计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(8):197−203 [doi: 10.15888/j.cnki.csa.007034]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

医学领域中基于注意力机制的查询扩展①

陈　素[1], 杨　燕[1], 胡琴敏[1,2], 贺　樑[1], 陈成才[3]

[1](华东师范大学 计算机与软件工程学院, 上海 200062)
[2](瑞尔森大学 计算机科学系, 多伦多 ON M5B 2K3)
[3](上海智臻智能网络科技股份有限公司 小 i 机器人研究院, 上海 201803)
通讯作者: 杨　燕, E-mail: yanyang@cs.ecnu.edu.cn

摘　要: 临床决策支持系统中, 通常使用电子病历中的病人描述作为查询检索, 进而辅助医生做决策分析. 我们提出了一个基于注意力机制的网络扩展查询方法以提高检索效果. 由于医学文本注释的难度和成本很高, 并受到了迁移学习理念的启发, 我们选择了非医学领域数据集学习句子与实体的关系, 迁移到医学领域数据集, 模型用 LSTM 获得句子表征并用注意力机制来获得实体表示. 我们提出的方法可以动态选择相关实体作为查询扩展, 同时我们不仅考虑单个实体作为扩展的影响, 也考虑了实体组合作为扩展的影响, 解决了选择固定数目实体的问题. 我们在 TREC Clinical Decision Support Track 三个标准数据集上进行实验, 实验表明本文提出的方法在实验结果上有显著的提升.

关键词: 查询扩展; 注意力机制; 迁移学习; 深度学习

引用格式: 陈素,杨燕,胡琴敏,贺樑,陈成才.医学领域中基于注意力机制的查询扩展.计算机系统应用,2019,28(8):197–203. http://www.c-s-a.org.cn/1003-3254/7034.html

Attention Based Network for Query Expansion in Medical Domain

CHEN Su[1], YANG Yan[1], HU Qin-Min[1,2], HE Liang[1], CHEN Cheng-Cai[3]

[1](School of Computer Science and Software Engineering, East China Normal University, Shanghai 200062, China)
[2](Department of Computer Science, Ryerson University, Toronto ON M5B 2K3, Canada)
[3](Shanghai Xiaoi Robot Technology Co. Ltd., Shanghai 201803, China)

Abstract: The aim of clinical decision support implementing electronic health records is to satisfy the physicians' information needs. We are motivated to propose an attention based network on query expansion. Considering the difficulty and cost of medical text annotation and inspired by the idea of migration learning, we chose the non-medical dataset for model training, and migrated to medical datasets. The model utilizes LSTM to obtain sentence representation and adopt attention mechanism to obtain entities representation. The proposed approach can dynamically select related entities as expansion of the query. At the same time, we not only consider the score of a single term as an expansion term, but also consider the score of term combination. We conduct the experiments on the three standard datasets of TREC Clinical Decision Support Track, where the approach has a promising overall performance over the strong baseline.

Key words: query expansion; attention mechanism; migration learning; deep learning

文本检索会议 (TREC) 临床决策支持任务 (CDS) 的目的是构建一个信息检索系统以支持临床决策. 系统接受由医生从电子病历 (EMR) 中总结的查询, 然后从在线的医学文献集合返回相关文献.

传统的信息检索系统是根据给定查询中关键词的统计量信息计算文章的相关性[1]. 根据文章的相关性得

① 收稿时间: 2019-02-20; 修改时间: 2019-03-08; 采用时间: 2019-03-18; csa 在线出版时间: 2019-08-08

分从高到底进行排序, 相关性最高的作为结果返回. 作为提高信息检索系统性能的有效方法之一, 查询扩展是将包含信息的词加入原始查询, 以便可以使用更多的统计信息来检索相关文章[2−4]. 在 TREC CDS 任务中, 大部分方法都是使用统计信息, 例如单词出现的次数, TF-IDF 分数等来进行扩展单词选择. 这种类型的方法不考虑在语义级别查询和扩展词之间的关系, 并且这类方法在选择扩展词数量的时候都是使用固定的 K 值, 但是不同的查询应该使用不同的 K 值. 另一种方法便是利用资源库对查询进行扩展, 例如 Zhu[5]先找到查询中的关键词, 再通过多个外部数据库对查询进行扩展. 国内的哈工大同义词林[6], 知网 (HowNet)[7], 国外的 WordNet[8]等也经常应用在查询扩展中. 但是在 Guo[9], Bacchin[10]的实验中发现, 使用医学知识进行查询扩展几乎没有声明改进甚至损害了检索性能. 矛盾的结果表明, 对于医疗信息检索而言, 并非所有的医学知识都适用于检索. 因此, 我们有动力去研究 CDS 任务中合适的医学知识对查询进行扩展.

本文提出的方法不仅考虑了统计量信息, 还考虑了查询和扩展词之间的关系. 考虑到医学文本注释的难度和成本, 我们使用非医疗数据集 STS 数据集作为训练数据来学习句子和实体之间的关系. 我们选择这些数据的原因是因为我们认为学习句子和实体之间的关系对计算查询和扩展医学实体之间的关系是有用的. 此外, 在测试部分, 我们不仅仅考虑单个医学实体作为扩展的影响, 还考虑了实体组合作为扩展的影响. 所以我们解决了使用固定 K 的问题. 因为 LSTM 能够很好的处理序列问题, 所以我们使用 LSTM 获得句子表征. 在实体部分, 我们使用注意力机制获得实体组表示, 因为注意力机制可以帮助模型集中于实体组中重要程度大的实体.

在本文, 我们提供了一种查询扩展方式来支持医学信息检索中的临床决策问题. 具体来说, 我们提出了基于注意力机制的神经网络来动态选择扩展医学实体. 我们利用了迁移学习的思想, 将在其他领域学习到的知识应用于医学领域, 以此来解决医学标注成本高, 标注难得问题. 此外, 我们选择最佳的医学实体组合作为扩展来解决固定 K 的问题, 并且能考虑到医学实体之间的影响. 本文方法使用注意力机制得到实体组表征, 能够让模型关注实体组中更重要的实体. 本文的其余安排如下: 我们首先介绍相关工作. 然后再介绍了提出的方法. 之后说明实验设置并展示和分析实验结果. 最后, 我们给出结论和展望.

1 基于注意力机制网络的查询扩展

本章将介绍在医学领域中基于注意机制网络的查询扩展方法. 该方法利用了迁移学习的思想, 将在其他领域学习到的知识运用到医学领域, 以此来解决医学领域中标注困难和成本高的问题. 由于在模型的训练和测试部分我们使用的是不同的数据, 所以我们将分两部分介绍概念. 在训练过程中, $S=\{w_1,w_2,\cdots w_n\}$ 用来表示 STS 数据集中的句子, w_i 代表 S 中第 i 个单词. $JE=\{je_1,je_2,\cdots,je_m\}$ 代表用 TagMe 工具进行标注的通用实体. 在测试过程中, $Q=\{qw_1,qw_2,\cdots,qw_n\}$ 代表查询, 其中的 qw_i 代表查询中的第 i 个词, $ME=\{me_1,me_2,\cdots,me_m\}$ 是利用 MeSH[11]词表映射得到的医学实体集合. 本文方法的目的是为 Q 选择合适的扩展医学实体. 在方法框架中, 我们利用谷歌搜索引擎和 MeSH[11]词表得到候选扩展实体集合. 为了在扩展候选实体中选择最佳的扩展实体, 我们不仅仅考虑到每个候选实体的得分, 我们还考虑到候选实体组合的得分.图 1 展示了本文方法的大体框架. 简单来说, 我们首先将原始查询提交到搜索引擎中并且选择前 N 个结果; 其次, 我们利用选择模型得到最佳的扩展实体; 最后, 我们将得到的最佳扩展实体加入原始查询中进行检索. 下面我们将分两部分介绍本文的方法, 一是如何得到候选扩展实体集, 二是如何选择出最佳的实体组合.

1.1 扩展候选医学实体组合

我们是利用网络资源和 MeSH 得到扩展候选实体. 我们选择使用网络资源的原因是我们认为搜索引擎返回给我们结果是和原始查询相关的并且结果是按照与查询的相关性从高到低进行排序的. 在之前的工作[12,13], 我们可以发现医学实体对医学查询有正向作用, 所以我们利用 MeSH 对查询进行映射, 得到扩展医学实体. MeSH[11]是一种广泛使用的医学本体数据库, 由 16 类医学概念组成. 作为外部知识资源, 如果医学实体可以在前 N 个搜索结果中找到, 我们就认为医学实体是候选扩展医学实体.

因为候选实体之间也是存在一定的联系和影响, 所以我们不仅仅考虑单个候选实体的影响, 而且考虑候选实体组合的影响. 考虑到计算的时间复杂性和空间复杂性, 我们只结合任何 $K\leqslant 5$ 个候选扩展医学实体

作为候选扩展医学实体组合, 这是因为如果将所有的扩展医学实体进行全排序, 则每个查询需要计算 2^n, 但是每个查询的候选扩展医学实体的数量都超过 10 个, 这样, 需要计算的实体组数量较大, 测试的过程需要消耗更大的空间和更多的时间. 例如, 我们有“chest pain, disease, fatigue, heart failure, hypertension, nausea”6 候选扩展医学实体, 那我们就会有C_6^1个只包含 1 个医学实体的扩展组合, C_6^2个包含 2 个医学实体的扩展组合$\cdots C_6^5$个包含 5 个医学实体的扩展组合, 其中C_n^m代表排列组合. 接下来我们便要计算每一个候选实体组合的得分.

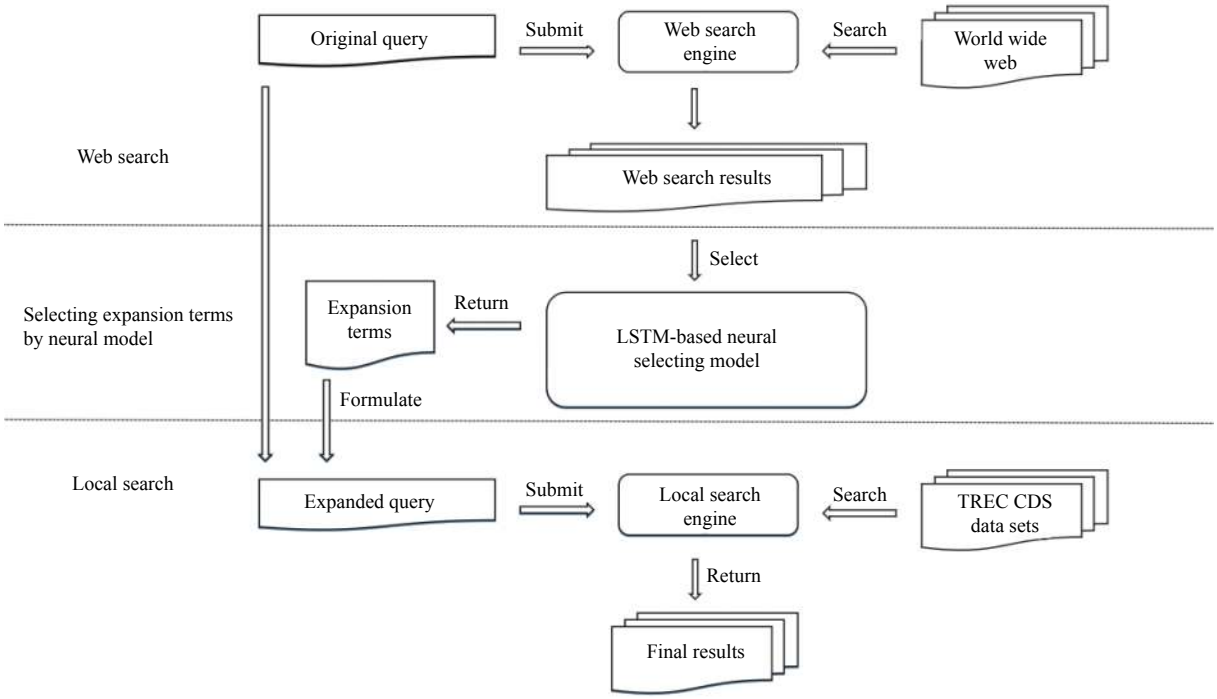

图 1 查询扩展方法框架

1.2 扩展医学实体组合的挑选

和之前只利用统计量信息不同, 我们使用神经网络模型自动的选择最佳扩展医学组合. 考虑到人工标注的困难和开销, 我们选择其他领域的数据集来训练模型. 我们之所以选择其他领域数据集训练模型是因为我们认为实体是句子理解的重要组成部分, 而医学实体是理解医学句子的重要组成部分, 所以学习到的实体和句子之间的知识也是可以用到医学实体和医学句子中. 图 2 显示了神经网络选择模型的结构, 包括了嵌入层 (Embedding Layer), LSTM 层 (LSTM Layer), 注意力层 (Attention Layer) 和预测层 (Predict Layer). 从图 2 我们可以看出, 最重要的部分是得到句子和实体组的表示.

嵌入层: 因为我们是学习其他领域的知识并将其运用到医学领域中, 所以只使用词嵌入是不合理的.如果我们只使用了词嵌入, 则在模型训练期间无法训练医学单词, 为了解决这个问题, 我们将词嵌入和字符嵌入结合起来表示句子和实体. 和词嵌入一样, 字符嵌入也是将每个单词映射到高维向量空间, 但是字符嵌入训练的是每个字符的向量. 我们利用卷积神经网络 (convolutional neural networks) 求得$S(Q)$, $JE(ME)$中每个词的字符嵌入. 在字符嵌入层中, 每个单词都被表示成$C \in R^{wl \times d}$, 其中wl表示单词长度, d代表向量维度. 我们将线下训练完成的词向量和字符向量结合, 得到每个单词的表示, 公式如下:

$$FW(w_i) = [C(w_i), W(w_i)] \quad (1)$$

其中$C(w_i)$和$W(w_i)$分别表示$S(Q)$和$JE(ME)$中每个单词的字符嵌入和词嵌入. $[a,b]$表示连接两个向量表示.

LSTM 层: 我们使用 LSTM 生成具有语义构成的句子表征. 众所周知, LSTM 是一种典型的递归神经网

络变体, 通过引入记忆细胞和门机制, 已被广泛用于长文档建模. 在每个位置t, 隐藏状态h_t以及记忆细胞c_t可以通过上一位置的隐藏状态h_{t-1}, 记忆细胞c_{t-1}以及当前位置的输入x_t来更新它们的信息. 公式如下:

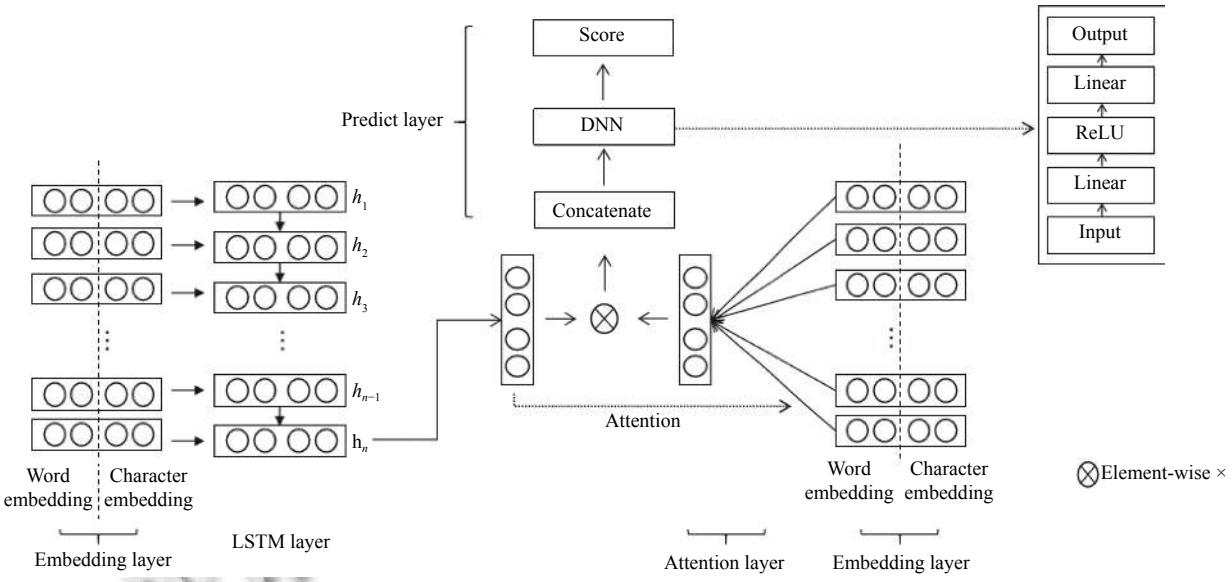

图 2 基于注意力机制的网络模型

$$i_t = \sigma(W_i x_t + U_i h_{t-1}) \quad (2)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1}) \quad (3)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1}) \quad (4)$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1}) \quad (5)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t \quad (6)$$

$$h_t = o_t \otimes \tanh(c_t) \quad (7)$$

其中, σ表示 Sigmoid 函数. i_t, f_t, o_t分别表示输入门, 遗忘门和输出门. 这些门用来保护和控制输入信息. 隐藏状态h_t表示当前位置信息的表征, 这个表征包含了当前位置的上文信息. $\otimes$代表点乘.

注意力机制层: 我们使用注意力机制来获取实体组中比较重要的实体. 受到神经机器翻译的影响, 注意力机制得到了广泛的使用, 因为它可以帮助模型集中注意力于输入信息中比较重要的部分. 实体组的表征S是所有实体的加权求和, 公式如下:

$$EP = \sum_{i=1}^{m} \alpha_i e_i \quad (8)$$

其中e_i代表第i个实体的表示, α_i表示第i个实体的注意力权重, 这个权重衡量了e_i在整个实体组中的重要性. α_i的计算公式如下

$$\alpha_i = \frac{\exp(E(e_i, FW))}{\sum_{K=1}^{m} \exp(E(e_k, FW))} \quad (9)$$

$$E(e_i, FW) = v^{\mathrm{T}} \tanh(W_h e_i + W_\alpha FW + b) \quad (10)$$

其中, E表示分数函数, 用来计算实体组中每个实体的重要性, v代表权重向量, v^{T}代表v的转置. W_h和W_a代表权重矩阵. FW代表查询的句子表征.

预测层: 在得到句子表征和实体组表征之后, 我们将句子表征和实体表征点乘并与句子表征, 实体表征拼接送入二层的深度神经网络, 具体公式如下:

$$y_1 = ReLU([SP, EP, (SP \otimes EP)] \cdot W_{v1} + b) \quad (11)$$

$$y = W_v \cdot y_1 + b_1 \quad (12)$$

其中, SP表示句子表征, 即 LSTM 层的输出. EP表示实体组表征, $\otimes$表示点乘, y表示实体组作为扩展的得分.

2 实验设置

这一章主要介绍实验的配置, 包括数据集、查询、参数设置以及评测指标.

数据集: STS 数据集包括了从 2012 年到 2017 年期间在 SemEval 中使用过的英文数据集的集合, 一共有 8628 个句子对. 表 1 显示的是根据类型和训练-验证-评测的具体划分情况. 在我们的方法中, 我们将 STS 的训练数据集和验证数据集都作为训练数据集, STS 的测试数据集作为验证数据集. 我们在 TREC CDS 2014, 2015 和 2016 上检验方法的效果. TREC CDS 中的文章都是来自于 Open Access Subset of PubMed Central

(PMC). PMC 是一个免费在线生物医学文章数据库, 所有的文章都是以 NXML 的形式呈现.

表 1 STS 数据集结构

	train	dev	test	total
news	3299	500	500	4299
caption	2000	625	625	3250
forum	450	375	254	1079
total	5749	1500	1379	8628

所有的实体都是用 TagMe[14]进行注释并且所有的注释都被保留. 这是 ClueWeb 上最广泛使用的基于实体的排名方法设置.

查询: 在 TREC CDS 中, 查询都是由专业查询开发者从患者的实际医疗记录中总结得到的自由文本. 检索到的文章对于回答每个查询的临床问题应该是有用的, 每一个 TREC CDS 任务都有 30 个查询.

参数设置: 我们使用 5 折交叉验证评估该方法. LSTM 设置: 隐藏状态设置为 100 维. 句子部分用 ReLU 激活函数. Batch 大小在{32, 64, 128}中选择, 学习率在{0.0005, 0.001, 0.002}中选择, 迭代次数在{5, 10, 20}中选择, 医疗实体组合的个数从{1, 2, 3, 4, 5}中选择. 另外, 我们为每个查询选择的是谷歌检索的前 10 个结果.

评测指标:在这次实验中, 我们选择 NDCG 和 P@10 作为评测指标. 其中 P@10 表示前 10 个检索结果的准确率, NDCG 则是衡量前 10 个检索结果排序的评测指标.

3 实验结果与分析

在一章中, 我们主要介绍比较方法和本文提出的方法在 2014 年, 2015 年, 2016 年数据集上的实验效果. 因为处理数据和建索引方法的不同, 并且在 TREC CDS 数据集中未有神经网络方法的实验结果可做对比, 为了验证本文方法的有效性, 我们选择了下面几种对比方法.

BM25: 只用原始 query 进行检索, 没有任何查询扩展. 检索模型为 BM25[15].

WebAssistance: 所有在谷歌检索结果中出现过的医学词都作为查询扩展.

WithoutCombination: 只单单选择前三个分数高的医学实体作为查询扩展. 该方法不考虑实体组合的效果.

LSTM+AdA: 在得到实体表示部分, 使用加权平均的方法获得实体表示, 并没有使用注意力机制的方法.

3.1 与对比方法的比较

表 2 展示了本文方法和对比方法在 TREC CDS 2014 年, 2015 年以及 2016 年数据上的实验效果. 从表 2 中可以看出, 和对比方法相比, 本文提出的方法有很大的提升. 本文提出的方法在 2014 年数据集上, NDCG 为 0.2521, 和 BM25 相比, 有 5.48% 的提升, P@10 值为 0.3, 和 BM25 相比, 本文提出的方法有 11.11% 的提升. 在 2016 年数据集上, 评测指标 NDCG 的值为 0.2172, P@10 的值为 0.2733, 同时我们可以看出, 和 BM25 相比, NDCG 有 9.42% 的提升, P@10 有 15.46% 的提升, 这是很显著的提升. 虽然在 2015 年数据集上的 P@10 的效果有所降低, 但是总体来说, 本文提出的方法是对提升检索的效果是有效的. 从表 2 中我们还可以看出, WebAssistance 的实验效果比 BM25 的效果好, 但是还是没有本文提出的方法效果好, 这是因为 WebAssistance 将所有的医学实体都作为查询扩展词, 这样会引入噪音, 这也就可以解释在 2015 年和 2016 年数据集上效果下降的现象. 同时, 我们也可以从表中看出, 考虑实体组合比只考虑单个实体的效果要好. 注意力机制也被验证是有效的, 因为 LSTM+AdA 忽略了实体之间的影响.

表 2 实验结果

	2014 TREC CDS Data Set		2015 TREC CDS Data Set		2016 TREC CDS Data Set	
Method	NDCG	P@10	NDCG	P@10	NDCG	P@10
Bm25	0.239(0)	0.27(0)	0.2893(0)	0.3933(0)	0.1985(0)	0.2367(0)
WebAssistance	0.2514(+5.19%)	0.3(+11.11%)	0.2608(–9.85%)	0.3333(–15.26%)	0.1962(–1.16%)	0.2333(–1.44%)
WithoutCombination	0.224(–6.28%)	0.26(–3.70%)	0.2815(–2.7%)	0.3633(7.63%)	0.1888(–4.89%)	0.2467(+4.22%)
LSTM	0.2489(+4.14%)	0.29(+7.14%)	0.3(+3.70%)	0.3833(–2.54%)	0.2053(+3.43%)	0.26(+9.84%)
LSTM+Attention	0.2521(+5.48%)	0.3(+11.11%)	0.2998(+3.63%)	0.3833(–2.54%)	0.2172(+9.42%)	0.2733(+15.46%)

3.2 参数学习

N是决定最大组合数量的参数. 考虑到计算时间复杂度和空间复杂度, 我们评测N从 1 到 5, 间隔为 1 对本文提出方法的性能影响. 根据实验结果, 在 2014 年和 2016 年, N最优为 4, 在 2015 年N的最优值为 2. 为了探寻这个结果的原因, 我们计算了 2014 年, 2015 年

和 2016 年查询的平均长度 (AL). 正如表 3 所显示, 我们发现 2015 年查询的平均长度是最短的, 同时 2015 年N的最优值为 2, 所以我们可以看出句子长度对N的选择是有影响的. 这是因为长查询需要更多的扩展词才能影响查询的效果.

表 3 各年份查询平均长度

	2014	2015	2016
AL	26	21	33

3.3 词嵌入和字符嵌入的比较

在本文中, 我们同样对词嵌入和字符嵌入对迁移学习的影响做了实验. 在这里, 使用词嵌入和使用字符嵌入的区别是在得到单词表示上. 词嵌入是一种向量训练, 首先通过预训练获得向量, 然后随着模型的训练进一步训练. 字符嵌入的方法是随机初始化字符向量, 然后随着模型的训练进一步训练出字符向量, 最后得到字符向量, 对字符向量的进一步操作才得到单词表示. 从图 3, 图 4 和图 5 中我们可以发现词嵌入和字符嵌入在一定程度上都是有效的. 在 2014 年和 2016 年数据集中, 使用词嵌入的方法要比原始的方法效果好, 使用字符嵌入的效果比使用词嵌入的效果好. 在 2014 年中, 只使用词嵌入方法的 NDCG 值为 0.2473, P@10 为 0.28, 从图 6, 图 7, 图 8 中我们可以看出和 BM25 方法相比, NDCG 有 3.47% 的提升, P@10 有 3.70% 的提升. 只使用字符嵌入方法的 NDCG 值为 0.2501, P@10 的值为 0.29, 和 BM25 相比, NDCG 有 4.64% 的提升, P@10 有 7.41% 的提升. 在 2016 年中, 同样可以发现, 使用字符嵌入的效果比使用词嵌入的效果要好. 这表明, 当我们将从其他领域学习到的知识应用于医学领域时, 字符嵌入的性能优于字嵌入. 这个结果是因为我们只使用了医学数据进行测试, 并没有使用医学数据进行训练, 所以在测试时会出现大量未列出的单词, 这也导致了词嵌入性能不显著. 但是无论是哪个领域, 里面的单词都是由字符组成, 并且字符表示会随着模型的优化而更新, 因此我们使用基于字符嵌入能获得更好的单词表示.

4 结论与展望

本文主要提出了一个基于知识的神经网络查询扩展模型以提高医疗信息检索效果. 考虑到医学文本注释的难度和成本, 我们将从其他领域学习到的知识应用于医学领域. 与以前的工作不同, 我们解决了固定K的问题. 我们不再是选择前K个扩展医学实体, 而是选择最佳的医学实体组合, 因此不同的查询可能有不同数量的扩展医学实体. 在本文中, 我们不仅展示了我们提出方法的有效性, 而且还比较了词嵌入和字符嵌入对迁移学习效果的影响. 在之后的研究中, 我们可以尝试使用不同的注意力机制得到实体组的表示以及不同的神经网络对查询扩展的影响.

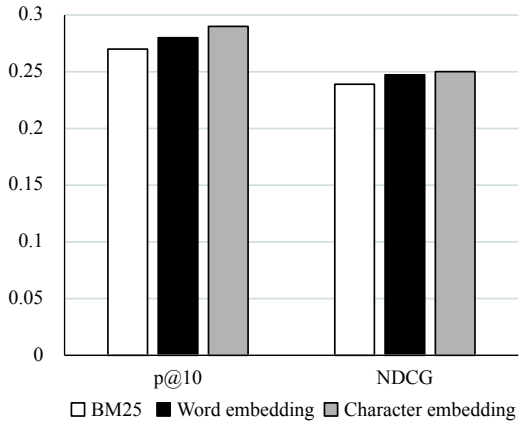

图 3 2014 年不同嵌入效果对比

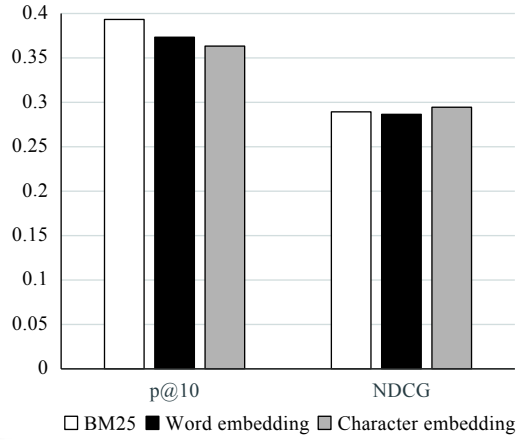

图 4 2015 年不同嵌入效果对比

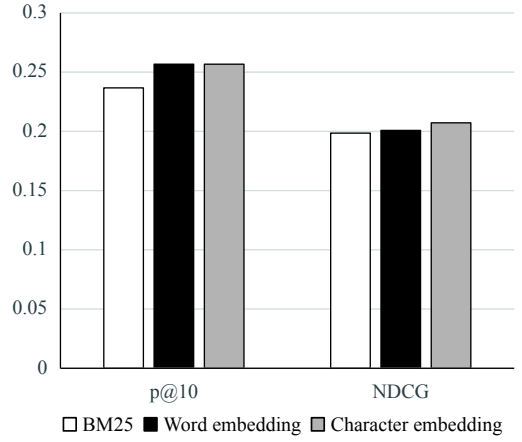

图 5 2016 年不同嵌入效果对比

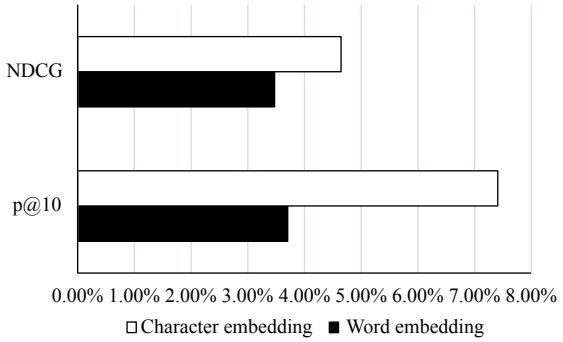

图 6 2014 年不同嵌入效果对比 (rate)

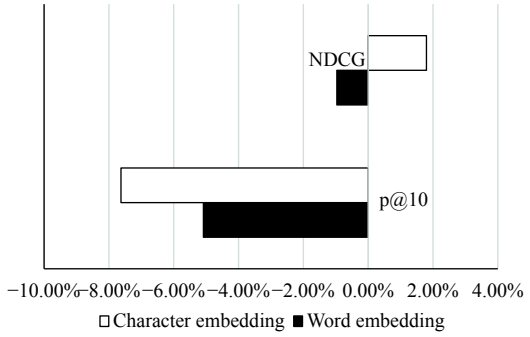

图 7 2014 年不同嵌入效果对比 (rate)

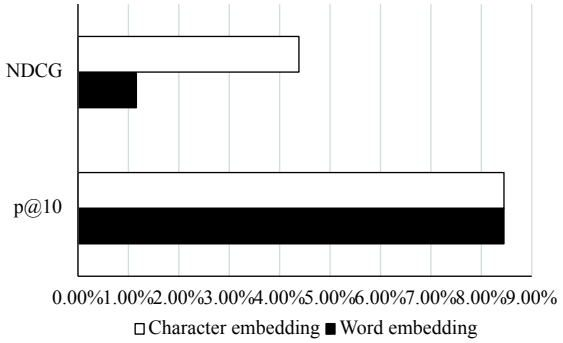

图 8 2014 年不同嵌入效果对比 (rate)

参考文献

1 Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 2009, 3(4): 333–389.

2 Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 1988, 24(5): 513–523.

3 Robertson S E. On term selection for query expansion. Journal of Documentation, 1990, 46(4): 359–364. [doi: 10.1108/eb026866]

4 Xu JX, Croft WB. Query expansion using local and global document analysis. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Zurich, Switzerland. 1996. 4–11.

5 Zhu DQ, Carterette B. Improving health records search using multiple query expansion collections. Proceedings of 2012 IEEE International Conference on Bioinformatics and Biomedicine. Philadelphia, PA, USA. 2012. 1–7.

6 赵静. 大规模汉语语义词典构建[硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2011.

7 Dong ZD, Dong Q. HowNet–a hybrid language and knowledge resource. Proceedings of International Conference on Natural Language Processing and Knowledge Engineering. Beijing, China. 2003. 820–824.

8 Miller GA, Beckwith R, Fellbaum C, *et al*. Introduction to WordNet: An on-line lexical database. International Journal of Lexicography, 1990, 3(4): 235–244. [doi: 10.1093/ijl/3.4.235]

9 Guo YK, Harkema H, Gaizauskas R. Sheffield university and the TREC 2004 Genomics track: Query expansion using synonymous terms. Proceedings of TREC. 2004.

10 Bacchin M, Melucci M. Symbol-based query expansion experiments at TREC 2005 genomics track. Proceedings of TREC. 2005.

11 Medical Library Association. Bulletin of the medical library association. Chicago: Medical Library Association, 1911.

12 Hoenkamp E, Bruza P, Song DW, *et al*. An effective approach to verbose queries using a limited dependencies language model. Proceedings of the 2nd International Conference on the Theory of Information Retrieval: Advances in Information Retrieval Theory. Cambridge, UK. 2009. 116-127.

13 Bendersky M, Croft WB. Discovering key concepts in verbose queries. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore. 2008. 491–498.

14 Ferragina P, Scaiella U. Fast and accurate annotation of short texts with Wikipedia pages. IEEE Software, 2012, 29(1): 70–75. [doi: 10.1109/MS.2011.122]

15 Manning CD, Raghavan P, Schütze H. An introduction to information retrieval. Cambridge: Cambridge University Press, 2008. 233.