_ext。模型训练中使用二次训练的方法F1提升了约1%，使用对抗学习F1提升了约0.5%，模型融合后的Dev数据上F1值达到了90.02%。具体参数和结果建下表：

模型	参数	dev F1
bert-wwm-ext_1round	batch_size=128, length=80, opoch=15, lr=1e-5	87.50%
bert-wwm-ext_2round	round 1: batch_size=128, length=80, opoch=2, lr=5e-6; round 2: batch_size=128, length=80, opoch=15, lr=1e-5	88.42%
bert-wwm-ext_2round_adv	round 1: batch_size=128, length=80, opoch=2, lr=5e-6; round 2: batch_size=128, length=80, opoch=15, lr=1e-5	88.89%
roberta_wwm_ext_2round_adv	round 1: batch_size=128, length=80, opoch=2, lr=5e-6; round 2: batch_size=128, length=80, opoch=15, lr=1e-5	88.54%
ernie_2round_adv	round 1: batch_size=128, length=80, opoch=2, lr=5e-6; round 2: batch_size=128, length=80, opoch=15, lr=1e-5	88.23%
roberta_wwm_large_ext_2round_adv	round 1: batch_size=32, length=80, opoch=2, lr=5e-6; round 2: batch_size=32, length=80, opoch=15, lr=5e-6	88.74%
模型融合的结果		90.02%

表. 2. 实体分类参数以及结果

4.3 实体消歧

本次使用了4个不同的BERT预训练模型，分别为bert-base、bert-wwm-ext、UER-base[12]以及roberta_wwm_large_ext。特征融合的方法是使用多折的方法训练了一个MLP的模型。具体参数和验证数据集下结果如下表:

模型	参数	dev F1
bert-base	batch_size=64, length=256, epoch=3 lr=2e-5	87.90%
bert-wwm-ext	batch_size=64, length=256, epoch=3 lr=2e-5	87.91%
UER-base	batch_size=64, length=256, epoch=3 lr=2e-5	88.04%
bert-base-adv	batch_size=64, length=256, epoch=3 lr=2e-5	88.43%
bert-wwm-ext-adv	batch_size=64, length=256, epoch=3 lr=2e-5	88.14%
UER-base-adv	batch_size=64, length=256, epoch=3 lr=2e-5	88.38%
roberta_wwm_large_ext	batch_size=16, length=256, epoch=3 lr=1e-5	87.66%
19个特征融合结果		89.29%

表. 3. 实体消歧参数以及结果

5 结论

本文中，介绍了小组针对CCKS 2020中文短文本的实体链指任务设计的一个多因子融合实体链指模型。首先采用了预训练的BERT来对短文本中的实体进行类别预测。利用预测类型构建一个仅包含类型特征的实体称为NIL_type实体，和知识库中其他可以检索到的实体构成完备候选实体集。然后对每一个候选实体进行多方位的特征因子抽取，特征主要包括文本上下文和候选实体描述文本的相似度计算，多个指称项之间的关联度计算和一些实体的统计特征。将这些特征因子利用一个多层感知机模型进行融合打分，选择分数最高的候选实体作为实体消歧预测结果。我们在验证数据集上F1值为89.29%，在最终的测试数据集上取得了89.53%的成绩。

与此同时，本文还有一些值得探索的地方有待完善。比如没有充分利用其它指称项的侯选实体信息，对其他指称项信息的利用仅仅停留在名称层面。另外，可以利用一些特征，先对候选实体进行一次排序，选择排序前几的候选实体进行下一步的消歧，这样分层消歧在候选实体过多的情况下不仅可以提高准确率，还能提高消歧效率。

References

1. 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.
2. Taylor W L. “Cloze procedure”: A new tool for measuring readability[J]. Journalism quarterly, 1953, 30(4): 415-433.
3. Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.
4. Sun Y , Wang S , Li Y , et al. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5):8968-8975.
5. Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
6. Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese bert[J]. arXiv preprint arXiv:1906.08101, 2019.

7. Ganea O E, Hofmann T. Deep joint entity disambiguation with local neural attention[J]. arXiv preprint arXiv:1704.04920, 2017.
8. Le P, Titov I. Improving entity linking by modeling latent relations between mentions[J]. arXiv preprint arXiv:1804.10637, 2018.
9. Raiman J, Raiman O. Deeptype: multilingual entity linking by neural type system evolution[J]. arXiv preprint arXiv:1802.01021, 2018.
10. Sil A, Kundu G, Florian R, et al. Neural cross-lingual entity linking[J]. arXiv preprint arXiv:1712.01813, 2017.
11. Goodfellow, Ian J, Shlens, Jonathon, and Szegedy, Christian. Explaining and harnessing adversarial examples.International Conference on Learning Representations(ICLR), 2015.
12. LZhao Z, Chen H, Zhang J, et al. UER: An Open-Source Toolkit for Pre-training Models[J]. arXiv preprint arXiv:1909.05658, 2019.