基于 BERT 的新冠概念图谱上下位关系预测方法

钟嘉伦[1]，孙昊海[1]，刘宇航[1]，韩旭[1]

华中科技大学, 湖北武汉, 430074[1]
{zhongjl,haohais,lyuhang,hans}@hust.edu.cn

摘要 知识图谱中的概念关系上下位预测能够将层次化的信息引入从文本提取的实体与概念之中，从而有助于上层应用。本次全国知识图谱与语义计算大会（CCKS）针对新冠领域的上下位关系预测设立了一项评测任务，即判断医药、疾病、症状、病菌相关的概念与实体之间的从属关系。针对这个任务，本文提出了一种基于词向量聚类的顶层设计与预训练模型结合的方法，在最终测试集上取得了 0.48178 的结果，排名第二。

Keywords: 新冠知识图谱 · 关系预测 · BERT

1 介绍

如何表示人类知识是人工智能研究中重要的一环。知识图谱近年来被研究者们重点关注，其通过一系列三元组（head/subject, relation/predicate, tail/object）去表示知识，被应用于搜索、推荐、问答等领域。为了更好的服务上层应用，如何构建一个完善的知识图谱是研究的重点之一。实体的概念信息预测、概念的上下位关系预测是构建完善的知识图谱的重要任务之一。2020 年，新冠疫情爆发，为了更好的“科技抗疫”，2020 年 CCKS 组委会发布了“新冠概念图谱的上下位关系预测”任务。

由于赛道没有给出任务的标注数据，因此我们通过爬取了相关网页的构建了一份基本数据集。同时为了发掘概念词汇之间的深层语义信息，我们通过预训练的词向量对概念和实体聚类。利用预构建的数据集，基于预训练模型 BERT[5]，设计了新的概念上下位预测方法，在最终评测中取得 0.48178（第二）的成绩。我们的工作主要为以下几点：

- 爬取了相关网页，构建了一份基础数据集；
- 利用预训练词向量进行概念实体聚类；
- 基于 BERT 设计了上下位关系预测的深层模型。

2 相关工作

2.1 模板匹配

早期 Hearst 等人 [2] 通过正则模板匹配的方式来挖掘实体的上下位关系，如 “w 是 [一个 | 一种] h”。但是由于中文语法的灵活性，这种方式只能提取出少数的关系，而不能覆盖大多数情形，除此之外，这种方法对距离较长或者存在修饰词的实体并不适用，模板难以判断实体的边界位置。

2.2 向量嵌入

由于采用固定模式的抽取方法泛化能力弱，上下位关系抽取的准确率很高但召回率很低，进而导致 F1 值很低。为了实现概念的上下位关系自动抽取，Fu 和 Guo 等人 [3] 在 2014 提出一种基于词向量映射的上下位关系预测模型，其采用概念的词向量来构建语义的层次化结构，通过学习从下位词到上位词的语义映射模型，可以实现将下位词的词向量投影映射至其上位词的词向量，进而用于预测概念的上位概念词汇。相较于传统的规则匹配方法，词向量中蕴含着前后文语义信息，保证了词汇上下位关系预测的准确性，此外，基于词嵌入映射的方法使得预测模型能够泛化到新的词汇，进而提升模型的鲁棒性和预测召回率，[3] 的实验结果也证实了采用词嵌入映射的方法能够极大提升上位概念预测的 F1 值。Wang 等人 [4] 则通过迭代学习的方式增加了预测的准确率。

2.3 预训练模型

随着 2018 年底 Google 提出预训练模型 BERT[5]，大量的预训练模型被提出，如 XLNet[6]、ALBERT[7] 等。这些预训练模型融合了大量的上下文信息，对语义信息作出了不错的表达，能够在很小的代价下迁移到新的学习任务上。KG-BERT[8] 就利用预训练模型进行知识图谱的补全，并取得了不错的结果。

3 方法

3.1 数据构建

为了利用在线的医学语料和医疗数据库知识，本方案从多个医学数据源进行了数据爬取和数据下载，主要的数据来源如下表：

表 1. 外部数据源。其中，前 5 行为医学名词，如疾病、药品的数据库，第 6 行为词向量数据库

数据源	数据形式
万方中医知识数据库	中药、疾病种类
哈工大词林	实体、概念的层次结构
思知知识图谱	实体、概念的描述与属性
Github	疾病、症状、药品列表
微医（挂号网）	疾病、药品的科室结构
Chinese Embedding	中文词向量[1]

根据表 1 中的中文词向量，给比赛中的概念词构建词嵌入表示，由于约有 400 个概念词并不包含在中文词向量的词库中，因此需要对这些概念词进行处理，本方案采用如下规则给概念词汇构建词向量：

- 如果概念词汇出现在中文词向量库中，那么直接使用该中文词向量；
- 如果概念词汇未出现在中文词向量库，那么分为以下两种情况：
  - 将词汇分割为前后两部分（即子串）s_1 和 s_2，两个子串都在中文词向量库中，对应的词向量分别为 e_1 和 e_2。那么词汇的词向量表示为两个子串词向量的平均：
$$e = \frac{e_1 + e_2}{2} \tag{1}$$
如**化学试剂**可以分割为**化学**和**试剂**两个子串，那么**化学试剂**的词向量为两个子串词向量的平均。
  - 如果词汇无法分割成两个子串使得两个子串都在词向量中，那么取词汇的最长后缀 s_{suf}，保证该后缀字符串在中文词向量库中，然后对整个词汇统计所有的 1-gram、2-gram、3-gram 和 4-gram 字符串，得到 n-gram 字符串集合 $\{s_1^{gram}, s_2^{gram}, ..., s_N^{gram}\}$。那么概念词汇的词向量表示为后缀字符串以及 n-gram 字符串词向量的加权平均：
$$e = 0.5e_{suf} + 0.5 * \frac{1}{N}\sum_{i=1}^{N} e_i^{gram} \tag{2}$$
如：**抗偏头痛药**的最长后缀为**头痛药**，但是**抗偏**不在词向量库中，因此，统计所有所有的 n-gram 子串集合 {抗，偏，头，...，偏头，头痛，头痛药，...}。

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

根据以上概念词汇的词向量，可以使用 sklearn 库中的 KMeans 进行聚类分析，聚类的簇数可以人为定义。设定聚类簇为 6，得到的聚类结构采用 PCA 降维并可视化的效果如下：

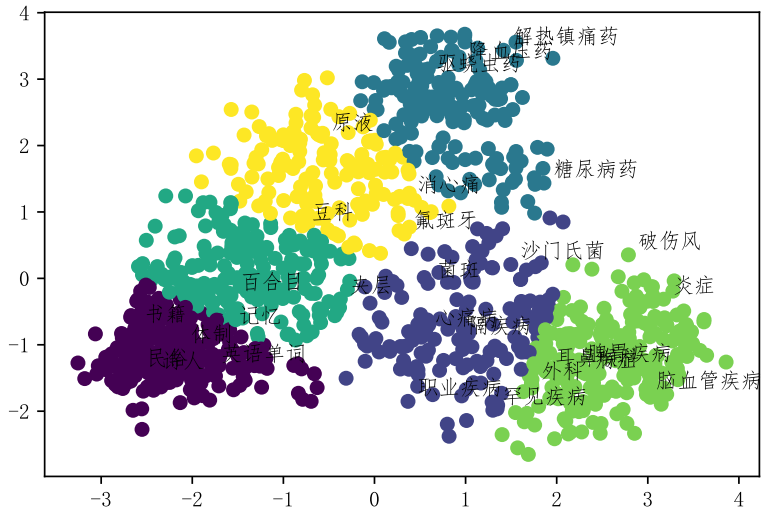

图 1. 采用 PCA 降维和 KMeans 聚类的可视化效果。

从聚类结果中可以看出，词向量能够较为清晰对概念词汇进行类别的区分，比如药品、疾病、工具等等。通过数据分布，我们按照如下的方式提取并构造了一部分上下位关系的数据样本：

- 通过后缀与规则获取数据，如“甲类传染病”属于“传染病”；
- 对思知知识图谱的实体/概念描述，采用模板提取一部分数据。对思知知识图谱的部分实体/概念属性，直接获取数据样本；
- 对结构化的数据源，直接获取数据样本；
- 跟据词向量的相似程度判断同义/近义名词，进行数据增强；
- 将聚类结果视为宏观知识，人工构建部分领域的上下位关系，如人属于生物，学科属于抽象事物等。为了防止构建的上下位关系对公平性的影响，我们粗略统计了 entity 中数据占比分布，植物与菌类约共占 5%，所以我们主要构建了植物、菌类的上下位关系、少量抽象事物的上下位关

系（主要为概念图中的根部概念）以及通过先验知识获取的少量稀疏的疾病上下位关系

3.2 预训练模型

由于预训练模型具有感知上下文的能力，我们采用 Bert[5] 作为上下位关系判断模型。如图 2 所示，对两个不同的实体/概念，我们将实体/概念的名称、描述、属性值进行拼接，组成它的上下文。接着用 “[SEP]” 进行分隔并在头部加上”[CLS]” 标记，模型的最后一层的头部输出经过一个全连接之后得到上位关系、下位关系以及无关系的概率。注意到头实体与尾实体具有先后顺序，所以我们的标签也是对称的。当下位关系得分高于某阈值且上位关系得分低于某阈值时，我们判断输入模型的头实体是尾实体的下位词。

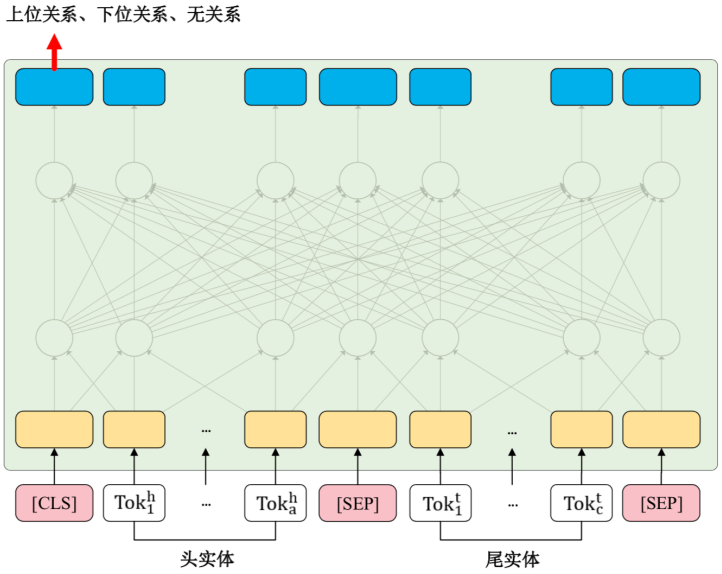

图 2. 预训练模型的结构表示

3.3 模型融合

通过 3.2 中的模型结构，采用不同的随机种子以及不同的取样策略分别训练模型，最终采用模型融合的方式得出最终的结果，由于计算资源的限制

以及本地验证的结果，本方案主要做了多个文件的投票，并按照出现占比阈值来判别是否为正样本。最后，采用人工规则来补充提交的数据。

4 实验结果

我们首先采用了向量映射的方式来判断上下位关系。我们尝试了 [3] 中提到的方法，但是该方法在本地中表现不佳。经过分析，原因大致有如下几点：

- 新冠知识图谱的数据具有比较强的领域性，而领域性的实体/概念对应的词向量出现概率较低，训练不充分。同时基于词林的训练集难以迁移到该数据集；
- 词向量可能具有较高的语义相似度，但不一定具有上下位关系；
- **模型趋向于将词向量映射至类簇的中心**：存在一种现象，就是将下位词词向量映射到上位词空间时，容易不断重复出现同一词汇，其语义相似但不具有上下位关系，有可能**模型趋向于将词向量映射至相似语义簇中心，而不是映射至上位词语义空间**。；
- 多义词以及 k-gram 组合词带来的问题，如**门**具有建筑的属性，也可以被认为是分类学中“界门纲目科属种”的一级，当它作为组合元出现在概念“**被子植物亚门**”之中时，在 top5 上位概念之中会引入**建筑物**这类概念

所以，我们只选用了词向量带来的领域知识，构建了一部分测试样本作为预训练模型的数据，从表格 2 之中可以看出，在使用了预训练模型之后，测试集的得分有了显著的提高，也证明了方法的有效性。

表 2. 不同训练方式的实验结果

训练方式	**提交得分（F1）**
开源数据	0.372
开源数据 +Bert	0.430
开源数据 +Bert+ 融合	0.477
开源数据 +Bert+ 融合 + 规则	**0.482**

5 总结

我们根据现有的方法与模型，提出了一种基于词向量聚类的顶层设计与预训练模型结合的方法，同时分析了不同方法的优劣之处，实验结果也证明了提出方法的有效性。当然我们的方法也存在着不足，如依赖于高质量语义、对多义词与实体的处理不够细致等，这也是之后可以研究改善的方向。

参考文献

1. 哈工大同义词词林 (扩展版), http://www.ltp-cloud.com/download.
2. Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. The 15th International Conference on Computational Linguistics.
3. Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1. 1199–1209.
4. Chengyu Wang, Yan Fan, Xiaofeng He, Aoying Zhou. 2019. Predicting hypernym-hyponym relations for Chinese taxonomy learning. Knowl. Inf. Syst. 58(3): 585-610.
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
6. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. NeurIPS 2019: 5754-5764.
7. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In International Conference on Learning Representations (ICLR).
8. Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. Computing Research Repository, arXiv:1909.03193.