基于模式扩充及 BERT 分类的上下位关系识别研究

苏江文

福建亿榕信息技术有限公司 福建福州 350003

{sujiangwen@sgitg.sgcc.com.cn}

摘要：针对医疗实体概念的上下位关系自动识别问题，提出了一种基于模式扩充及 BERT 分类的上下位关系预测的方法，基于模式扩充从外部数据中抽取潜在的上下位关系对，利用 BERT-Attention-Bi-LSTM 模型对潜在上下位关系对进行预测，获得主要上下位关系预测结果。方法在 CCKS2020 新冠知识图谱中概念实体的上下位关系识别任务评测中排名第一，验证了方法的有效性及可行性。

关键词：上下位关系；BERT 分类；模式扩充；关系抽取

Research on Recognition of Hypernym Relations Based on Pattern Expansion and BERT Classification

Su Jiangwen

Fujian Yirong Information Technology Co Ltd, Fuzhou 350003, Fujian, China

Abstract: Aiming at the problem of automatic recognition of the hypernym relation of the medical entity concept, a method for predicting the hypernym relation based on pattern expansion and BERT classification is proposed. Based on the pattern expansion, potential hypernym relation pairs are extracted from external data, using BERT-Attention-Bi-LSTM model predicts the potential hypernym relation pairs and obtains the prediction results. The method ranks first in the evaluation of the hypernym relation recognition task of CCKS2020, which verifies the effectiveness and feasibility of the method.

Key words: hypernym relation; BERT classification; pattern expansion; relation extraction

1 引言

上下位关系(Hypernym Relation)是自然语言处理任务中较受关注的语义关系之一,对知识库与知识图谱的构建具有重要的意义。在知识图谱构建过程中，人工预定义的实体类型覆盖程度有限且不易更新，当涉及新的领域时，实体类别体系可能需要重新定义。通过在网络中动态的获得实体的概念类别,并自动化识别类别之间的上下位关系不但可以解决人工预

通达作者：sujiangwen@sgitg.sgcc.com.cn

定义的缺陷，还可以使知识图谱更加立体丰满，有助于上层应用。

在如今的信息化时代，互联网中实体类别多样化，且粒度更细并具有层次，相对于类别有限的传统命名实体，人们开始将目光转向开放域实体挖掘。所以，上下位关系也常常出现在与知识图谱相关的任务中。本文基于CCKS 2020新冠知识图谱构建与问答评测的子任务“新冠概念图谱的上下位关系预测”的实践，提出一种基于模式扩充及BERT分类的上下位关系预测的方法，经评测，提交结果排名第一，一定程度上验证了方法的可行性。

2 相关研究

2.1 基于规则的方法

基于规则的方法使用词汇句法模式从文本中识别上下位关系。该研究领域中最早且最具影响力的是Marti Hearst教授,她手工定义了7种用来从英文语料中识别“is-a”关系的语言模式[1]。这些模式引起了广泛关注,并且今天也经常使用。“[C] such as [E]”、“[E] is a [C]”是典型 Hearst 模式,其中[C]和[E]是名词短语占位符,分别代表了在“is-a”关系(x,y)中的上位词(类)y和下位词(实体)x。Probase就是利用Hearst模式从数十亿个网页中抽取“is-a”关系构建的,它包含了256万个概念和2076万对“is-a”关系。Kozareva等也采用了相似的方法,他们利用Hearst模式从网页中提取了用来学习分类法的“is-a”关系。

基于规则的方法所定义的模式很精确,并且对英语语料中的上下位关系有很高的覆盖率,但由于自然语言的歧义性和复杂性,这些太具体的模式无法覆盖所有的语言情况,因此往往召回率很低。而且简单模式匹配常常由于惯用表达式、解析错误、不完整或无用信息的提取以及模棱两可的概念而出现错误。所以，在如何基于规则的方法再提升准确性和召回率方面，也开展了一些研究，总结介绍如下：

（1）召回率提升相关方法

一是模式泛化。这种方法主要有两种途径。一是通过语言规则来扩展原始的 Hearst 模式， 二是通过语料来学习更广义、更泛化的词汇句法模式[2-3]。二是关系推断[4-5]。关系推断克服了术语对(x,y)必须在同一个句子中出现的限制，其假设y是x的上位词， 而x1与x及其相似，则 y 很有可能是 x1的上位词。而x与x1的相似性,可以通过词嵌入表示后，转换成基于余弦相似度等相似性判定方法。

（2）准确性提升相关方法

一是置信度评估。在提取出候选上下位关系对(x,y)之后,可以使用统计方法来计算置信

度分数,得分低的关系对将会被过滤。例如，根据搜索引擎查询结果的命中次数来估算 x 和 y 的逐点互信息(PMI),或者考虑外部因素,如 WordNet 等词典中所包含概念以及数据源的可信赖度。二是基于分类的验证。这些方法通过训练分类器 f 来预测所抽取关系对(x,y)的正确性,选用的典型模型主要包括支持向量机(SVM)、逻辑回归以及神经网络。而分类器 f 所用特征大概可以划分为以下几类:表层名称、语法、统计信息、外部资源等。

2.2 基于统计的方法

基于统计的方法通过对大规模语料库的统计处理发现规律，从而识别上下位关系。基于统计的方法主要分为分类及词嵌入投影。

分类方法集中在机器学习算法的研究上，机器学习也是研究成果中出现最多、应用最广泛的信息抽取技术，主要涵盖支持向量机、条件随机场、决策树、朴素贝叶斯、神经网络等。考虑到中文语言的特性，分类方法通常会结合额外的语言规则、句子结构特征、句法特征、词典以及知识库等。多种特征以及知识库的融合在一定程度上能够有效提高识别的准确率，但分类方法人面临着一些挑战，如特征构建的过程随机、难以复制且不可控，知识库的维护更新代价太高等。

另一一种基于统计的中文上下位识别方法是基于词嵌入的投影模型，其在多个英文数据集上取得了 State-of-the-art 的结果，在中文上的表现还有很大提升的空间。它们不需要利用各种相似性度量方法，而是从语料中获取词语特征来识别上下位关系。典型工作 Wang 等[6]设计的深度学习模型在中英文症状和疾病语料库上都取得了不错的结果。孙佳伟等[7]利用简单的前馈神经网络和 softmax 结构来实现了关系的分类。深度神经网络可以有效地学习上下位关系和非上下位关系的复杂非线性投影。

3 基于模式扩充及 BERT 分类的上下位关系预测

本文的方法是基于模式扩充从外部数据中抽取潜在的上下位关系对，而后利用基于 BERT 预训练模型的分类技术对潜在上下位关系对进行预测，获得主要上下位关系预测结果。在此基础上，结合对领域数据的观察，编写规则对上下位关系补充与后处理，整合形成最终结果。主要包括以下三个阶段。

3.1 基于外部数据的上下位关系种子发现

本步骤的目标，是基于外部数据和规则模式，形成少量准确的上下位关系种子，为后续的全量潜在上下位关系分类判断提供基础样本。

（1）上下位关系识别模式扩充

由于任务未提供种子上下位关系作为模型训练的，本文首先开展了种子关系的识别。根据中文的语言学特征，制定规则模板，可以较大程度的利用知识、准确率较高。然而，原始的 Hearst 模式给出了面向英文的抽象模式。相较于英文，从中文文本语料库中识别上下位关系更是一项艰难的挑战。从语言学的角度看，中文是表意文字的一种形式，其词的结构、语义和语法是灵活和不规则的。通过对相关语料的观察及任务提交探测，本文给出了一组面向医疗实体概念上下位模式规则集合。表 1 展示了其中的主要部分。

表 1：医疗实体概念上下位模式规则集

上下位模式规则
(.*?)指的是.+的(.+)
(.*?)为.+的一种(?+)
(.*?)是.+的.*
(.*?)包含(.*?)
(.*?)是(.*?)的一种
(.*?)指.+的(.+)
(.*?)是(.*?)的术语之一

（2）基于新冠知识图谱概念数据集的上下位关系种子发现

“CCKS 2020：新冠知识图谱构建与问答评测（一）新冠百科知识图谱类型推断”提供了新冠开放知识图谱的概念数据集。概念数据集提供包括了维基百科等相关外部网页全文，其中包含了大量明确的上下位关系。具体方法是：利用上述扩充的医疗实体概念上下位模式规则集，在本数据集上针对所有文章执行全文规则匹配，获得高准确性的上下位关系。

（3）基于《大词林》的种子上下位关系补充

除了上述基于规则及外部数据的种子发现外，本文还从《大词林》[8]中获取上下位关系作为补充。《大词林》是一个开放域命名实体知识库自动构建系统，系统从 Web 搜索结果、在线百科和命名实体字面等多个信息源挖掘命名实体的类别，并从 Apriori 关联项、后缀上位词、分类层次化和词汇分布表示等多个角度学习获取类别之间的层次化关系。开源的《大词林》中的 75 万的核心实体涵盖了常见的人名、地名、物品名等术语，概念词列表则包含了细粒度的实体概念信息，可作为 NLP 相关任务的良好数据基础。

本文基于《大词林》全量数据，以层次遍历的方式，获得了一部分医疗实体概念上下位关系，与面向外部数据的规则匹配发现的结果整合，形成上下位关系模型训练种子语料。

3.2 基于 BERT-Attention-Bi-LSTM 的上下位关系分类

主要包括两个步骤。一是基于任务提供的待预测数据，两两匹配生成全量的待预测上下位关系数据集。二是以 3.1 节识别的上下位关系模型训练种子语料作为正例，从带预测关系数据集中辅以简单规则识别出 1：1 的负例，基于基于 BERT-Attention-Bi-LSTM 模型开展上下位关系分类模型训练，对待预测数据集进行预测。其中第一个步骤比较简单，不再展开阐述，以下重点对基于 BERT-Attention-Bi-LSTM 模型的上下位关系分类进行说明。

本文最终采用了 BERT-Attention-Bi-LSTM[9]模型，其结构主要分为三部分：将待预测的上下位关系词对拼装为序列文本，而后通过 BERT 模型训练获取序列文本的语义表示，再将文本中每个字的向量表示输入到 Attention-Bi-LSTM 模型中,进行进一步语义分析。最后，将 softmax 层输出文本标签，0 代表非上下文关系，1 代表是上下位关系。

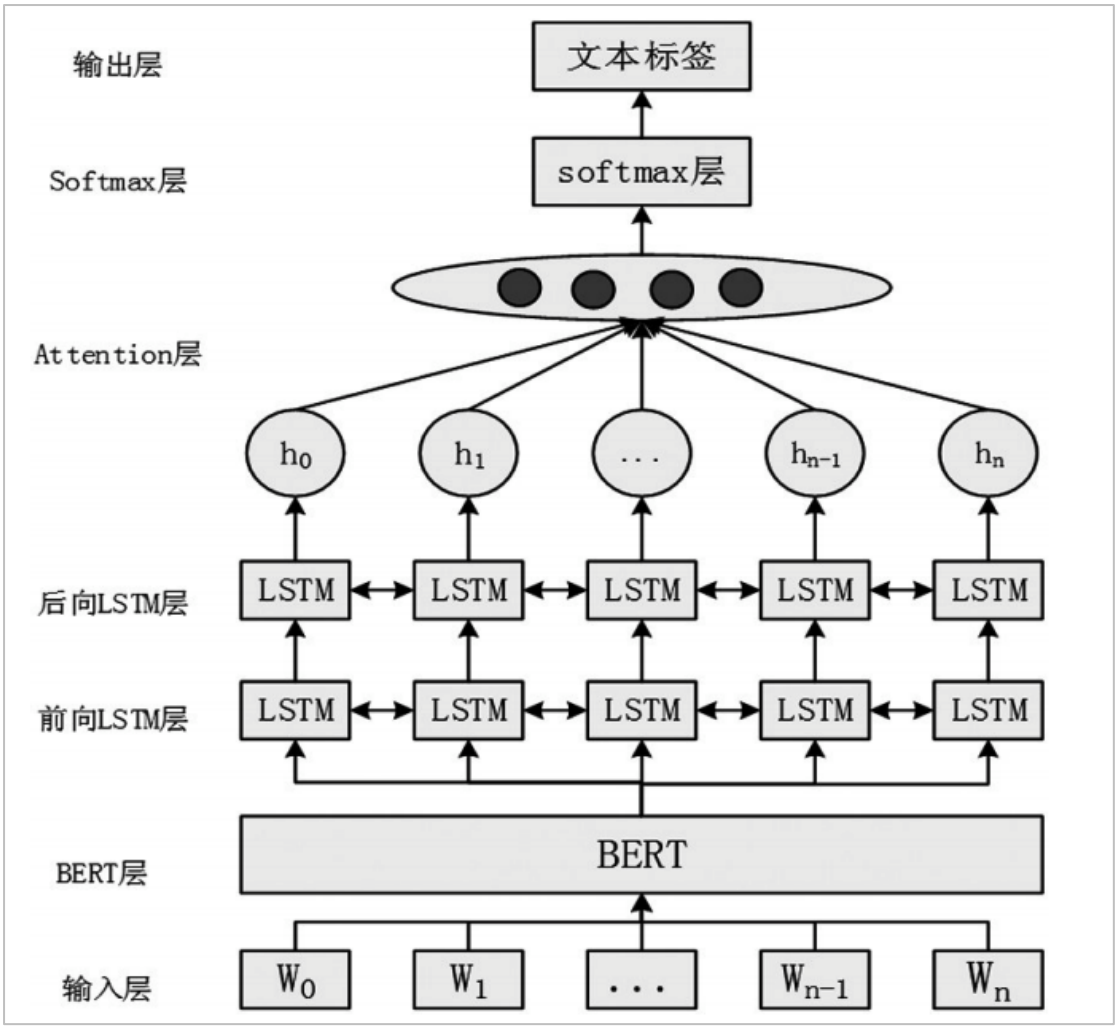

图 1：BERT-Attention-Bi-LSTM 模型

在 BERT-Bi-LSTM 基础上增加 Attention 层的目的在于生成注意力向量。通过与输入向量进行相似性计算，更新各个维度的权重值，提升重点词语在句子中的价值，使模型将注意力

集中在重点词上，降低其他无关词的作用，进一步提高文本分类的精度。实际评测中能提升约 2 个百分点的 F1 值。

3.3 基于概念后缀词的规则匹配结果扩充及合并

通过 3.2 工作的开展，已经识别了大部分最终提交的上下位关系集。为进一步优化结果，增加本节作为后处理步骤，对生成的结果进行增补与删减，主要采用基于概念后缀词的匹配。观察语料可分析，中文的部分概念定义习惯性的将语言元核放在概念词序列的后面，即后缀词相似的概念可能具有一定的上下位关系。具体做法是对概念集的任意一对概念进行遍历和后缀匹配，相同或同类后缀词则加入作为作为结果。图 2 展示了部分基于概念后缀词的判定源码。

```
if ent[-1] == '病':
    map_entity_concept[ent].append("疾病")
elif ent[-1] == '炎':
    map_entity_concept[ent].append("疾病")

elif ent[-1] == '素':
    map_entity_concept[ent].append("药物")
    map_entity_concept[ent].append("药品")
elif ent[-1] == '菌':
    map_entity_concept[ent].append("微生物")
    map_entity_concept[ent].append("生物")
    map_entity_concept[ent].append("细菌")
elif ent[-1] == '属':
    map_entity_concept[ent].append("科学")
    map_entity_concept[ent].append("自然科学")
    map_entity_concept[ent].append("生物")
elif ent[-3:] == '综合症':
    map_entity_concept[ent].append("综合症")
    map_entity_concept[ent].append("疾病")
elif ent[-1:] == '症':
    map_entity_concept[ent].append("疾病")
elif ent[-1:] == '散':
    map_entity_concept[ent].extend(['中药', '药品', '药物'])
elif ent[-1:] == '膏':
    map_entity_concept[ent].extend(['中药', '药品', '药物'])
```

图 2：基于概念后缀词判定的上下位关系发现

4 实验结果与分析

4.1 实验任务及评测指标

本任务聚焦新冠知识图谱中概念实体的上下位关系识别。任务提供了约 20000 个实体、1000 个概念（类型），要求基于上下位关系，识别出实体-概念之间的类型关系，以及概念-概念之间的上下位关系，前者是后者的子概念。

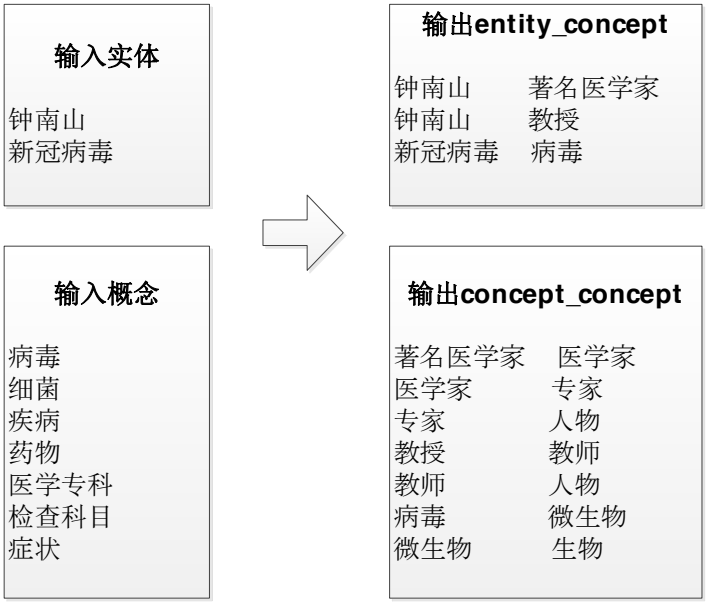

图 3：上下位关系输入输出样例

任务本身的设置是无监督的，因此不提供训练集。测试集是主办方通过自动化实体类型推测和人工检验进行标注的，采用公开数据集的子集作为测试集（500 左右实体，500 左右概念）。任务采用精确率（Precision, P）、召回率（Recall, R）、F1 值（F1-measure, F1）来评估效果。

4.2 模型及参数设置

本文最终采用 BERT-Attention-Bi-LSTM 模型作为上下位关系预测模型。作为实验对照，引入了 BERT- Bi-LSTM 模型作为比较。参数设置如表 2 所示。

表 2：模型参数设置

模型	BERT-Attention-Bi-LSTM	BERT- Bi-LSTM
句子最大长度	100	128
Batch Size	32	32

学习率	2e-5	2e-5
迭代次数	3	3
激活函数	gelu	gelu
隐层大小	768	768
隐层层数	12	12
Attention 层大小	100	

4.3 实验结果

按照 80%/20%的比例划分训练集和测试集，BERT-Attention-Bi-LSTM 模型的预测结果 F1 值为 67.3%，BERT- Bi-LSTM 模型的预测结果 F1 值为 65.4%。提交到线上基于验证集评测的结果为 0.484392619341443，这说明了前期识别的种子词样本仍然存在一些噪声。

6 总结与展望

针对医疗实体概念的上下位关系自动识别问题，本文中提出了一种基于模式扩充及 BERT 分类的上下位关系预测的方法，基于模式扩充从外部数据中抽取潜在的上下位关系对，而后利用 BERT-Attention-Bi-LSTM 模型对潜在上下位关系对进行预测，获得主要上下位关系预测结果。最终，与基于概念后缀词的规则匹配结果扩充结果一起，合并形成提交结果，在 CCKS2020：新冠知识图谱中概念实体的上下位关系识别任务中评测的 F1 值为 0.484392619341443，在所有提交中排名第一。从模型训练过程的评估指标与线上验证集评测指标之间的差异可知，模型的训练样本仍然混杂了较多噪声，如何提升训练种子整负例的准确性，并尝试多标签分类以及其他如“基于词嵌入投影”等模型方法，是后续的提升改进方向。

参考文献

[1]Hearst M A. Automatic acquisition of hyponyms from large text corpora[C]//Proceedings of the 14th conference on Computational linguistics-Volume, Association for Computational [2]Linguistics,1992: 539-545.Ritter A, Soderland S, Etzioni O. What Is This, Anyway: Automatic Hypernym Discovery[C]//AAAI Spring Symposium: Learning by Reading and Learning to Read,2009: 88-93.

[3]Anh T L, Kim J, Ng S K. Taxonomy construction using syntactic contextual evidence[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language

Processing (EMNLP),2014: 810-819.

[4]Ritter A, Soderland S, Etzioni O. What Is This, Anyway: Automatic Hypernym Discovery[C]//AAAI Spring Symposium: Learning by Reading and Learn ing to Read,2009: 88-93.

[5]Anh T L, Kim J, Ng S K. Taxonomy construction using syntactic contextual evidence[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),2014: 810-819.

[6] Wang Q, Xu C, Zhou Y, et al. An attention-based Bi-GRU-CapsNet model for hypernymy detection between compound entities[C]//2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).IEEE, 2018: 1031-1035

[7] 孙佳伟,李正华,陈文亮,张民.基于词模式嵌入的词语上下位关系分类[J].北京大学学报(自然科学版),2019,55(1):1-7.

[8]刘燊. 面向《大词林》的中文实体关系挖掘[D].哈尔滨工业大学,2016.

[9]周文烨,刘亮亮,张再跃.融合多层注意力机制与双向 LSTM 的语义关系抽取[J].软件导刊,2019,18(07):10-14+18.]