# 中文短文本的实体链指研究

徐国进

电子科技大学,成都,610031, 中国
xgj_012@163.com

**摘要.** 面向中文短文本的实体识别与链指，简称ERL（Entity Recognition and Linking），是NLP领域的基础任务之一，即对于给定的一个中文短文本识别出其中的实体，并与给定知识库中的对应实体进行关联。传统的实体链指任务主要是针对长文档，长文档拥有在写的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。针对该问题，本文给出了解决了该任务即实体识别和实体链指两个子任务的方法。对于实体识别这一子任务，本文采用BERT-BiLSTM-Dense的半指针半标注的结构，这一结构通过灵活的解码来提升实体识别的性能。对于实体链指这一子任务，根据实体描述文本长度对候选实体进行有效的筛选获得一个小的候选集，然后将实体消歧转为化在这个小的候选集上的多分类的问题。在CCKS 2019 中文短文本的实体链指这个评测任务上，该文提出的方法在最终的评测数据上达到F1为0.79654的成绩。

**关键词:** 实体链接，实体识别，实体消歧，BERT，多分类

# Research on Entity Chain Reference of Chinese Short Texts

XU Guojin

University of Electronic Science and Technology,Chengdu 610031, China
xgj_012@163.com

**Abstract.** Entity Recognition and Linking (ERL) is one of the basic tasks in NLP field. It is to recognize the entities in a given short Chinese text and associate them with the corresponding entities in a given knowledge base. Traditional entity chain refers to the task mainly for long documents, long documents have context information in writing, which can assist entity disambiguation and complete the chain finger. In contrast, the entity chain finger for short Chinese texts is a big challenge. To solve this problem, this paper presents a method to solve the task, that is, entity recognition and entity chain referring to two subtasks. For the sub-task of entity recognition, this paper adopts the structure of BERT-BiLSTM-Dense semi-pointer and semi-annotation, which improves the

performance of entity recognition by flexible decoding. For the sub-task of entity chain, a small candidate set is obtained by effectively screening candidate entities according to the length of entity description text, and then entity disambiguation is transformed into multi-classification on this small candidate set. In the evaluation task of entity chain finger of CCKS 2019 Chinese short text, the method proposed in this paper achieves the result of F1 0.79654 on the final evaluation data.

**Key words:** entity links, entity recognition, entity disambiguation, BERT, multi-classification

## 1 引言

近年来，随着互联网的普及和迅速发展，越来越多的短文本数据产生，如搜索Query、微博、用户对话内容、文章标题等。识别出这些短文本中的实体并确定这些实体链指到给定知识库中的目标实体具有重要的意义。

传统的实体链指任务主要是针对长文档，主要利用词袋子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，长文档拥有在写的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。

针对中文短文本的实体链指的这些问题，实体识别本文采用了基于预训练的bert的半指针半标注的方法，通过灵活的解码方式来提高识别的性能；实体消歧这一任务，本文首先对待链指的候选实体数目过多的情况下进行一道筛选并保证召回率，然后对剩下的候选实体采用多分类的方法来进行链指。

## 2 相关研究

中文短文本的实体链指的核心就是实体识别与实体链指。中文的命名实体识别较之英语要更为复杂困难，这主要表现在汉语文本中没有表示词语边界的分隔符号，命名实体的识别效果很大程度受自动分词结果影响[1]，而汉语自动分词的效果往往也受制于命名实体的识别。早期NER研究，人工构建有限规则，再从文本中寻找匹配这些规则的字符串，是一种主流的方法。同时研究者们也试图借助机器自动地发现和生成规则，这其中最具代表性的便是Collins 等[2]提出的DLCoTrain方法，类似的还有使用Bootstrapping 进行规则自动生成的方法[3]。而试图通过制定有限的规则来识别出变化无穷的命名实体，这样的方法愈发显得笨重。更不用说规则对领域知识的极度依赖，使得当领域差别很大时，制定的规则往往无法移植，不得不重新制定规则。随后随着机器学习在NLP领域的兴起，经典机器学习分类模型如HMM[4]、ME[5]、CRF[6]都被成功地用来进行命名实体的序列化标注，且获得了较好的效果。近年来，深度学习技术成

为机器学习领域新的热潮，基于深度学习的NER也有很多的研究，很多由基于深度神经网络和CRF相结合的方法，如Bi-LSTM-CRF[7]、ID-CNN-CRF[8]等都取得了很好的效果。而近年出现的BERT[9]在各大任务上都超越了以往模型的成绩，因此本文采用基于BERT的实体识别的方法，此外从模型解码灵活度角度考虑，本文的实体识别没有与CRF结合，而是采用半指针半标注的结构[10]

实体链接的核心是计算实体指称项和候选实体的相似度, 选择相似度最大的候选实体作为链接的目标实体[11]。Bagga[12]等人用词袋模型来解决人名歧义的问题。Fleichman[13]等人利用网络信息等特征训练最大熵模型来解决实体歧义问题。这些方法都是通过衡量指称项上下文文本与目标实体文本之间的相似度来判定两者是否一致。在这些方法中很大一部分都是利用词袋模型或者类似于词袋模型的方法，然而词袋模型只能捕捉表层字面匹配信息无法捕捉深层语义。为了解决这个问题，基于深度学习的实体消歧可以采用类似句子相似度或者句对匹配的方法，本文采用多分类的方法来进行实体的消歧，即实体指称项与所有候选实体同时进行匹配。

# 3 实体识别与实体消歧

## 3.1 预处理

CCKS 2019 中文短文本的实体链指这个任务提供的训练数据中的有一些标注错误，预处理主要是对这些标注错误进行了一些处理，同时也对kb_data中的实体data项的predicate与object进行了拼接作为实体的描述文本。

**标注错误处理**

针对训练数据中实体中包含实体的标注错误，经观察发现这部分错误往往都是索引的错误，例如：{"text_id": "855", "text": "《摩登家庭》里出现的这种相互理解宽容的家庭,真实吗?...", "mention_data": [{"kb_id": "37876", "mention": "摩登家庭", "offset": "1"}, {"kb_id": "385888", "mention": "家庭", "offset": "3"}]}。即offset为3的那个"家庭"不是"摩登家庭"中的，而是后面那个家庭，于是我们直接将这种类似的错误，对其offset进行修正处理。训练数据中的训练集中有2.592%的实体名在实体库中无法匹配，例如：1. 安妮•海瑟薇：文本中间有特殊符号；2. 新浪微薄：输入文本中实体名错误；3. 下载, : 实体名标注错误，引入了符号；4. 国家质检总局：别名不在知识库中。这部分错误的实体和知识库中的实体往往只是有细微的差别，比如：多一个符号，错了一个字。针对这部分错误我们直接根据与此错误实体编辑距离最小的对应kbid中的实体来代替它即可。

**实体描述文本的构建**

根据知识库中每个实体的data项，对data项中的predicate和object进行拼接，例如数据：{"alias": ["无尽武道"],"subject_id": "10007", "subject": "无尽武道", "type": ["CreativeWork"], "data": [{"predicate": "摘要", "object": "《无尽武道》是一部爱情类的小说，作者是穿越梦竹，小说仍在连载中，连载网站是晋江文学城。"},

{"predicate": "连载网站", "object": "晋江文学城"}, {"predicate": "小说进度", "object": "连载"}, {"predicate": "作者", "object": "穿越梦竹"}, {"predicate": "中文名", "object": "无尽武道"}, {"predicate": "小说类型", "object": "爱情"}, {"predicate": "义项描述", "object": "穿越梦竹著作小说"}, {"predicate": "标签", "object": "书籍"}, {"predicate": "标签", "object": "小说作品"}, {"predicate": "标签", "object": "中国文学"}, {"predicate": "标签", "object": "文学作品"}, {"predicate": "标签", "object": "小说"}]}。

将所有predicate和object通过拼接得到如下实体描述文本：标签：小说。标签：文学作品。标签：中国文学。标签：小说作品。标签：书籍。义项描述：穿越梦竹著作小说。摘要：《无尽武道》是一部爱情类的小说，作者是穿越梦竹，小说仍在连载中，连载网站是晋江文学城。连载网站：晋江文学城。小说进度：连载。作者：穿越梦竹。中文名：无尽武道。小说类型：爱情。

由于标签与意向描述是比较重要的信息，描述文本中都将标签项、意向描述项放在了文本的最前面，以防止文本过长经截断而丢失。对于过长的文本，采用按最大长度728截断处理。

### 3.2 实体识别

**模型结构**

实体识别本文采用的是BERT+BiLSTM+Dense的一个半指针半标注的模型结构，其结构如下所示：

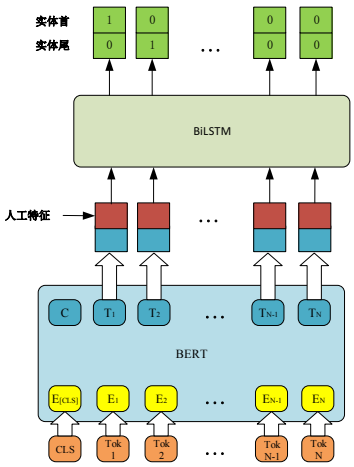


**图. 1.** 实体识别模型结构

其中BERT采用的是中文BERT-wwm[14]的预训练模型fine tuning，首先通过BERT编码得到每个字的向量表示，然后拼接人工特征，经过一层BiLSTM，最后经过Dense，激活函数采用sigmoid，每个字得到一个二维的向量，即表示此字是实体的首和尾的概率。

**人工特征**

由于数据中实体长度为2、3、4、5的实体占比很大。人工特征本文提取的是：序列中的字是否为长度为2、3、4、5的kb_data中的实体的首尾的字特征。这个特征依靠结巴分词后来判断分词后的每个词是否在kb中，如果在kb中且长度在2、3、4、5范围内，则将此词对应的首和尾位置分别标1，其他位置标0，这样能得到8个字特征，用来作为实体识别的先验信息。

**解码**

解码采用设置阈值的方式来判断每个字是否是实体的首和尾，对于某个首位置，我们遍历其后面所有的尾位置，再通过判断这样得到的实体是否在kb中来做排除。通过这样的方式我们可以提高实体的召回率。对召回来的实体我们又做了最大粒度的选择，这样用来提高实体的准确率。

### 3.3 实体消歧

**链指框架与多分类模型结构**

实体链指采用候选集筛选然后通过多分类的方式来消歧，其框架如下所示：

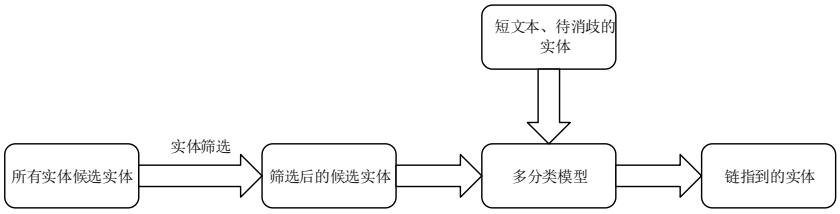


**图. 2.** 实体链指框架

本文采用多分类的模型来进行消歧，其模型结构如下所示：

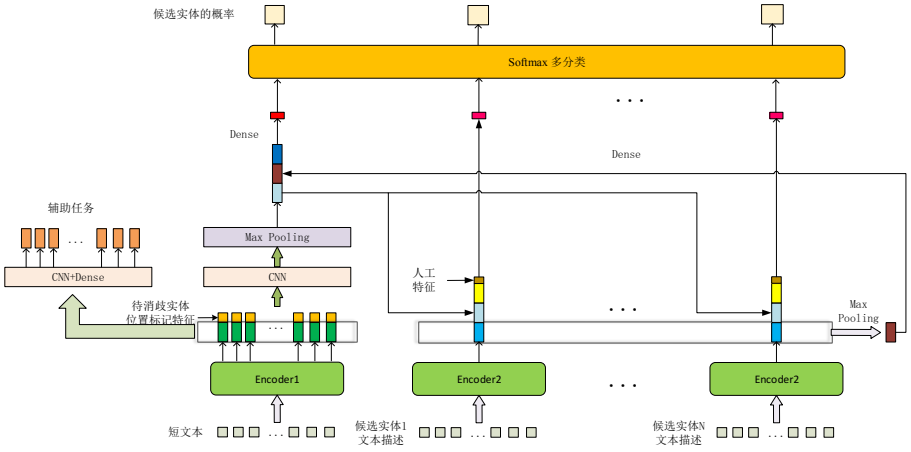


**图. 3.** 实体消歧模型结构

其中短文本通过Encoder1编码，候选实体描述文本用Encoder2编码。其中短文本对应的softmax输出表示为NIL的概率，每个候选实体描述文本的softmax输出表示为此链指到此候选实体的概率。待消歧实体位置标记是一个0，1的字特征，即待消歧实体的位置上标记1，其他位置标0。短文本通过Encoder1编码后拼接待消歧实体位置标记特征后经过CNN层然后再进行最大池化得到短文本-待消歧的编码向量 $\mathbf{d}$ 。候选实体n描述文本用Encoder2编码得到一个编码向量 $\mathbf{x}_n$ 。然后对所有候选实体描述文本的编码向量经过最大池化操作，得到一个同样长度的编码向量 $\mathbf{x}$ ，此向量用来表示所有候选实体描述文本的信息。通过拼接 $[\mathbf{d},\mathbf{x},\mathbf{d}\circ\mathbf{x}]$ ，然后进行Dense得到NIL的Softmax输入，通过拼接 $[\mathbf{x}_n,\mathbf{d},\mathbf{d}\circ\mathbf{x}_n,\mathbf{f}_n]$ 然后进行Dense得到候选实体n的Softmax输入，其中 $\mathbf{f}_n$ 是实体n的人工特征。模型中的辅助任务是对短文本中的实体首尾位置以及实体类型的的学习，其结构是采用实体识别中的半指针半标注的结构。

模型中Encoder1是一个经过字词向量对齐拼接后经过两层BiLSTM的操作，其结构如下所示：

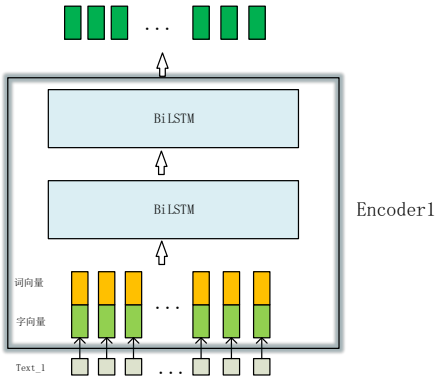


**图. 4.** Encoder1模型结构

其中字、词向量的训练语料由CCKS 2019 中文短文本的实体链指这个任务提供的train数据、develop数据、test数据中的text以及由kb_data构建的实体描述文本组成。

Encoder2是一个经过字词向量对齐拼接，然后复制subject_type先验信息并拼接，接着再拼接字特征，经过两层BiLSTM的操作，最后通过最大池化操作得到一个编码向量，其结构如下所示：

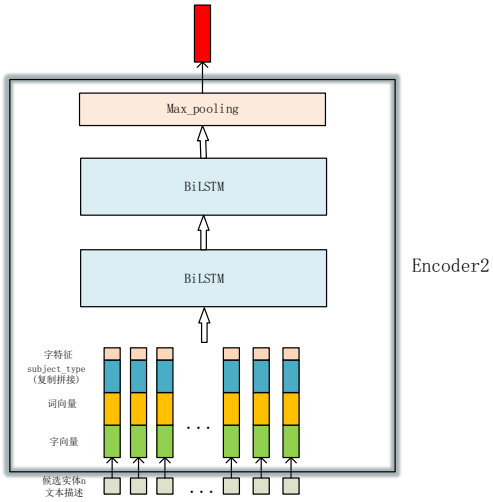


**图. 5.** Encoder2模型结构

**特征**

实体类型
kb_data中的实体类型的类别总共有51种。我们将subject_type编码为一个长度为51的0,1向量，每一位中1表示属于此种类型，0表示不属于此种类型。

人工字特征
字特征有两个，一个用于表示实体描述文本中predicate位置的0,1字特征：实体描述文本中predicate位置标记为1，其他的全标为0；另一个字特征为实体描述文本中的实体位置0,1字特征，文本中出现此实体或其别名则标记为1，其他全标记为0。

候选实体人工特征
候选实体人工特征 主要主要是从实体流行度的角度考虑的。实体描述文本长度越长，其流行度越高，被链指到的概率也越大，于是我们从实体描述文本长度的角度提取了两个特征：一个是在候选实体描述文本长度的排序特征，另一个是代表候选实体描述文本绝对长度的特征。此外一个实体别名越多，也一定程度上反应了实体的流行度，于是我们也提取了别名排序特征，候选实体中别名最多的为1，其他为0。

**候选集的筛选**
由于有一定比例的待消歧的候选实体数目非常多，最多的甚至超过1000。而多分类的消歧模型无法直接对这种候选实体数目太大的待消歧实体在内存有限的

情况下完成，于是本文考虑先在所有可能的候选实体中排除一部分实体，并同时保证召回率。

统计发现实体描述文本长度是一个很好的用于候选集选取的特征，当我们把最大候选集设置为64时，通过选择实体描述文本长度最长的64个作为候选集时，总体的召回率接近100%。

# 4 实验结果及分析

## 4.1 实验数据

本文实验采用的训练数据、评测数据均由CCKS 2019 中文短文本的实体链指这个任务提供。该任务的知识库包括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KB-ID，一个字符串名称，上位type信息及与此实体相关的一系列三元组<subject，predicate，object>信息形式。知识库中每行代表知识库的一条记录，每条记录的格式为一个json格式。predicate-id和object-id的值都为0，subject-id的值为一个正整数，说明subject总是对应知识库中的一个实体。训练数据中包括9万条短文本标注数据，数据通过百度众包标注生成（人工进行评估，其平均准确率95%以上）。数据主要来自于：真实的互联网网页标题数据，是用户检索Query对应的有展现及点击的网页，短文本平均长度为21.73中文字符，覆盖了不同领域的实体（包括各垂类的实例、概念），如人物、电影、电视、小说、软件、组织机构、事件等垂类，以及通用概念。

## 4.2 实验参数设置

实体描述文本长度本文采用最大长度728来截断。字、词向量通过gensim分别训练128维的字、词向量模型，训练语料为train数据、develop数据、test数据中的text以及kb_data中的实体描述文本，其中参数设置为：窗口window为6、采用skip-gram算法、最小词频min_count为2，训练迭代30轮。实体识别解码时阈值设置为0.35。实体消歧主辅任务的损失函数权重比为1:0.2。

## 4.3 实验结果

实体识别与实体消歧均采用9折交叉验证，对3万条评测数据随机打乱并且分成9份，实体识别与实体消歧分别训练9个模型，然后分别对结果做平均。首先通过实体识别模型解码得到实体，然后再对识别的实体通过多分类的消歧模型，得到类别为NIL的实体在评测时去掉。训练集上实体识别、实体消歧交叉验证的结果以及实体链指的评测结果如下所示：

**表 1.** 实验结果

| 任务 | 准确率 | 召回率 | F1 |
|---|---|---|---|
| 本地实体识别 | 0.8575 | 0.8615 | 0.8595 |
| 本地实体消歧 | 0.9031 | 0.9104 | 0.9067 |
| 整体的实体链指评测 | 0.7913 | 0.8018 | 0.7965 |

## 5 结论

本文介绍了小组参加CCKS 2019 中文短文本的实体链指评测的基本情况。实体识别方面采用的半指针半标注的结构通过灵活的解码方式提升了效果；实体消歧方面本文没有采用基于二分类的方法，而是通过候选实体筛选然后通过多分类的方法来达到消歧。从评测结果上看，本文的模型达到了不错的性能。

## 参考文献


1. 赵军. 命名实体识别、排歧和跨语言关联[J]. 中文信息学报,2009, 23(2): 3-17.
2. Collins M, Singer Y. Unsupervised models for named entity classification[C]// Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999: 100-110.
3. Cucerzan S, Yarowsky D. Language independent named entity recognition combining morphological and contextual evidence[C]// Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC, 1999: 90-99.
4. Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder[C]// Proceedings of the Fifth Conference on Applied Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 1997: 194-201.
5. Borthwick A E. A maximum entropy approach to named entity recognition[D]. New York: New York University, 1999.
6. McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Stroudsburg: Association for Computational Linguistics, 2003,4: 188-191.
7. Lample G , Ballesteros M , Subramanian S , et al. Neural Architectures for Named Entity Recognition[J]. 2016.
8. Strubell E , Verga P , Belanger D , et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions[J]. 2017.
9. Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
10. 苏剑林. (2019, Jun 03). 《基于DGCNN和概率图的轻量级信息抽取模型 》[Blog post]. Retrieved from https://kexue.fm/archives/6671
11. 赵军, 刘康, 周光有, 等. 开放式文本信息抽取. 中文信息学报, 2011, 25(6): 98–110
12. Amir S M , Bagga A . Entity-Based Cross-Document Coreferencing Using the Vector[C]// Meeting of the Association for Computational Linguistics & International Conference on Computational Linguistics. Association for Computational Linguistics, 1998.

13. Fleischman M, Hovy E. Multi-document person name resolution[C]//Proceedings of the Conference on Reference Resolution and Its Applications. 2004: 1-8.
14. Cui Y, Che W, Liu T, et al. Pre-Training with Whole Word Masking for Chinese BERT[J]. arXiv preprint arXiv:1906.08101, 2019.