文章编号：1003-0077(2018)05-0089-08

基于共指消解的实体搜索模型研究

熊 玲，徐增壮，王潇斌，洪 宇，朱巧明

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：实体属性挖掘(slot filling，SF)旨在从大规模文档集中挖掘给定实体(称作查询)的特定属性信息。实体搜索是 SF 的重要组成部分，负责检索包含给定查询的文档(称为相关文档)，供后续模块从中抽取属性信息。目前，SF 领域关于实体搜索的研究较少，使用的基于布尔逻辑的检索模型忽略了实体查询的特点，仅使用查询的词形信息，受限于查询歧义性，检索结果准确率较低。针对这一问题，该文提出一种基于跨文档实体共指消解(cross document coreference resolution，CDCR)的实体搜索模型。该方法通过对召回率较高但准确率较低的候选结果进行 CDCR，过滤不包含与给定实体共指实体的文档，提高检索结果的准确率。为了降低过滤造成的召回率损失，该文使用伪相关反馈方法扩充查询实体的描述信息。实验结果显示，相比于基准系统，该方法能有效提升检索结果，准确率和 F_1 分别提升 5.63%、2.56%。

关键字：共指消解；伪相关反馈；实体搜索

A Coreference Resolution Based Entity Search Model

XIONG Ling, XU Zengzhuang, WANG Xiaobin, HONG Yu, ZHU Qiaoming

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: The goal of Slot Filling (SF) is extracting certain attribute value of given entity(query) from large scale corpus. Entity search, as an important component of SF, retrieves documents referring to the given entity for other components to extracting attribute values from them. In contrast to the existing entity search based on boolean logic, we propose a cross document coreference resolution (CDCR) based entity search model. This CDCR improves the precision of IR results by filtering documents which do not contain mentions referring to the given entity. To minimize the loss of recall in filtering process, we introduce the pseudo relevant feedback method to augment the information of given entity. Experimental results show that our model outperforms the baseline by increasing the precision and F1 score by 5.63% and 2.56%, respectively.

Keywords: coreference resolution; pseudo relevant feedback; entity search

0 引言

SF 是文本分析会议(text analysis conference)发起的知识库构建任务(knowledge base population，KBP)下的第二项子任务。作为信息抽取的一个重要方向，SF 旨在从开放的数据源中抽取查询实体(简称查询)的属性值[1]。SF 系统包含三个基本模块：①文档检索模块；②属性抽取模块；③后处理模块[2]。其中，文档检索模块旨在从海量的文档集中检索出包含查询实体的相关文档；属性抽取模块旨在从相关文档中抽取出查询实体的候选属性值。例如对于查询实体“Alexander Haig(亚历山大·海格)”，从检索得到的文档包含的句子“Former US Secretary of State Alexander Haig died on Saturday in hospital at 85.(美国前国务卿亚历山大·海格于周六在医院去世，享年 85 岁。)”中可以抽取属性值如表 1 所示；后处理模块旨在从候选属性值

收稿日期：2015-12-06 **定稿日期**：2016-03-25

基金项目：国家自然科学基金(61373097, 61672367, 61672368)

中合并同一属性的结果,并筛选出最终属性值。

表1 查询"Alexander Haig"属性值

属性	属性值
title(头衔)	Secretary of State(国务卿)
date_of_death(死亡日期)	Saturday(星期六)
place_of_death(死亡地点)	Hospital(医院)
age(年龄)	85

文档检索的结果是SF的基础。因此,如何有效地提高文档检索模块的性能是提高SF整体性能的关键[3]。本文把SF中查询的相关文档检索问题定义为实体搜索问题,即给定一个实体(人或组织)的名称(称为查询),从指定文档集中检索所有包含该实体的文档,即相关文档。基于布尔逻辑的检索模型[4],在应用到实体搜索问题上时存在以下两个问题:①实体同名歧义问题,例如,在维基百科中,与查询"John Graham(约翰·格拉汉姆)"同名的人物共有25个。检索得到的文档虽然包含与查询实体名称相匹配的字符串,但是,该字符串在文中指代的实体不是给定的查询实体;②实体别名问题,即同一个实体可能对应多个别名。例如,对于查询"Public Library of Science(美国科学公共图书馆)"存在别名"PLoS",文档中可能出现与给定查询形式不同的别名,前者导致检索结果的准确率较低,后者导致检索结果的召回率较低。

在信息检索领域,通常使用查询扩展方法,在本问题上使用实体的别名(称为扩展查询)对查询进行扩展,以达到提高检索召回率的目的[3]。但是该方法无法解决问题1,且该方法会加重问题1的影响,因为使用的扩展查询本身也会存在歧义。例如,对于查询"National Restaurant Association(全国餐饮协会)"的扩展查询"NRA",还可能表示"National Rifle Association(全国枪支协会)"。目前,SF领域关于实体搜索的相关研究较少,虽然已经能够较好地解决问题2,但是尚未有较好的对问题1的解决方法。本文将重点研究问题1的解决方法。

针对实体搜索的实体同名歧义问题,本文提出一种基于跨文档实体共指消解的实体搜索方法,该方法以基于布尔逻辑的检索模型为基础,共分为两个阶段:

(1) 使用查询扩展方法对原始查询(即给定查询)进行扩展,然后使用基于布尔逻辑的检索模型从大规模文档中获取召回率较高的候选文档集合;

(2) 通过使用伪相关反馈方法扩充查询的描述信息、使用CDCR过滤候选文档集合中由同名歧义问题引起的错误结果(不相关文档),优化得到最终检索结果。

实验结果表明,通过使用伪相关反馈检索方法可以提高检索结果的召回率,通过使用CDCR可以提高检索结果的准确率,二者的结合可以获得比现有检索模型更好的性能。

围绕基于共指消解的实体搜索方法研究,本文其余部分组织结构如下:第2节介绍相关工作;第3节详细介绍本文提出的基于共指消解的实体搜索方法;第4节给出对本文方法的实验验证与分析;最后,第5节总结全文并简单阐述未来工作。

1 相关工作

现有SF研究聚焦于属性抽取方法。使用的方法主要包括基于远程监督分类的属性抽取方法,如Roth等[5]和Angeli等[6],基于依存模式匹配的属性抽取方法,如Yu等[7]和Li等[8],以及基于OpenIE属性抽取方式,如Stephen等[9]。

但是,在对实体搜索方法方面研究工作较少[3],现有的检索模型只是简单使用附带查询扩展的基于布尔逻辑的检索模型,差异主要在扩展查询的获取方式。常用的查询扩展方式主要分为两类,利用知识库信息的查询扩展和基于规则的查询扩展。

基于知识库的查询扩展方法主要利用现有的实体知识库(如维基百科)及给定的参考文档信息获取扩展查询。Roth等[5],Xu等[10]以及Yu等[7]通过统计维基页面的重定向关系信息,构建重定向词典,通过重定向词典的查找获得给定查询的候选扩展。Xu等[10]从给定查询的参考文档中搜索查询的缩写形式和全称形式来获取扩展查询。这类方法对知识库有较强的依赖性,对未收录的实体无法进行有效处理。

基于规则的查询扩展方法使用人工编写的规则对查询字符串进行变换,获得扩展查询。

Li等[8]通过去除特殊后缀(如"Ltd."、"Corp."等),利用参考文档中的实体共指信息,以及基于规则从参考文档中抽取别名,获取扩展查询。Min等[11]使用特殊后缀去除、人名改写(如将查询"George Walker Bush"变为"George W. Bush")等方法来获取扩展查询。这类方法集中于对词一级实体名称的变形和抽象,忽略了内容更为丰富的文档一级的特征对实体的刻画能力,获得的扩展查询仅

有助于发现实体名称的多种变形，能提高检索结果召回率，而无法保证结果的准确率。

本文在附带查询扩展的基于布尔逻辑的检索模型基础上，增加了对检索结果的过滤，既能有效保留附带查询扩展的检索模型召回率高的优点，也能有效解决查询歧义性引起的问题，从而获得更好的检索结果。

2　基于共指消解的实体搜索方法

基于共指消解的实体搜索方法以基于布尔逻辑的检索模型为基础，通过运用伪相关反馈方法和跨文档实体共指消解方法，对基本的检索结果进行优化。图 1 给出了该方法的流程图。

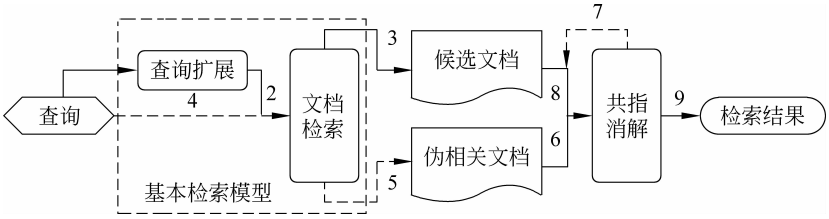


图 1　基于共指消解的实体搜索方法流程图

图 1 中使用不同线型表示模型的两个阶段：①虚线部分表示伪相关文档检索与共指消解，其中 7 表示伪相关反馈；②实线部分表示候选文档检索与共指消解。前者旨在使用伪相关文档扩充查询实体的描述信息，提高最终结果的召回率；后者旨在从大规模数据集中检索获得最终检索结果的候选集合，并通过过滤候选集合中的错误结果，即未包含给定查询实体的文档，以提高最终结果的准确率。接下来先介绍本文使用的跨文档实体共指消解方法，然后详细介绍本文提出的方法的两个处理阶段。

2.1　跨文档实体共指消解

实体共指消解分为单文档共指消解（within document coreference resolution，WDCR）和跨文档共指消解（cross document coreference resolution，CDCR）。WDCR 旨在判断同一篇文档内部的多个实体名称是否指代同一个实体，然后形成多个实体链，每个实体链中的实体名称都指代同一个实体。CDCR 旨在判断不同文档中多个实体描述是否指代同一个实体。为了降低问题复杂度，CDCR 通常在 WDCR 的基础上进行，即判断各个文档中不同的实体链是否指向同一实体。现有研究主要使用聚类算法实现实体共指消解，区别主要在于聚类算法的选择及实体的相似度度量标准。

本文 WDCR 方法使用 Manning 等[12]的方法，使用的 CDCR 方法借鉴 Rao 等[13]，该方法相比于传统的层次聚类算法，具有算法复杂度低的优点，适用于本文这种需要对大量文档进行共指消解的场合。其算法如下：

算法 CDCR-cluster

输入：WDCR 后的文档集合，表示为序列$<D_i, E_i>$，
其中，E_i 为待共指消解的实体在 D_i 中的实体链。

```
1  C←φ
2  For each entity chain <D_i, E_i>:
3    For each cluster c_k in C:
4      sim_k ← Similarity( <D_i, E_i> , c_k )   //①
5    n ← the k maximize the sim_k
6    if sim_n > T:
7      c_n ← c_n ∪ {<D_i, E_i>}   //将文档加入最相似类簇
8    else
9      C ∪ {<D_i, E_i>}          //创建新的类簇
```

输出：C

算法 CDCR-cluster 表示对每篇输入文档计算与所有类簇的相似度（第 4 行），如果最大相似度值大于设定的阈值 T（依据经验，本文设定该阈值为 0.1），该文档加入取得最大相似度值的类簇（第 7 行），否则创建新的类簇（第 9 行）。最终输出聚类结果。其中①处相似度计算如式(1)所示。

$$\begin{aligned}&\text{Similarity}(<D,E>,C)\\&=\alpha\text{Jaccard}(C,E)+(1-\alpha)\cos(C,D)\end{aligned}\quad(1)$$

式(1)中 Jaccard(C,E)表示实体链 E 包含的所有实体名称包含的单词的 unigram 集合与类簇 C 中所有实体名称包含的单词（忽略代词）的 unigram 集合的 Jaccard 系数，计算如式(2)所示。

$$\text{Jaccard}(C,E)=\frac{|\text{unigram}(C)\cap\text{unigram}(E)|}{|\text{unigram}(C)\cup\text{unigram}(E)|}\quad(2)$$

$\cos(C,D)$表示文档 D 的特征向量与文档 C 的特征向量的余弦相似度。文档特征向量采用向量空间模型表示，聚类结果的特征向量使用类簇中所有

文档的特征向量之和表示，也称为该类簇的中心向量。α表示平衡系数，依据经验设定为0.2。因此，式(1)的含义就是：实体名称以及其上下文与类簇中实体名称与上下文共同信息越多，实体与类簇越相似。

2.2 候选相关文档检索

候选相关文档指包含给定查询实体名或其别名的文档，但无法确定此实体名是否指代给定的实体。本文对候选相关文档集进行优化，获得最终的检索结果。首先，候选相关文档集决定了本文最终结果的文档范围，因此要求其具有较高的召回率，即获取候选相关文档集的检索模型需要很好地解决实体别名问题。正如相关工作中所述，现有的在SF中使用的检索模型对此已有研究，且该问题不在本文研究范围之内，本文直接借鉴目前较好的系统的模型。具体的，本文使用萨尔大学口语系统研究所(Spoken Language Systems at Saarland University, LSV)在2013年SF任务中使用的检索模型[①]作为基本检索模型，用以获取候选相关文档。LSV在2012、2013年SF评测中均获得第一的成绩，其检索模型可以代表目前较高水平。该检索模型先获取原始查询的扩展查询，采用的方式如下：

(1) 使用重定向词典。根据维基页面的重定向关系，整理得到页面标题的重定向词典。如对于标题为“Obama(奥巴马)”页面被重定向到页面“Barack Obama(巴拉克·奥巴马)”，这两个页面标题构成的二元组为重定向词典的一条记录；

(2) 人名扩展。对于人名查询，使用人名姓氏作为扩展。例如，将“Jobs(乔布斯)”作为查询“Steve Jobs(斯蒂夫·乔布斯)”的扩展；

(3) 组织名后缀替换。对于组织名查询，通过改变表示组织名查询的后缀获得扩充查询。如对于查询“ABC Corp.(美国广播公司)”可以获得扩展查询“ABC Inc.(美国广播公司)”。

然后，该检索模型利用原始查询和扩展查询进行检索，要求检索到的文档或者包含原始查询或者包含扩展查询。扩展查询使用方式如下：

(1) 选择与原始查询互信息最大的扩展查询对原始查询进行扩展，从大规模文档中进行检索；

(2) 如果检索结果中文档数量小于阈值(设定为500)，则使用所有扩展查询对原始查询进行扩展，重新进行检索。

通过使用上述的检索模型本文获取候选相关文档集，且该检索结果的召回率较高。但是，由于该检索模型没有考虑查询歧义性问题，其准确率较低，需要对结果进行优化，接下来，本文将详细描述如何使用本文提出的方法优化该结果。

2.3 候选相关文档集优化

候选相关文档集虽然较好地解决了实体别名问题，具有较高的召回率，但无法解决实体同名歧义问题，因此其准确率较低。本文使用CDCR方法，对其进行优化。具体的，通过将候选相关文档集进行共指消解聚类，可以将不同实体的相关文档聚成不同类簇，对应给定查询实体类簇中所有文档都是相关文档。

从上文CDCR-cluster算法中可以看出，实体共指消解需要在具体的文档中进行，文档与查询字符串无法进行共指消解。另外，CDCR-cluster算法只能将实体分成不同实体的实体链，并不能给出哪个实体链指代本文所需的实体。此时，KBP为每个查询提供的参考文档，即包含给定查询实体的文档可以发挥作用。具体的，本文对CDCR-cluster算法进行一些修改：初始时将参考文档作为一个类簇，将聚类结果中处于该类的文档都看作相关文档。

然而，这些参考文档包含的相关查询实体的描述信息并不是很充分，有些仅仅只是提到相关查询实体，即使人工阅读也很难总结出有效的用于区分实体的信息。这些参考文档在CDCR时由于与候选文档相同信息较少，过滤过程造成召回率损失较大。为此，本文采用伪相关反馈的方式对查询实体的描述信息进行扩充。伪相关反馈指使用初步检索结果中相关度前几位的文档(称为伪相关文档)对查询进行扩充。相关研究表明，当查询信息不足时，使用伪相关反馈可以获得更好的检索结果[14]。本文借助基本检索模型，不使用其查询扩展功能，仅仅使用文档检索功能，检索结果按照文档与查询的相关度排序。参照伪相关反馈的一般方法，本文从基本检索结果中选取排序靠前的部分文档作为伪相关文档，用以扩充查询描述信息。具体扩充方法为：将伪相关文档和参考文档同时加入CDCR初始类簇中。

但是，伪相关文档中同样会存在同名歧义问题，如果不对此问题进行处理，必然在扩充查询信息时引入较多噪声，影响最终结果的准确率。因此，本文

① https://github.com/beroth/relationfactory

使用参考文档作为初始类簇，先对所有伪相关文档进行共指消解，将聚类结果中保留的文档作为最终使用的伪相关文档。

至此，本文提及多个文档集合，图 2 解释了它们之间的关系：使用参考文档作为初始条件提高伪相关文档集 a 的准确率，并用得到的结果 b 对候选文档集 c 进行优化，得到准确率与召回率兼顾的最终结果 d。

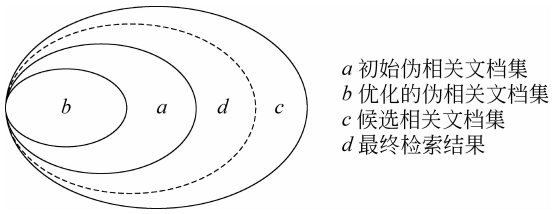


图 2　各个文档集之间的关系

算法 Two-stage-CDCR 描述了本文对候选相关文档的优化过程：

算法 Two-stage-CDCR

输入：候选相关文档集$\{<D_i, E_i>\}$，
　　伪相关文档集$\{<D_j, E_j>\}$，
　　参考文档$<D_r, E_r>$，
　　输入文档均经过 WDCR 处理。

1　$C \leftarrow <D_r, E_r>$
2　For each entity chain $<D_j, E_j>$:
3　　sim $\leftarrow$ Similarity ($<D_j, E_j>$, C)
4　　if $sim > T$:
5　　$C \leftarrow C \cup \{<D_j, E_j>\}$
6　For each entity chain $<D_i, E_i>$:
7　　$sim \leftarrow$ Similarity ($<D_i, E_i>$, C)
8　　if $sim > T$:
9　　$C \leftarrow C \cup \{<D_i, E_i>\}$

输出：C

从算法 Two-stage-CDCR 可以看出，本文对候选结果的优化分为两个阶段：首先，本文通过对伪相关反馈的结果进行共指消解(1～5 行)，从而解决其自身存在的同名歧义问题。然后，本文将伪相关文档共指消解的结果作为候选文档集共指消解的初始条件，对候选文档集进行共指消解(6～9 行)，从而克服参考文档包含的有关查询的信息不充分的问题，保证结果的召回率。

3　实验设置与分析

3.1　实验数据集与评价方法

本文采用 KBP2013 SF 任务评测所用的查询作为输入查询，共包含 100 条查询，其中人名实体查询和组织名实体查询各为 50 条。为了验证本文方法在歧义性较大的查询上的作用，本文又从 KBP2014 SF 任务评测所用的查询中选取存在歧义的 15 个查询[3]作为输入查询。本文采用 KBP2013 SF 任务评测提供的文档集(LDC2014E13)作为相关文档的检索范围。该文档集共包含文档 2 099 319 篇，包含文档类型如下：新闻文本(1 000 257 篇)、网页页面(999 999 篇)和论坛数据(99 063 篇)。

KBP 对所有 SF 输出结果进行人工标注，本文从 KBP2013 标注结果中抽取被标注为正确的文档编号形成 2013 年 100 个查询检索结果的参考答案集，并且对 Yu 等[7]和 Roth 等[4]系统检索模块对此 100 个查询的检索结果进行人工标注，作为参考答案集的补充。对于 2014 年的输入查询，由于缺乏 KBP 的标注数据，使用对 Yu 等[7]和 Roth 等[4]系统检索模块对此 15 个查询的检索结果的人工标注结果作为参考答案集。本文对检索结果采用的评价指标为准确率(Precision，P)，召回率(Recall，R)和 F_1 值(F_1)。

3.2　实验系统设置

为验证本文方法的有效性，本文设置了如下几个系统作为对比系统：

(1) Baseline：该系统采用 LSV 在 2013 年 SF 评测使用的系统的检索模块；

(2) Sys_R：该系统仅将参考文档加入第二阶段聚类的初始类簇，不使用伪相关反馈方法；

(3) Sys_PF：该系统将所有伪相关文档加入第二阶段聚类的初始类簇，省去第一阶段聚类；

(4) Sys_R_PF：完整使用本文的方法，即使用参考文档作为初始条件进行第一阶段聚类，优化伪相关文档集，将此结果作为第二阶段聚类的初始条件，再对候选文档集进行聚类。

LSV 参与了 2012 年与 2013 年的 SF 评测，并均取得了第一名的成绩，本文认为 LSV 系统的检索模型能够在一定程度上代表目前 SF 中检索模型的较优水平，因此将其作为基准系统。设置 Sys_R 的目的在于，通过与 Sys_R_PF 的结果进行对比，可以分析伪相关反馈的作用。设置 Sys_PF 的目的在于，通过与 Sys_R_PF 的结果进行对比，可以分析人工标注的参考文档的作用。

3.3　实验分析

表 2 给出了各个系统的性能数据。从表中数据

可以看出,根据本文提出的基于实体共指的实体搜索方法实现的系统 Sys_R,Sys_PF,Sys_R_PF 相比与根据基于布尔逻辑的检索模型实现的实体搜索系统 Baseline,F_1 值都有一定程度的提升,分别提升 1.79%,2.34%,2.56%。可见,本文提出的方法对现有的实体搜索方法有一定改进。下面本文分别从共指消解的作用、参考文档的作用与伪相关反馈的作用三个方面分析本文方法优于 Baseline 的原因。

表 2 各个系统性能

系统名称	P	R	F_1
Baseline	61.60%	89.55%	72.99%
Sys_R	67.04%	84.55%	74.78%
Sys_PF	66.09%	87.58%	75.33%
Sys_R_PF	67.23%	86.22%	75.55%

3.3.1 共指消解的作用

表 3 给出了对各个系统在候选文档集中文档过滤准确率的统计,过滤准确率即被过滤文档中确为不相关的文档数量占过滤总数的比例。从表 3 中数据可以看出,通过使用实体共指消解方法,对候选文档的过滤准确率基本都能达到 80%左右,即能够排除候选文档中大量的错误,所以能够有效提升结果的准确率,由此,F_1 值得到提升。

表 3 各个系统过滤准确率统计

系统名称	过滤数量	过滤准确率
Sys_R	1 367	79.52%
Sys_PF	1 214	86.41%
Sys_R_PF	1 280	85.00%

3.3.2 参考文档的作用

根据表 2 数据,通过 Baseline 与 Sys_R 的对比可以看出,引入参考文档对候选文档集进行共指消解,可以有效提升检索结果的准确率,虽然召回率受到了损失,但是总体的 F_1 值有所提高。同样,对比 Sys_PF 与 Sys_R_PF 的结果,通过引入参考文档对伪相关反馈结果进行提升,使得最终检索结果的准确率与 F_1 值也能得到有效提升。可见,通过人工标注给出查询的正确相关文档,在提升检索结果的性能上至关重要。而查询一篇相关文档的标注工作量较小,在实际应用中完全可以接受。因此本文提出的方法在给定查询没有参考文档的应用场景下依然具有使用价值。

3.3.3 伪相关反馈的作用

然而,从另一个角度观察表 3 中的数据,由于实体共指的结果并不是 100%的准确,导致对候选结果的过滤中丢弃了一部分相关文档,表 2 的数据反映了相同的问题,即相比于 Baseline,其余各个系统的召回率都有不同程度的损失。本文通过引入伪相关反馈降低该问题的影响。

Sys_R、Sys_R_PF、Sys_PF 在对候选文档集进行聚类时,使用的伪相关文档数量依次增加,其中 Sys_R 不使用伪相关文档,Sys_R_PF 使用部分伪相关文档、Sys_PF 使用了全部伪相关文档,对比表 2 中数据可以看出,随着使用伪相关文档的数量的增加,检索结果的召回率也在提高。从另一个角度看,相比于 Baseline,这些系统的召回率都较低,提高这些系统的召回率可以保证尽量减少候选文档中相关文档的损失。此外,从表 3 数据也能大致看出伪相关文档的作用:仅将参考文档用于共指消解,过滤准确率相对较低,而使用伪相关文档可以有效提升过滤准确率,证明使用伪相关反馈是本文方法优于 Baseline 的一个关键原因。

综合以上分析,本文方法优于 Baseline 的原因主要有三点:

(1) 通过使用实体共指消解过滤不相关文档,提升了检索结果的准确率;

(2) 共指消解中引入参考文档,进一步提升过滤的准确率;

(3) 使用伪相关反馈降低过滤过程中的召回率损失,保证了 F_1 的提升。

表 4 给出在 SF 2014 年的 15 个歧义性较大的查询上的性能。需要说明的是,由于 2014 年查询的参考答案标注范围小于 2013 年(缺少 KBP 的标注数据),因此,表 4 中所有系统的召回率都偏高,但是对于分析系统之间的差异依然具有一定的参考价值。

表 4 歧义查询上各个系统的性能

系统名称	P	R	F_1
Baseline	63.57%	96.78%	76.73%
Sys_R	81.28%	74.76%	77.88%
Sys_PF	71.45%	90.49%	79.85%
Sys_R_PF	79.84%	83.87%	81.80%

从表 4 中数据可以看出,在歧义性较大的查询

上，系统之间的差异更加明显。相比 Baseline，Sys_R_PF 的准确率、F_1 值分别提升 16.27%、5.07%。可见本文方法在歧义性较大的查询上作用更加明显。相比 Sys_R，Sys_R_PF 准确率有所下降，与表 3 中体现的规律不一致，主要原因在于，当查询歧义性较大时，伪相关的准确率也较低，使用伪相关文档会引入一些噪声，这从 Sys_PF 的准确率明显低于 Sys_R、Sys_R_PF 也能得到证明。Sys_R_PF 相比于 Sys_R，召回率提升明显，且 F_1 值提升 3.92%，再次证明伪相关反馈机制的作用。

4　总结与展望

针对目前 SF 任务中使用的检索模型无法处理查询歧义性、检索结果准确率较低的问题，本文提出一种基于跨文档实体共指消解的实体搜索模型，该模型借助伪反馈机制获取的文档对查询的参考文档进行补充，丰富用于共指消解的信息，使用实体共指消解方法对候选结果进行优化，过滤不相关文档，获取比目前常见的仅基于布尔逻辑的检索模型更好的检索性能。实验结果表明，本文提出的方法能够有效过滤不相关文档，进一步提升检索结果的准确率和 F_1 值。

但是，从实验数据可以看出，本文的方法依然还有很大的提升空间，检索结果的准确率依然较低。本文的未来工作围绕以下几方面进行：

（1）改变单一的使用别名对原始查询按照“或”逻辑扩展的查询扩展方式，使用从参考文档中提取的关键词对原始查询按照“与”逻辑扩展，提高候选相关文档的准确率；

（2）改进共指消解聚类所用的相似度计算方法，如融入文档主题信息、使用实体上下文替代整篇文档、区分不同类型词对相似度计算的重要性等，使得相似度值能够更加客观地反映文档间实体共指情况。

参考文献

[1] Dang H T, Surdeanu M. Task description for knowledge-base population at TAC 2013[C]//Proceedings of the 6th Text Analysis Conterence, 2013.

[2] Ji H, Grishman R. Knowledge base population: Successful approaches and challenges[C]//Proceedings of the 49th ACL: Human Language Technologies-Volume 1, 2011: 1148-1158.

[3] Surdeanu M, Ji H. Overview of the English slot filling track at the TAC2014 knowledge base population evaluation[C]//Proceeding of Text Analysis Conference (TAC2014), 2014.

[4] Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. Cambridge: Cambridge university press, 2008: 1-18.

[5] Roth B, Barth T, Wiegand M, et al. Effective slot filling based on shallow distant supervision methods [C]//Proceedings of the Sixth Text Analysis Conference, 2013.

[6] Angeli G, Chaganty A, Chang A, et al. Stanford's 2013 KBP system[C]//Proceedings of the Sixth Text Analysis Conference, 2013.

[7] Yu D, Li H, Cassidy T, et al. RPI-BLENDER TAC-KBP2013 knowledge base population system[C]//Proceedings of the Sixth Text Analysis Conference, 2013.

[8] Li Y, Zhang Y C, Li D Y, et al. PRIS at knowledge base population 2013[C]//Proceedings of the Sixth Text Analysis Conference, 2013.

[9] Soderland S, Gilmer J, Bart R, et al. Open information extraction to KBP relations in 3 hours[C]//Proceedings of the Sixth Text Analysis Conference, 2013.

[10] Xu S, Zhang C X, Niu Z D, et al. BIT's slot-filling method for TAC-KBP 2013[C]//Proceedings of the Sixth Text Analysis Conference, 2013.

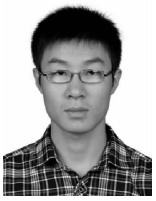
[11] Min B, Li X, Grishman R, et al. New York University 2012 system for KBP slot filling[C]//Proceedings of the Fifth Text Analysis Conference. 2012.

[12] Manning Christopher D, Mihai Surdeanu, John Bauer, et al. The Stanford CoreNLP Natural Language Processing Toolkit[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014: 55-60

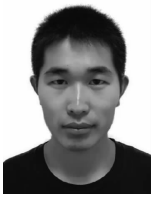
[13] Rao D, McNamee P, Dredze M. Streaming cross document entity coreference resolution[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 1050-1058.

[14] Xu J, Croft W B. Query expansion using local and global document analysis[C]//Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1996: 4-11.

熊玲(1997—),学士,主要研究领域为自然语言处理,信息抽取。
E-mail: lingxiong97@gmail.com



王潇斌(1991—),硕士,主要研究领域为自然语言处理,信息抽取。
E-mail: czwangxiaobin@foxmail.com



徐增壮(1993—),硕士,主要研究领域为自然语言处理,信息抽取。
E-mail: nedxuwork@gmail.com

(上接第 88 页)

[22] Zhang J, Liu B, Tang J, et al. Social influence locality for modelingretweeting behaviors[C]//Proceedings of the International Joint Conference on Artificial Intelligence. AAAI Press, 2013: 2761-2767.

[23] Ma H, King I, Lyu M R. Learning to recommend with social trust ensemble[C]//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2009: 203-210.

[24] Li L F, Fan X Z, Li H Q. Domain-specific QA driven by computation of semantic similarity[J]. Journal of Beijing Institute of Technology, 2005, 25(11): 958-962.

[25] Ru L, Wang Z, Li S, et al. Chinese sentence similarity computing based on frame semantic parsing[J]. Journal of Computer Research & Development, 2013, 50(8): 1728-1736.

高亨德(1990—),硕士,主要研究领域为社会媒体数据挖掘、自然语言处理。
E-mail: 476908322@qq.com

李茹(1963—),教授,博士,博士生导师,主要研究领域为中文信息处理、信息检索。
E-mail: liru@sxu.edu.cn

王智强(1987—),博士研究生,主要研究领域为社会媒体数据挖掘、自然语言处理。
E-mail: zhiq.wang@163.com