混合语义相似度的中文知识图谱问答系统

骆金昌, 尹存祥, 吴晓晖, 周丽芳, and 钟辉强

深圳市南山区粤海街道海天一路 6 号
{764666804,553658947,343270711,1113697731,1032079470}@qq.com

摘要 为了解决知识图谱问答中实体识别和答案路径匹配的语义漂移问题，在文本中，我们提出了一种混合多种语义相似度的中文知识图谱问答系统。系统主要由四个组件构成：指称识别、实体链接、模板匹配和路径排序，各个组件充分融合了问题和答案之间的字词粒度与句子粒度的词义和语义混合相似特征。实验结果表明，我们提出的系统在测试数据集上取得了很好的泛化效果，F1 达到了 73.54%。

Keywords: KBQA · Entity Linking · Mention Recognition · Sematic Matching.

1 引言

在本文中，我们提出了一个高效精准的中文开放式图谱问答系统。在该系统中，我们提出四个创新的组件来解决图谱问答问题。首先是实体指称识别，通过构造指称词典来切分问句中的文本，召回大量实体指称，并用多个策略进行指称的粗力度筛选，保证了在保留正确指称的基础上，尽可能地去掉会引入噪音的指称。其次是焦点实体排序，我们使用了一个模型对指称所对应的所有实体进行了排序，在 dev 集上精度达到了 94%。最后通过排序后的实体，选取 Top2 的实体，进行路径的生成，并且通过模型来对路径进行排序。路径排序是一个计算匹配得分的过程，得分越高，就越有可能是正确的路径。因此，模型对问句和路径的匹配的建模能力就是效果提升的关键一环。文本的匹配包括了字面匹配，统计数值匹配和语义匹配，对于开放式图谱问答来说，通用语义的匹配显得尤为重要，这是因为开放式问句中使用的问法和词语是不可能完全存在于训练数据中的。对于通用语义，我们使用了词向量和 BERT 句子向量，使得整体模型的精度得到进一步提升。

2 系统

我们提出的系统的框架流程如图 1所示, 整个系统包含了四个主要的组件: 指称识别、实体链接、模板匹配、路径排序。

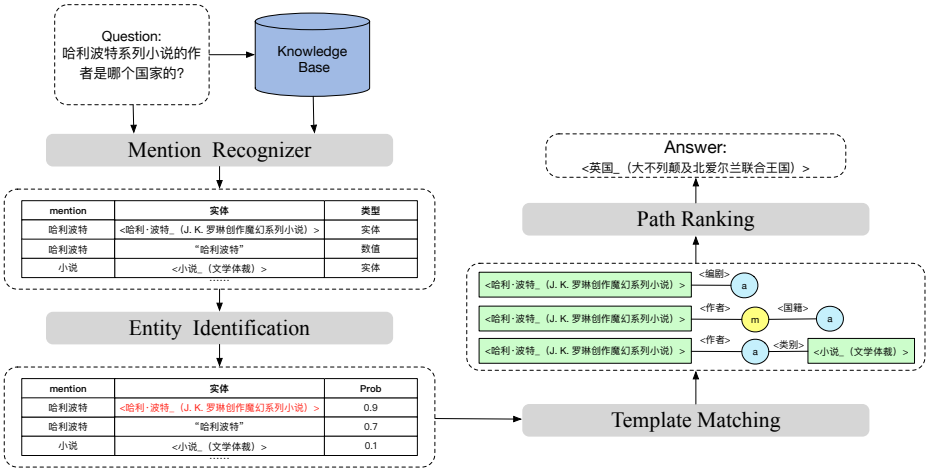

图 1. 系统整体流程图, 本系统包括指称识别、实体链接、模板匹配和路径排序四个组件。

2.1 指称识别组件

我们提出三种识别指称的方法。指称识别的目标是召回候选的实体列表，因为候选实体列表的质量会影响实体链接的效果，甚至会导致真实的实体被遗漏，因此识别的指标即要保证精度，也要尽量提高召回率。

方法 1 通过子串匹配的方法识别指称。先生成问题的全部子串，再从指称库召回所有匹配的指称。为了提高召回的精度，我们提出了多种最后剪枝的策略，如指称的长度至少为 2 个字符，指称不能被别的指称完全包含等。

方法 2 通过命名实体识别器召回指称。本系统采用命名实体识别器召回人名指称，进一步提高实体的召回率。

方法 3 通过启发式方法识别指称。方法 1 召回最大长度的指称，导致正确的指称被删除，例如问题"西湖景区的湖中二塔是指?"，方法 1 同时匹配中"西湖"和"西湖景区"两个指称，但"西湖景区"长度更长，因此正确的指称"西湖"被删除。针对该问题，我们召回短的指称对应的所有实体，再召回实体的一度关系，例如指称"西湖"，系统召回其中一个三元组为:（< 西湖 _（浙江省杭州市西湖）> ，< 湖中二塔 > ，" 保俶塔" ）。当一度关系名与问题完成匹配时，我们保留该指称。

2.2 实体链接组件

指称识别组件从问题中召回了多个候选的实体，实体链接组件需要从这些实体中找到问题最核心的实体，即焦点实体。我们采用排序模型来联合解决实体链接与焦点实体识别的任务。对于每个候选的实体，我们挖掘了多组特征，包含语义特征、字面特征等:

实体与问题匹配特征：候选实体的信号与问题越是匹配，该实体越可能是焦点实体。我们综合考虑了三类匹配度：实体名称与问题的匹配度、实体二度子图与问题的匹配度、实体类型与问题的匹配度。匹配度采用多种方式计算：集合距离、word2vec 语义相似度等。

流行度特征：我们定义了两种流行度：实体在图谱中出现的频率；实体不同的一度关系的个数。实体的出现的频次越高或者不同的关系边越丰富，说明该实体越重要。

指称重要度特征：指称越重要，说明对应的候选实体越重要。我们挖掘指称多种特征来表示重要度，例如指称是否被引号或书名号包含、指称是否在开头或结尾、指称跟疑问词的距离、指称是否包含数字或字母等。

其他特征：我们还挖掘了问题本身特征：问题字个数、问题词个数；实体本身的特征：实体全名是否在问题出现、实体的长度等。

我们实验了基于 pointwise 与 pairwise 算法，包括 LR、GBDT、RankSVM 等算法，最后我们选择了 LambdaRank 的排序算法。为了增加路径召回率，我们选择得分最高的 Top2 实体作为链指结果。

2.3 模板匹配组件

上一步骤识别出了问题中包含的图谱实体，接着我们对这些实体召回具体的子图，然后生成包含正确答案的路径。首先我们召回每个实体的二度关系子图。召回子图的边过多，容易降低路径排序模型的效果，因此我们提出多种剪枝的策略。

剪枝策略 1 根据实体的流行度剪枝。当实体的流行度超过一定阈值，慢删除该节点的关联边。例如"< 中华人民共和国 >"的一度关系数量为 707,879，如果保留所有的边，将大大降低排序模型的效果。

剪枝策略 2 根据边的方向剪枝。统计发现，某些路径的方向未在训练集中出现过，可删除这些路径，例如二度关系路径，方向为（入向，入向）。

我们共采用三种模板来生成答案的路径，分别为一度关系模板、二度关系模板和联合实体模板，具体见图 2，其中 e 表示链接的目标实体，r 表示

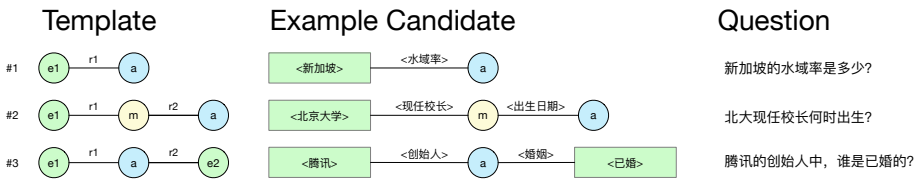

图 2. 本系统采用的生成候选路径的模板、路径样例以及相应的问题。模板 1、模板 2、模板 3 在训练集分别命中 54.9%、21.4%、8.1% 的问题。

关系，a 表示答案节点，m 表示中间节点。我们的模板方法和模板和论文 [1] 类似，但我们也提出许多创新点。

模板 1 匹配实体的一度关系，模型 1 匹配的路径都加入答案路径的候选集；

模板 2 匹配实体的二度关系，匹配中的路径数据非常巨大，因此我们提出了启发式的策略进行剪枝，例如关系 r2 的名称未在问题中出现，路径被丢弃；

模板 3 我们创新性地提出了联合两实体且关系为一度的模板，该模板命中训练集 8.1% 的问题。

2.4 路径排序组件

基于路径或者子图排序选择最优的匹配答案，是常用的图谱答问方法 [4][1]，我们采用传统的 learn-to-rank 排序模型对路径进行排序。对于每个候选答案路径，系统抽取了 39 个特征，主要包括：路径与问题的匹配特征、实体与问题的匹配特征、BERT[2] 语义匹配特征、答案类型匹配特征等。

路径与问题的字面匹配特征 我们先把问题和候选的答案路径表示成字符串，计算两个文本的字面相似度包含：Jaccard 距离、编辑距离等。为了提高匹配的精度，我们保留字分割和词分割的句子切分结果。

路径与问题的语义匹配特征 除了计算 fasttext 词向量词袋匹配特征，为了提高路径与问题的语义匹配能力，我们采用 BERT 抽取路径与问题的向量，并采用 cosine 函数计算两向量的相似度。在我们的训练集上，BERT 语义匹配特征可以提高约 1.5% 的排序精度。

答案类型匹配特征 正确答案的类型往往与问题的意图一致。例如问题“谁发明了电灯?”，该问题的意图是查询人物；“< 爱迪生 >”实体的类型是人物，那么该实体可能是正确答案。我们通过规则的方式构造了问题的意图分类器，包括五种意图：人物、地点、时间、数量和其他。最后因为知识库中的实体类型存在比较多的错误，我们通过启发式的方法对答案的类型进行分类。

实体链接的概率 实体链接采用的是 pairwise 的排序模型，故并不能直接得到链接的概率。我们提出一种创新的方法估计链接的概率。首先我们对链接器输出 Top20 个预测的实体进行打分，排序越前，得分越高；接着统计每个实体的总分，最终选择 Top2 实体；最后，对两实体的得分归一化得到链接的概率。

候选路径自身的特征 我们一共考虑了两类路径的特征。第一类是路径的匹配中的模板，显然模板 1 匹配中的路径相比于模板 3 匹配中的路径更可能是正确的路径。第二类是路径的方向类型特征，出向的类型比入向的路径类型正确的概率更大。

我们模型采用了 Pairwise LambdaRank 的排序算法，选取 Top1 路径对应的答案作为输出答案，如果候选路径为空，则无输出答案。另一方面，我们也实现了深度学习的排序算法，如 ESIM、BiMPM[3] 等算法，但在我们的场景下，深度学习算法的精度未超过 LambdaRank 算法。

3 实验

我们在 CCKS 2019 年的中文开放知识图谱问答数据集中验证了我们提出的方法。该数据集包含了一个问答数据集和一个开放知识图谱。其中，问答数据集包含了 2,298 条训练集，766 条验证集和 766 条测试集。其次，开放知识图谱使用了一个大型的中文知识图谱 PKU-base，该图谱包含了 41,009,141 条三元组，25,182,627 条实体类型和 13,930,117 条实体别称。本次评测的指标是 Macro F1。由于训练数据过少，为了减少模型方差，我们使用了 10 折交叉验证。

3.1 实体链接

表格 1展示了我们的实体链接模型的表现，结论如下：1. 我们最好的模型在 Top2 上准确率达到了 90.73%，2. 去掉 fasttext 的语义向量匹配后，效果衰退了 0.66%，说明通用语义对模型效果提升有一定作用。3. 去掉指称上下文的词袋特征，效果衰退了 0.92%。说明上下文的引入能够更好地帮助消歧。

3.2 路径排序

表格 2展示了我们的路径排序模型的表现，从中我们得到发下结论：1. 最好模型的 F1 值是 73.54；2. 在去掉 BERT 特征，效果衰退了 1.35，证明

表 1. 实体链接的 Acc@1- Acc@3

Model	Acc@1	Acc@2	Acc@3
Our Model	**82.11**	**90.73**	**92.95**
Our Model w/o fasttext	81.98	90.07	92.42
Our Model w/o mention context	80.93	89.81	92.16

表 2. KBQA 系统的的 F1@1

Model	F1	△
Our Model	**73.54**	
Our Model w/o BERT	72.19	-1.35
Our Model w/o BERT and fasttext	71.78	-1.76
Our Model w/o BERT and fasttext and bag of char	69.02	-4.51

BERT 的句子向量中的通用语义特征能带来很好的收益；3. 去掉了 fasttext 词向量特征，模型效果继续降低，证明 fasttext 的词向量携带了共现的语义相关性；4. 最后，去掉了路径与问题的词袋特征，效果大幅降低，证明模型能自动学习词组的字面匹配信息。

参考文献

1. Bast, H., Haussmann, E.: More accurate question answering on freebase. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 1431–1440. ACM (2015)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Wang, Z., Hamza, W., Florian, R.: Bilateral multi-perspective matching for natural language sentences. arXiv preprint arXiv:1702.03814 (2017)
4. Yih, W.t., Chang, M.W., He, X., Gao, J.: Semantic parsing via staged query graph generation: Question answering with knowledge base. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1321–1331 (2015)