基于知识增强的上下位关系识别

李如寐，王思睿，张鸿志

美团, 北京市朝阳区 100020
{lirumei,wangsirui,zhanghongzhi03}@meituan.com

Abstract. 上下位关系识别是大多数知识图谱构建的必备环节，但图谱中的节点数量往往很大，导致数据标注成本增加。在本文中，我们提出了一种结合开源数据与知识增强的上下位关系识别系统。系统主要有两个模块构成：爬虫模块和关系识别模块，爬虫模块负责爬取互联网中的开源数据，关系识别模块则利用爬取的知识进行关系识别。实验结果表明，我们所爬取的无监督知识对于上下位识别取得了一定的效果，F1达到了24.24%。

Keywords: 知识图谱，上下位识别，预训练语言模型

1 引言

上下位关系识别不仅需要模型具备浅层语义的理解能力，更需要模型捕获到词语的深层次知识。但本次测评数据集中的词语普遍较短，且存在很多专有名词，直接使用在开源数据上预训练的BERT很难达到好的效果。为了获取更多知识，同时节省标注成本，我们提出了基于开源数据与知识增强的上下位关系识别系统。系统中包含两个模块，首先是负责爬取开源数据的爬虫模块，之后是基于知识增强进行关系识别的模块。我们利用互联网上的海量知识，对模型对词语理解能力进行增强，使系统可以识别到各类专有名词的上下位关系。

2 系统

我们的系统流程如图1所示，分别是爬虫模块和上下位关系识别模块。

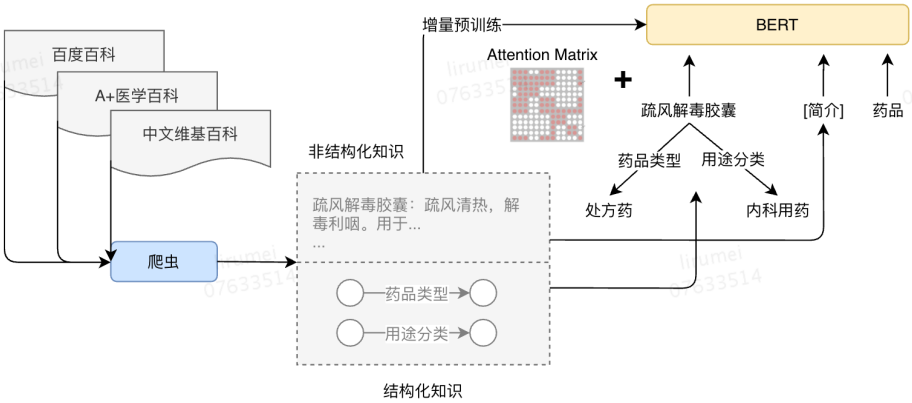

Fig. 1. 基于知识增强的上下位识别系统流程图

2.1 百科爬虫

因为测评没有提供监督数据，所以我们选择通过爬虫对百度百科、A+医学百科和中文维基百科三个网站进行爬取，尽可能收集监督数据与知识。监督数据的获取主要通过以下两种方法：

方法1 通过百度百科的简介与结构化卡片获取上位词标签。在百度百科网站的简介或者下方的结构化卡片中，部分时候会存在上位词信息。我们用测评提供的上位词词典进行匹配，可以得到一些监督数据。

方法2 通过匹配获取上位词标签。对于部分领域内词汇，从字面上就可以获取到部分信息，例如“益肝灵胶囊”是“胶囊”的下位。我们根据测评数据进行两两匹配，根据包含关系得到部分上下位。

2.2 上下位关系识别

通过百科爬虫和部分规则，我们获取了一些有监督数据。之后，我们选择基于BERT进行上下位识别。由于我们采用的开源BERT是基于通用语料训练的，而测评数据中有较多的医疗类实体。因此我们将开源BERT在爬取的非结构化数据上进行了增量预训练，提升BERT对医疗类实体的理解能力。

实验中我们发现，只通过词语本身进行上下位识别的效果并不好，模型很可能只是单纯地记忆上位词。因此我们在输入中拼接了对下位词的简介和百科卡片信息，卡片信息是以键值对的形式存在的，拼接时我们参考了K-BERT的方式，将三元组加入句子中，同时修改注意力矩阵，只有实体本身可以与结构化节点进行计算。上述方法让模型的输入包含更多背景知识，得到了一定的效果提升。

3 实验

在上下位关系识别中，我们对于模型进行了多种尝试，效果如表1所示：

Table 1. 上下位识别系统的效果

模型	F1
BERT上下位识别模型	0.2133
+模型预训练	0.2265
+词语简介	0.2308
+结构化知识	0.2424

可以看到，对输入增加知识后模型效果有明显的提升。说明上下位关系的识别需要更多语义和知识信息。

参考文献

1. Roller S, Kiela D, Nickel M. Hearst patterns revisited: Automatic hypernym detection from large text corpora[J]. arXiv preprint arXiv:1806.03191, 2018.
2. Chang H S, Wang Z, Vilnis L, et al. Distributional inclusion vector embedding for unsupervised hypernymy detection[J]. arXiv preprint arXiv:1710.00880, 2017.
3. Chen H Y, Lee C S, Liao K T, et al. Word relation autoencoder for unseen hypernym extraction using word embeddings[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 4834-4839.
4. Held W, Habash N. The Effectiveness of Simple Hybrid Systems for Hypernym Discovery[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3362-3367.
5. Wang C, He X. BiRRE: Learning Bidirectional Residual Relation Embeddings for Supervised Hypernymy Detection[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 3630-3640.
6. Liu W, Zhou P, Zhao Z, et al. K-BERT: Enabling Language Representation with Knowledge Graph[J]. arXiv, 2019: arXiv: 1909.07606.