第36卷第8期 中国科学技术大学学报 Vol. 36, No. 8
2006年8月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Aug. 2006

文章编号:0253-2778(2006)08-0906-03

面向语义关系的生物文本检索算法*

李 姣,黄民烈,丁石林,余 浩,朱小燕

(清华大学计算机科学与技术系智能技术与系统国家重点实验室,北京 100084)

摘要:面向语义关系的生物文本检索算法通过从生物文本中自动生成满足一定语义关系的模板,将语义关系提取与文本信息检索技术有机融合,以满足用户对生物语义关系查询的需求. 在国际性评测会议 TREC Genomics 提供的标准数据集上的实验结果表明,该算法可以显著地改善生物文本信息检索的性能(平均检索精度提高 15.34%).

关键词:生物文本检索;语义关系提取;融合策略;标准评价

中图分类号:TP311 **文献标识码**:A

Semantic relationship-oriented biological text retrieval

LI Jiao, HUANG Min-lie, DING Shi-lin, YU Hao, ZHU Xiao-yan

(*State Key Lab of Intelligent Technology and System, Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: Based on natural language processing (NLP) technique which can automatically generate patterns from biological texts, relation extraction and text retrieval were fused to enhance the results of relationship-oriented biological text retrieval. Experimental results of TREC (text retrieval conference) genomics track show that the performance of relationship-oriented biological text retrieval can be improved by 15.34% in terms of mean average precision (MAP).

Key words: biological text retrieval; relation extraction; fusion strategy; standard evaluation

研究快报

生物实体间的关系,如基因-基因、基因-疾病、蛋白质-蛋白质等,对整个生物知识网络建立、生物体关系的预测、新药的研制等具有重要意义. 从海量生物文本中检索到包含特定语义关系文本的需求变得非常迫切,成为生物信息领域一个极具挑战性的课题. 但目前广泛应用的生物文本信息检索系统如 PubMed[1]、E-BioSci[2]、Textpresso[3] 等,由于检索模型相似度度量方法等自身的限制,使得系统返回相似度高的文档未能很好地体现生物实体间的语义关系.

为解决上述从生物文本中检索出包含实体间特定语义关系文本的问题,本文提出了面向语义关系的生物文本检索算法,将信息提取研究中模板的概念引入到检索任务中. 本文进行了以下两个方面的研究,如图 1 所示.

(Ⅰ)借助于基于字典的生物命名实体识别技术,从生物文本中识别出包含用户感兴趣的某类特定生物实体对的句子集. 对集合中的句子进行词性标注等自然语言处理后,采用基于序列对齐的思想,从生物文本中自动提取出描述生物实体关系的模

* 收稿日期:2006-07-03;修回日期:2006-07-19
基金项目:国家自然科学基金(60572084,60321002)和中国博士后科学基金(2005038088)资助.
作者简介:李姣,女,1981 年生,博士生. 研究方向:生物文本挖掘. E-mail:jiao-li04@mails.tsinghua.edu.cn
通讯作者:朱小燕,博士/教授. E-mail:zxy-dcs@tsinghua.edu.cn

板.

(Ⅱ)对于用户提交的生物实体语义关系的查询,检索系统返回文档相似度列表和包含用户关心的生物实体的句子集合. 该集合与上步自动生成的模板进行匹配,得到匹配相似度的列表. 将上述两列表融合成一个体现语义关系相关程度的文档列表返回给用户.

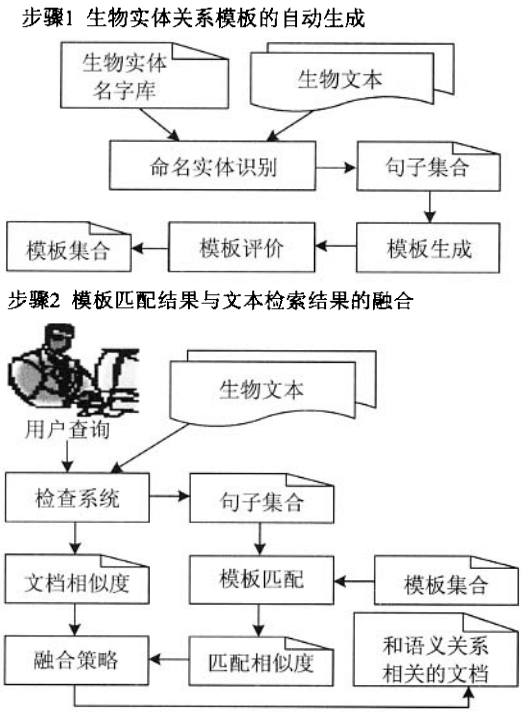

图 1　面向语义关系的生物文本检索过程图

Fig. 1　Semantic relationship-oriented biological text retrieval flowchart

基于序列对齐的模板生成算法的基本过程如下:

(Ⅰ)对从文本信息检索系统中返回的至少包含一个生物实体对(NE_1,NE_2)的句子集合,利用 Brill Tagger[4] 对上述的句子进行词性标注(part of speech,POS).

(Ⅱ)将这些词性标注两两对齐. 为了计算词性标注后句子 X 和 Y 的最佳对齐方式的得分,可采用动态规划的思想[5]. 对齐后的相同部分可作为候选模板.

(Ⅲ)利用模板评价算法[6]对这些模板进行过滤、筛选和优化后,输出模板集 P.

本文采用的模板结构定义为(pattern):

P ::= (pre-filler, NE_1, mid-filler, NE_2, post-filler),

其中,pre-filler、mid-filler、post-filler 分别是 NE_1 前面的、NE_1 和 NE_2 之间的、NE_2 后面的词串. 如"" ("", protein, interact with, protein,"")是描述两个蛋白作用关系的模板. 类似地,句子也被结构化为如下形式:

S ::= (prefix, NE_1, infix, NE_2, suffix),

其中,NE_1 和 NE_2 是句中两个命名实体的语义类别,prefix、infix、suffix 分别是 NE_1 前面的、NE_1 和 NE_2 之间的、NE_2 后面的句子片段. 若一个句子有两个以上的命名实体,则对应不同的命名实体组合,同一个句子有多个结构.

我们计算一个句子 S ::= (prefix, NE_1, infix, NE_2, suffix),和一个模板 P ::= (pre-filler, NE_1, mid-filler, NE_2, post-filler)的匹配得分情况,即计算句子与模板的相似度.

$$\mathrm{match}(\mathrm{sen},\mathrm{pat}) = \sum_{i=1}^{3} \mathrm{sim}(\mathrm{fix}_i^S, \mathrm{filler}_i^P) \quad (1)$$

其中, $\mathrm{fix}_{123}^S = (\mathrm{prefix}, \mathrm{infix}, \mathrm{suffix})$, $\mathrm{filler}_{123}^P = (\mathrm{pre\text{-}filler}, \mathrm{mid\text{-}filler}, \mathrm{post\text{-}filler})$;sim(fix, filler)函数用以衡量句子的 fix 部分与模板结构 filler 部分的相似程度[7].

若文档 D 包含 m 个句子, $D=\{s_1, s_2, \cdots, s_m\}$, 而模板集合为 $P=\{p_1, p_2, \cdots, p_n\}$,则可根据如下公式,对文档是否描述特定的语义关系进行排序,即定义文档相似度.

$$\mathrm{sim}(D) = \max_{s_i \in D, p_j \in P} \{\mathrm{match}(s_i, p_j)\} \quad (2)$$

查询(query)与文档(document)之间相似度的计算是该 document 与 query 中每个 term 的相似度之和.

$$\mathrm{sim}(D,Q) = \sum_{t \in Q} \lambda_t w_t \quad (3)$$

其中,D 为被检索的文档;Q 为用户的查询; w_t 为 Q 中的一个查询词 t 与文档 D 的相似度(使用了 BM2500 概率模型[8]);这里 $\lambda_t=1$.

式(2)、(3)在相同文档集合上和在不同粒度、不同精度上,对文档与用户查询的相似度加以判定. 前者的查询粒度更细,精度更高;后者返回了尽可能多的相关文档. 本文提出基于基础增强的合成方法(如算法 1),通过将二者融合来提高系统的总体性能. 其中,我们用排序倒数来代替相似度的取值,将上述两个不同的打分机制归整到一个框架下.

$$\hat{s}_i = 1/r_i \quad (4)$$

其中,r_i 是结果列表中文档 d_i 的排序,$\hat{s}_i$ 是归一化

后相似度的取值.

算法 1 基于基础增强的结果合并算法

For each $d_i \in L_1$ {
  if $d_i \in L_2$
    $s_i = \lambda \hat{s}_{i1}$;
  else $s_i = \hat{s}_{i1}$;
}

算法中,L_1是文本信息检索的结果(基础表);L_2是生物实体关系提取结果(增强表);文档 d_i在 L_1和 L_2中经过排序倒数的相似度分别为 $\hat{s}_{i1}$和 $\hat{s}_{i2}$;经过合并后的文档 d_i的相似度为 s_i. 由此可见,基于基础增强的结果合并算法,实际上是根据可信度较高的结果列表(增强表)对信息更充分的结果列表(基础表)进行重新排序(re-ranking)的过程.

本文利用 TREC2005 Genomics Track 提供的公共数据集和评测标准对我们的方法加以验证[9]. 实验中,我们选用了一组关于基因和疾病的查询请求,其具体描述如下:

检索出描述基因 G 和疾病 I 关系的文档. 其中,G 和 I 为给定的 10 对基因和疾病名称,例如:G: DRD4;I: alcoholism(酒精中毒).

我们的实验目的是通过融合生物实体关系提取结果(L_2)改善文本信息检索的结果(L_1)的性能. 将 L_1作为本实验的基准线(baseline),以平均检索精度(mean average precision,MAP)作为面向语义关系的生物文本检索性能的综合评价指标.

在基于基础增强的融合方法中,通过调节增强系数的大小,提升 L_1中体现语义相关性的文档的排序,其平均检索精度随增强系数的变化情况如图2所示(算法中的 λ 是横轴的指数函数). 由图 2 可见,该方法极大提高了平均检索精度(+15.34%). 权值在 (6, 10)区间上的性能已经趋于稳定,主要因为权值足够大将 L_2中的所有文档以其在 L_1中的顺序排在返回结果的最前面.

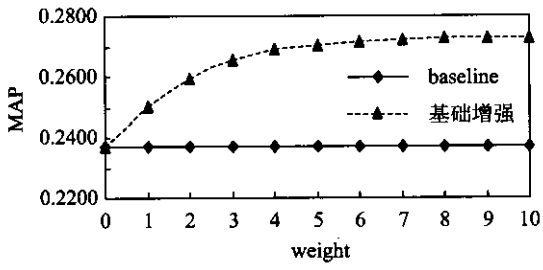

图 2 基于基础增强的融合方法的平均精度性能

Fig. 2 MAP performance of base-enhancing fusion method

本实验说明了生物实体关系提取和生物文本信息检索的融合为面向语义关系生物文本信息检索提供了一种行之有效的方法. 后续研究将围绕无监督的生物实体关系模板自动生成算法和模板匹配结果与文本检索结果融合算法展开.

参考文献(References)

[1] PubMed (access 2006)[DB]. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed.

[2] Grivell L. E-BioSci: semantic networks of biological information[J]. Information Services and Use 2003, 23(2-3):179-182.

[3] Muller H M, Kenny E E, Sternberg P W. Textpresso: an ontology-based information retrieval and extraction system for biological literature[J]. PLoS Biology, 2004, 2(11):1984-1998.

[4] Brill E. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging[J]. Computational Linguistics, 1995, 21(4):543-565.

[5] HUANG M L, ZHU X Y, HAO Y, et al. Discovering patterns to extract protein-protein interactions from full texts[J]. Bioinformatics, 2004, 20(18):3 604-3 612.

[6] HAO Yu, ZHU Xiao-yan, HUANG Min-lie, et al. Discovering patterns to extract protein-protein interactions from the literature: part II [J]. Bioinformatics, 2005, 21(15):3 294-3 300.

[7] HUANG M. Semantic relation discovery and its application on bioscience text processing [D]. Department of Computer Science and Technology of Tsinghua University, Beijing. 2006:135.

[8] Ricardo B Y, Berthier R N. Modern Information Retrieval[M]. New York: Addison Wesley, 1999.

[9] Hersh W, Cohen A, Yang J, et al. TREC 2005 genomics track overview[C]// Proceedings of 14th Text Retrieval Conference. Gaithersburg, 2005:14-23.

研究快报