多因子融合的实体识别与链指消歧

祝凯华[1,2]，戴安南[1,2]，范雪丽[1,2]

[1] 上汽集团人工智能实验室，上海，中国
[2] 赛可智能科技（上海）有限公司， 上海，中国
zhubloom@gmail.com {daiannan,fanxueli}@saicmotor.com

Abstract. 实体识别与链指消歧，又称为Entity recognition和Entity linking，是自然语言处理领域的基础任务之一。针对百度发布的面向中文短文本的实体识别与链指比赛数据集，本论文首先采用了预训练的Bert来对短文本中的实体进行提取，然后根据提取出的实体，采用DeepType来预测实体类型信息，DeepMatch对实体的上下文和知识库进行文本匹配，最后用DeepCosine来结合知识库实体向量的预测及其他数值特征，比如流行度等弱消歧模型进行融合进而可以产生一个非常强的实体消歧预测结果。

Keywords: Bert, DeepMatch, DeepType, 模型融合.

1 引言

为了更好的让机器理解文本，机器常常需要识别文本中的实体，同时将文本中的实体与对应的知识库中的实体一一对应。知识库中的同名实体往往很多，因此就需要根据一些模型去做一些实体链指消歧工作。

在整个实体识别与链指消歧的过程中，常见的是把这个任务分成两部分，即先进行实体的识别，然后再进行实体的消歧[1-3]。最近也有部分工作强调要用端到端的方式统一两个任务[4]。最近基于语言模型的预训练模型变的越来越受欢迎，比如Bert[5]，XLnet[6]等等。这种通过大数据预训练的方式产生的语言词汇向量表征相比于传统方法前进了一大步。因此基于预训练模型的实体识别结果也提高了很多。得益于预训练模型强大的实体识别能力，本文因此采用两步走的方式来进行实体识别和链指消歧。因为实体识别的准确率足够高，因此对后面的消歧结果产生的False Positve样本影响会小很多，同时可以降低联合模型的计算空间。

命名实体识别任务多在识别文本中的事物的名称，例如人名、地名和机构名。本文主要在互联网文本领域下处理命名识别，比如识别电影名称、书名等等。以Bert预训练模型为基础并引入CRF（条件随机场）从文本中提取出标注样本的线性空间转率概率。Bert模型采用了最新的参数优化方案[7]，通过这样迁移权重和在训练样本微调的方式训练，最后只需要两轮训练模型就达到最优效果。Bert结合CRF的实验也远远超过了传统的lstm+crf的实验结果。尽管采用经典字向量模型可以手动设计很多特征，比如pos特征，词特征等等，这些特征确实帮助模型达到更好的输入表征效果。但是Bert等超过规模预训练的方式得

到的字向量表征在实验中比传统精细设计方法的效果更好，而且模型结构设计更加简便。因此未来的深度学习模型极有可能都是建立在预训练语言模型基础上构建。

实体链指消歧是指在知识库中找到候选的正确实体描述。百度CCKS2019数据集多为互联网搜索文本。在这些文本中出现了大量的作品名称，这些作品有可能是小说，有可能是改编后的电影或者电视剧，如表1所示。实体链指的目的就是根据上下文找出最有可能的知识库实体。最近有不少这方面的优秀工作。比如Phong Le[8] 强调了上下文其他实体对该实体消歧的帮助是很重要的。Jonathan Raiman[9]则依靠建立DeepType的系统来达到消歧的目的。这部分工作本文也借鉴了其中的设计思路。Yorktown Heights[10]则设计了一个很好的匹配上下文和候选实体上下文的算法来帮助消歧。在候选实体的向量表征方面，Xiao Huang[11]设计一个基于实体向量寻找的知识图谱问答系统，里面寻找候选实体的时候利用了实体间的距离来作为辅助特征。本文也利用了这个信息来帮助实体消歧，主要提取实体向量，同时用候选实体向量和当前向量的余弦距离作为重要的消歧因子，称为DeepCosine。

表1. 文本“这个夏天去哪里玩比较好”中“夏天”对应的候选实体

候选实体序号	候选实体描述
1	《夏天》是 2008 年上映的德国爱情电影，麦克马茨克执导……
2	《夏天》是李荣浩创作的歌曲，发行于 2010 年 7 月……
3	夏天，四季中的第二个季节，英语为 summer……
4-53	其他 49 个名称为夏天的实体描述

本文的主要工作和创新就是在于充分利用了实体的上下文信息和知识库信息，构建了DeepType、DeepMatch、DeepCosine三种模型来从三个不同方面进行实体消歧，充分利用了候选实体类型、上下文相关和候选实体向量这三个方面的信息。这些模型单个的效果并不完美，但是结合在一起之后消歧的能力便大大增强。

2 命名实体识别

本文设计和比较了两种实体识别的模型即经典方法word embedding+bilstm+crf和基于大规模语料预训练的方法Bert（finetuned）+crf。实体的编码方式采用BIESO的方式编码。单个字实体为S，非实体为O，连续片段则用BIE进行标注。图1（A）是我们的经典方法示意图，其中的分词使用的是开源的Jieba模型。图1（B）则是使用了BERT模型进行预训练的方法。

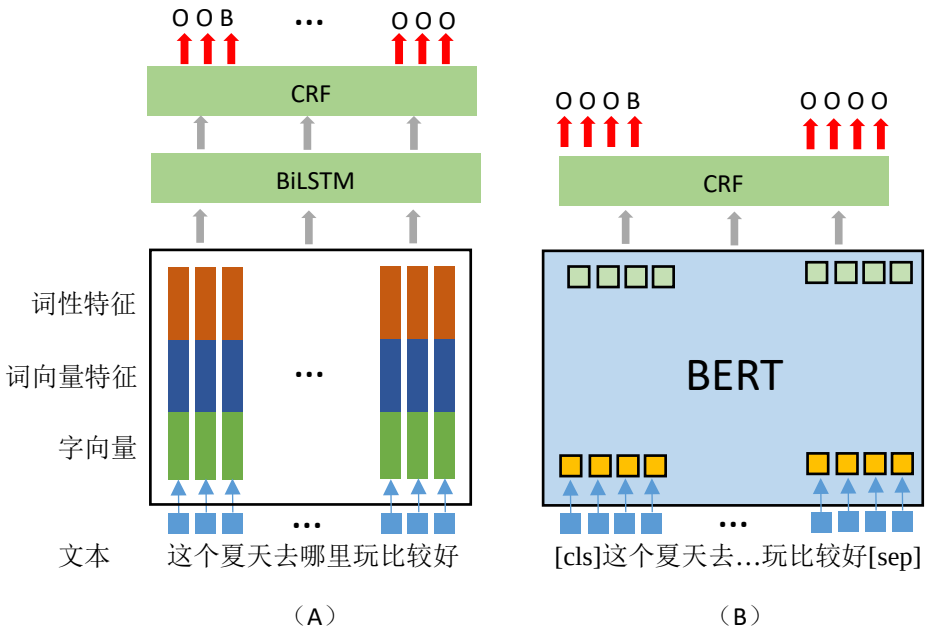

图1. 命名实体识别模型的设计。（A） 基于经典方法设计的实体识别模型，字、词向量采用了基于百度百科的300维词向量[1]。词向量特征为该字对应的分词结果中的词向量，同理词性特征为随机生成的100维向量；（B） 基于BERT Finetune[5]的识别实体模型。在原来BERT的基础上，增加了一层CRF层来实现更好的标注学习。

3 多因子融合的实体链指消歧

实体链指消歧的模型设计必然要和知识库或者实体库的结构和内容密切相关。本文所使用的知识库中的结构如图2所示。每个实体会有一个‘subject_id’字段，为该实体在知识库中唯一id。‘type’字段表示该实体类型。‘Predicate’中摘要则为介绍该实体的一段话，最后该知识库还会有其他属性信息表示该实体。图2只展示了部分实体属性信息。

1 https://github.com/Embedding/Chinese-Word-Vectors

```
{
    'alias': ['夏天'],
    'subject_id': '33119',
    'subject': '夏天',
    'type': ['thing'],
    'data': [{
        'predicate': '摘要',
        'object': '《夏天》是王浚懿写于2014年02的一首诗歌作品。'
    }, {
        'predicate': '作者',
        'object': '王浚懿'
    }, {
        'predicate': '作品名称',
        'object': '夏天'
    }]
}
```


图2. 知识库中实体结构分布。

3.1 DeepMatch模型

对于输入的文本“这个夏天去哪里玩比较好”，我们将此文本和所有候选实体一一配对。从中找出正确的配对的过程是一个二分类问题。因为输入语句上下文对实体消歧有很大的帮助[8]，因此本文构建了一个DeepMatch模型来匹配输入语句的上下文和候选实体的说明语句。候选实体的说明语句采用了摘要中的第一句话作为该候选实体的说明。采用的模型设计结构则是参考经典的ESIM[12]架构进行改进。如图3。其中输入语句和摘要文本中第一句话的encoder是基于百度百科的字向量。

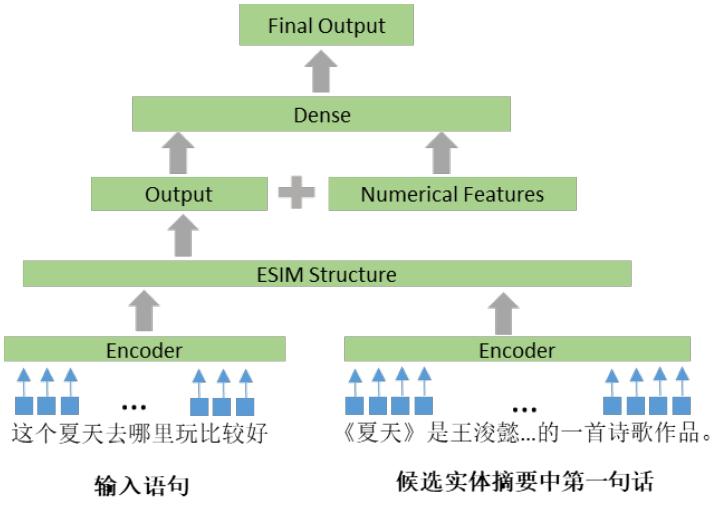

图3. DeepMatch模型结构示意图

DeepMatch模型中的Numerical Features为数值特征。该数值特征是人为提取的一些特征，其中重要的几个特征是历史点击率、该语句中其他实体是否在候选实体的摘要中、摘要的长度（类比于流行度）等。

3.2 DeepType模型

Jonathan Raiman[9]设计了一个Neural Type模型指导实体的消歧。文章中很重要的一个观点就是当我们知道了候选实体的类型之后，这个消歧的任务便被解决得差不多了。因此本文针对知识库中的‘type’字段设计了一个DeepType的预测系统。即根据训练集中已有的正确标注样本，我们可以知道该实体的类型是哪种。最终目的就是输入一句话并且给定潜在实体，该DeepType系统要能够预测出这个实体的类型。

DeepType模型的设计思路见图4。输入语句经过Bert获取到上下文相关字向量后，提取出实体区域（Entity Span）中第一个字和最后一个字的向量连接在一起，最后进行全连接（Dense）输出到各个候选类型进行多分类。分类层最后经过softmax归一化后取交叉熵作为损失函数。

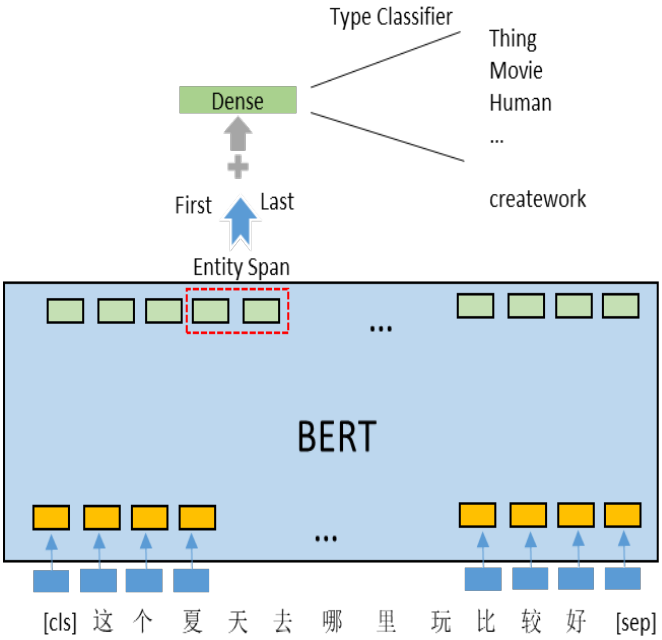

图4. DeepType模型对实体的类型进行预测

3.3 DeepCosine模型

知识库实体向量的表征对知识库中实体的识别至关重要。类似OpenKE[13]的工具对于帮助生成实体向量很有帮助。本文所使用的文本虽然具备一定的三元组结构，但是该三元组的末段即宾语结构部分并不常常是一个实体，而是一段描述文本。因此为了获得每个知识库中的实体表征，本文采用gensim中的word2vector方式，将知识库中的每个三元组，即（subject，predicate，object）都当成单独的一句话。Subject部分则用‘subject_id’代替形成一个完整的token。如下图5所示。最后生成的’33119’对应的词向量即被认为是该实体的实体向量。

‘33119’ 摘要 《夏天》是...的一首诗歌作品。
‘33119’ 作者 王浚懿。
‘33119’ 作品名称 夏天

图5.训练实体向量的数据准备形式

DeepCosine模型结构和DeepType的模型结构类似，参见图6。他们最后一层的目标不同，不是原先的分类模型而变成了如今的回归模型，回归目标就是该实体对应的实体向量（Entity Embedding）。损失函数也变为余弦距离损失。

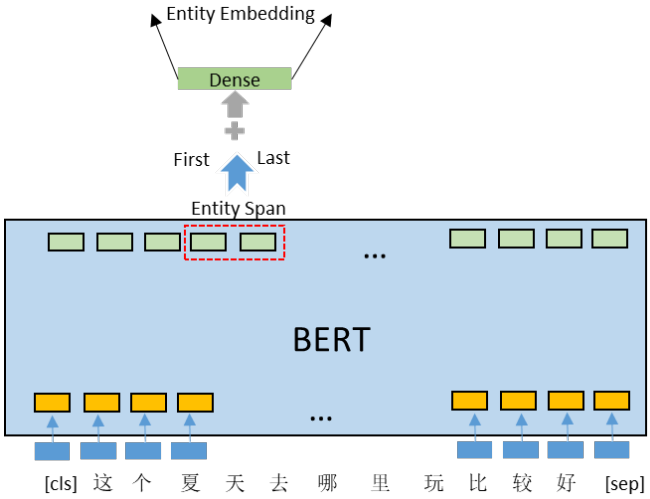

图6. DeepCosine模型的设计思路

3.4 模型融合

本文从三个不同方面刻画了实体和候选实体的相关性。因此最后需要通过模型融合（stacking）的方式来结合这三种特征以及其他一些数值特征来帮助模型进行消歧。最后二分类预测模型采用是lightgbm[14]这种梯度提升树。DeepType模型预测自身形成的特征是实体类型和候选实体类型的交叉熵损失函数（Type loss）。DeepMatch预测自身生成的特征是实体和候选实体的匹配程度

（0-1）。DeepCosine预测自身生成的特征则是实体和候选实体的向量的余弦距离。结合这三个特征和其他数值特征比如历史点击率，摘要的长度等，同时对这些特征相对于实体进行排序，得到他们的排序特征。这些特征工程完成后由lightgbm树模型输出他们的二分类预测结果。

4 实验结果

4.1 命名实体识别

实体识别的模型结构见图1。本文分别用A/B两种方法进行了实验。实验数据为百度CCKS2019的9万数据集。随机取其中1万作为我们实验的验证集，剩下的8万数据用来训练。实验结果如下表格2。NIL表示识别到的实体不在知识库中，受限于知识库的规模，会有相当一部分实体不被知识库包含。这部分实体会在后续的实体链指消歧中被去掉。从表格中可以看到基于Bert预训练的模型B相对于传统方法提升了很多。在本文的实验中，BERT模型的参数参考自文献[7]。比如学习速率取2e-5，这样的学习速率既可以避免灾难性遗忘，又可以很好的学习新数据的特征。本文的Bert模型学习了2个epoch就收敛至最佳效果。因此最后的实体识别模型采用的是全部训练数据训练2轮后的单模型。

表2. 两种实体识别算法在测试集上的表现

F1 值	含 NIL	去除 NIL
模型 A	0.8	0.82
模型 B	0.832	0.851

基于Bert的实体识别模型取得了很大的提升，但是仍然有一部分实体数据没有被很好的识别出来。因此本文试图对这些错误数据进行错例分析。表3中的错例很好的代表了模型所有识别错误的情况。比如“艺术”这个词在训练集中有一定概率被标注，因此模型只能最大似然的估计这个词是否需要标注为实体，受限于标注人员的标注习惯，必然会有部分实体被认为错误标记。实体识别模型在训练的过程之中也是在学习标注人员的标注习惯。

表3. 实体识别结果错例分析

句子	Bert 模型预测结果	标注结果
歌曲《乡音乡情》艺术分析	[歌曲，乡音乡情，艺术]	[歌曲，乡音乡情]
郑保国：助推企业创新发展的动力之源	[郑保国，企业]	[郑保国，企业，动力]

4.2 实体链指消歧

在实体链指消歧任务中，本文令每个候选实体和输入语句中的实体一一配对，形成一个二分类问题。将9万训练集一一配对后得到的总的二分类任务数据条目是150万以上。这么大数据量的分类任务采用lightgbm这种高效的梯度提升树来建模是非常有效的。对于这些分类任务中的特征，主要采用了DeepMatch、DeepCosine、DeepType三种模型做预测自身而生成。如图7所示，先把数据分成5份。取其中四份数据和对应的label训练一个模型model1。该模型对part5进行预测，得到自身的预测部分pred5。同理，循环这个过程，分别得到5个模型对Part1-5进行预测生成Pred1-5。这些预测结果连接在一起就可以构成该模型形成的一个特征。同时用这五个模型对测试集进行预测求平均，则得到测试集的特征。

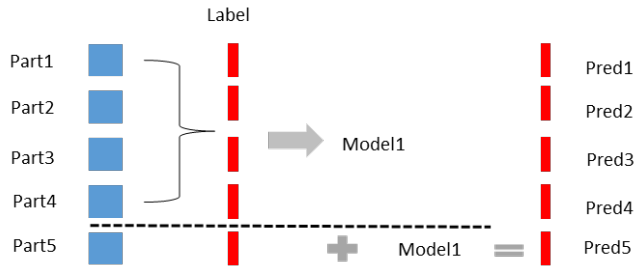

图7. Stack模型框架设计

这些特征在lightgbm的二分类模型下对应的特征重要性如下图8。可以看到DeepMatch的模型重要性最高，重要性的评价指标为树模型中划分过程中的该特征的平均增益。这些特征对应的中文含义对应表4。随机选一折数据用作测试集，得到的实体消歧的f1=0.92，去掉DeepMatch特征后的f1值迅速下降到0.905。可见DeepMatch为模型的提高贡献了很多指导价值。整个数据集的实体消歧的基线f1值是0.5（采用随机选取的方式）。当然我们可以看到，候选实体的摘要字数也很有价值，我们认为它相当于流行度这个特征。因为摘要越完善，说明知识库对它的维护越好，越说明该实体比较受重视。其他的特征比如一些排序特征也发挥了重大价值。因为大多数情况下正确的实体是从候选实体中选取一个作为标准答案，所以如果能把这个问题变成一个理想的排序问题的话相信结果也会进一步提高。

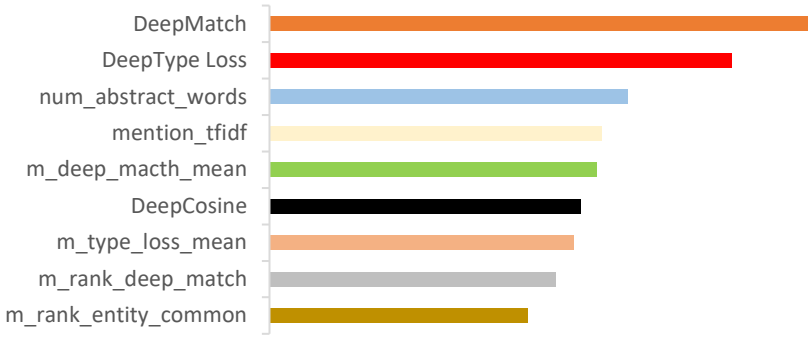

图8. lightgbm输出的前9个特征重要性排行

表4. 前9个重要特征的中文解释

m_rank_entity_com-mon	**输入语句中其他实体是候选实体摘要中的次数排序**
m_rank_deep_match	所有候选实体的DeepMatch预测值的排序
m_type_loss_mean	所有候选实体的类型损失
DeepCosine	DeepCosine模型预测的余弦距离
m_deep_macth_mean	所有候选实体的DeepMatch均值
mention_tfidf	输入实体的tfidf值
num_abstract_words	候选实体的摘要字数，类比于流行度
DeepType Loss	DeepType模型的Type交叉熵损失
DeepMatch	DeepMatch模型预测的匹配程度

5 总结与讨论

本文对实体识别与实体链指消歧方面做了一些有益的探索。在输入语句的词汇表征上，再一次证明了Bert的预训练模型已经超过经典固定的word2vector方法。同时对于实体链指消歧这个任务，本文综合了当前一些优秀的解决方案，通过模型融合的方式极大地提高了实体消歧的准确率。

与此同时，本文还有一些值得探索的地方有待完善。比如没有充分利用好Bert预训练过程中的NSP（Next Sentence Prediction）任务。该任务中用大量语料训练了上下句相关性，此方法可以移植用于实体消歧。另外，实体消歧很多时候是排序问题。因此把某个实体的所有候选实体一一配对形成一个batch，然后最后输出的时候在batch维度进行softmax归一化，这样排序后的loss可能会有更好的解释性。

参考文献

1. Piccinno F, Ferragina P. From TagME to WAT: a new entity annotator. InProceedings of the first international workshop on Entity recognition & disambiguation 2014 Jul 11 (pp. 55-62). ACM.
2. Daiber J, Jakob M, Hokamp C, Mendes PN. Improving efficiency and accuracy in multilingual entity extraction. InProceedings of the 9th International Conference on Semantic Systems 2013 Sep 4 (pp. 121-124). ACM.
3. Steinmetz N, Sack H. Semantic multimedia information retrieval based on contextual descriptions. InExtended Semantic Web Conference 2013 May 26 (pp. 382-396). Springer, Berlin, Heidelberg.
4. Kolitsas N, Ganea OE, Hofmann T. End-to-end neural entity linking. arXiv preprint arXiv:1808.07699. 2018 Aug 23.
5. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
6. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237. 2019 Jun 19.
7. Sun C, Qiu X, Xu Y, Huang X. How to Fine-Tune BERT for Text Classification?. arXiv preprint arXiv:1905.05583. 2019 May 14.
8. Le P, Titov I. Improving entity linking by modeling latent relations between mentions. arXiv preprint arXiv:1804.10637. 2018 Apr 27.
9. Raiman JR, Raiman OM. DeepType: multilingual entity linking by neural type system evolution. InThirty-Second AAAI Conference on Artificial Intelligence 2018 Apr 27.
10. Sil A, Kundu G, Florian R, Hamza W. Neural cross-lingual entity linking. InThirty-Second AAAI Conference on Artificial Intelligence 2018 Apr 27.
11. Huang X, Zhang J, Li D, Li P. Knowledge graph embedding based question answering. InProceedings of the Twelfth ACM International Conference on Web Search and Data Mining 2019 Jan 30 (pp. 105-113). ACM.
12. Chen Q, Zhu X, Ling Z, Wei S, Jiang H, Inkpen D. Enhanced lstm for natural language inference. arXiv preprint arXiv:1609.06038. 2016 Sep 20.
13. Han X, Cao S, Lv X, Lin Y, Liu Z, Sun M, Li J. Openke: An open toolkit for knowledge embedding. InProceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations 2018 Nov (pp. 139-144).
14. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. InAdvances in Neural Information Processing Systems 2017 (pp. 3146-3154).