硕 士 学 位 论 文

基于评论挖掘的药物副作用发现

The Discovery of Adverse Drug Reactions Based on Comment Mining

作 者 姓 名：程亮喜

学科、 专业：生物医学工程

学　　号：21109275

指 导 教 师：林鸿飞 教授

完 成 日 期：2014.6

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年____月____日

摘　要

随着药物副作用带来的危害越来越大，药物安全问题日益受到人们的重视并逐渐成为医学界和民众关注的热点，因此如何发现药物的副作用具有重大的理论与实用价值。而 Web 2.0 技术的发展使得互联网上出现了不少医疗健康类社交网站，人们在上面分享用药经历并对药物进行评论。这些网站上的用户评论数据日益丰富，其中蕴含的药物副作用相关信息开始受到研究人员的关注，并逐渐形成从用户评论中挖掘副作用信息这样一种快捷、有效的药物副作用发现机制。

在从用户评论中挖掘药物副作用时，由于人们可能采用不同的表述方式来描述副作用，而新药的上市与用药者的差异性会造成新的副作用出现，因此从评论中识别新的副作用名称并进行标准化十分重要。针对该问题，本文第 3 章工作利用条件随机场模型识别评论中的副作用，对识别出的副作用名称进行标准化，最后获取药物的副作用。实验结果显示，条件随机场模型可以识别出已知的与新的副作用名称，而标准化技术将副作用名称进行聚合与归并，有利于药物副作用的发现。本文通过将挖掘出的药物已知的副作用与数据库记录进行对比验证了本文方法的有效性，同时得到一个按评论中的发生频率排序的药物潜在副作用列表。

从用户评论中识别副作用名称是药物副作用发现中基础却关键的步骤，但由于评论内容在语法上的不规范性与副作用名称的多样性，从评论中识别副作用实体具有较大的挑战性。针对该问题，本文第 4 章实现了一个融合不同方法的副作用实体识别系统。第一种方法将滑动窗口中的短语与词典中的名称进行词袋匹配识别副作用实体，并在匹配时考虑了编辑距离；第二种方法利用条件随机场模型进行识别，其中应用了向前选择法找出最佳的特征集合，并通过试验找出效果最好的词语上下文特征组合方式。将两种方法的识别结果进行融合，得到的融合后结果比单一方法具有较大提升，说明通过融合可以弥补单一方法识别的不足。与其他文献中的副作用实体识别方法相比，本文方法的识别性能与之相当甚至可能优于他们，从而证明本文提出的融合方法的有效性。

关键词：药物副作用发现；文本挖掘；命名实体识别；条件随机场；实体标准化

The Discovery of Adverse Drug Reactions Based on Comment Mining

Abstract

As the harm of adverse drug reactions (ADRs) grows, drug safety issues draws more people's attention and is becoming the focus of the medical professionals and the public, so how to discover the ADRs is of great theoretical and practical value. With the development of Web 2.0 technologies, many healthcare social networking sites appear on the Internet and people share medication experiences and give comments on drugs there. As the comments on those websites become increasingly rich, researchers begin to pay attention to the ADR information within the user comments, and gradually develop a quick and effective ADR discovery mechanism through mining the comments.

When mining adverse drug reactions (ADRs) from the comments, it is very important to recognize novel ADR expressions and normalize them, since people probably adopt different expressions to describe adverse reactions and new adverse reactions may emerge with the listing of new drugs as well as the diversity of drug users. For this case, the work in Chapter 3 utilized the conditional random field (CRF) model to recognize adverse reaction entities, and proposed a normalization method for them. Experimental results indicated that the CRF was able to identify both known and novel adverse reaction entities, and the normalization merged those entities, which benefitted the ADR discovery. The similarity between mined results of known ADRs and database records verified the effectiveness of this mining method, and a list of potential ADRs sorted by occurrence frequency in comments was obtained finally.

The recognition of adverse reaction entities from user comments is a basic but crucial step in the discovery of ADRs. Because of the grammatical irregularity of the comments and the diversity of the adverse reaction expressions, it is rather challenging to recognize adverse reaction entities from the comments. To solve this problem, the work in Chapter 4 implemented an entity recognition framework which integrated the results of different recognition methods. The first method recognized entities by matching the bag of words within a sliding window with the word bag of lexical terms, where the edit distances of words were considered; the second one adopted the CRF model to recognize entities, which applied feature selection to find the best internal feature set, and identified the most effective feature combination through repeatedly trials. The recognition results of the two methods were integrated and the integrated performance is greatly improved compared with either result of the two recognition methods, indicating that the integration can compensate the deficiencies of a single recognition method. Compared with other adverse reaction entity recognition

methods, the performance of this method is comparable or maybe even better than theirs, demonstrating the effectiveness of the proposed method.

Key Words：Adverse Drug Reaction Discovery; Text Mining; Named Entity Recognition; Conditional Random Field; Entity Normalization

目　　录

1 绪论

1.1 研究背景

随着人们生活水平的提高与科学技术的快速发展，各个国家对生物医学领域的研究越来越重视，在这方面的投入逐渐加大，因此与生物医学相关的文献记录呈现飞速、爆炸式的增长。例如，生物医学领域权威的文献数据库 MEDLINE 已收录了来自世界约 5,600 个相关期刊的超过 2,100 万篇文献，并且每个工作日都会新增 2,000-4,000 篇的文献[1]。面对如此海量的数据，人们单凭人工从中发现有价值信息显得有些力不从心，这就迫切需要自动分析与提取药物信息的技术。而随着计算机技术的发展与普及，生物医学领域的研究人员开始结合应用数学、统计学、信息学和计算机科学的方法研究生物医学问题，从而诞生了生物信息学这一门新的交叉学科并使其在处理与利用生物医学海量数据方面大显身手，这其中就包括在药物研发方面的应用。

众所周知，新药的研发周期漫长，花费巨大，其中对药物安全性的评估是一个重要的环节。即使药物上市后，相关的监管部门也会持续跟进它可能产生的不良反应，降低它给人们带来的危害。目前制药工业正日益成为以知识为基础的行业，研究人员在研发药物的过程中需要从已有的知识库中获取与药物相关的各种信息。在面对拥有海量数据的药物相关的知识库时，人们开始借助生物信息学的技术与手段来研究与解决药物的相关问题，其中包括对药物安全性的评估。在 Campillos 等人[2] 发表于《科学》（Science）上的论文中，为了推断两种药物是否具有相同的靶标蛋白，他们没有利用药物的分子结构或在细胞中的活动状态等常规的方法，而是根据药物副作用的相似性进行推断，并通过生物医学实验检测推断结果的正确性，证实了利用表型信息推测上市药物的分子作用关系以及发现新功效的可行性。在 Lounkine 等人[3] 发表于《自然》（Nature）网络期刊上的论文中，他们利用计算机模型而非药物实验的手段对 656 种上市药物在 73 个靶标蛋白上的副作用进行了预测，其中约有一半的预测被其他来源的数据证实是正确的。

药物安全性评估是为了检测、发现药物使用过程中可能产生的副作用，从而规避潜在的风险。药物副作用（Drug Side Effect，或称药物不良反应，Adverse Drug Reaction, ADR）一般是指在正常条件下由药物引起的非预期的、有害的反应[4]。药物副作用涵盖的范围很广，可以是用药后的过敏反应或耐药性的改变，也可以是药物成瘾或因用药导致原有病情的加重等。由药物的副作用引起的病患占所有医院病患的 5%，药物副作用已成为导致医院死亡的第五大原因[5]。据估计，美国每年用于治疗药物副作用引发疾病的费用高达 1360 亿美元，而其他国家也面临着与此类似的情况[6]。用药安全问题日益受

到人们的重视，药物副作用逐渐成为医学界和民众关注的热点，因此如何判断和预测药物的副作用具有重大的理论和实用价值。

药物副作用的发现通常有两种机制，一是新药上市前的临床试验；二是上市后通过疾病流行和控制中心采集副作用的报告，这些机构通过设置采集信息平台，获取来自医院、制药公司、患者等的报告。例如美国的食品药品监督管理局（Food and Drug Administration, FDA）负责监督管理上市药物的安全问题，所有与药物不良反应相关的信息都会反馈到 FDA 的不良事件报告系统（Adverse Event Reporting System, AERS）。然而仅仅通过上述两种途径来收集与发现药物的不良反应信息是不够的。对于新药上市前的临床试验，由于药物作用机理的复杂性，很难做到检测的全面性，不可能发现药物所有的不良反应。通过官方机构的信息采集平台难以收集全面的药物不良反应信息，而且它确定药物副作用的时间周期一般较长，发现速度比较慢。

随着 Web 2.0 技术的发展，互联网上出现了社区、论坛、博客、微博、Wiki 等各种形式的用户生成内容（User-Generated Content，UGC），它们极大地丰富了网络，并扮演着越来越重要的角色，这其中就包括用户对物品的评论。近年来互联网上出现了一大批的医疗健康类的社交网站，如 DailyStrength、PatientsLikeMe、Health & Wellness Yahoo! Groups 等，这些网站积聚了众多的用户，他们在这里发表、分享自己或身边亲人、朋友的用药体验与评论，形成了关于用药反应的十分有价值的来自用户的第一手资料。调查显示，在 2012 年里美国 35% 的成人试图通过互联网对药物的相关情况进行判断，增进对感兴趣药物的了解[7]，这表明网络正逐渐成为人们交流与获取药物相关信息的一种重要方式。

这些医疗健康类的社交网站汇集了大量来自用户的用药体验与评论，其中蕴含的最新副作用信息越来越受到研究人员的重视，并逐渐形成从用户评论中挖掘药物副作用的研究方向。从用户评论中挖掘药物副作用信息是一种快捷、有效的渠道，其相关理论和方法的发展将对药物副作用的发现产生重要的推动作用。基于社交网络中用户评论挖掘的药物副作用发现和预测更能确切反映民众的需求，服务于民众的健康生活，获取的最新药物副作用信息对于患者的用药安全以及制药公司均具有重要的现实意义。

1.2 研究现状

2010 年 Leaman 等人[6] 以 DailyStrength 上用户的用药评论作为语料，通过计算滑动窗口中的评论内容与词典中副作用名称之间的相似度进行实体识别，对识别出的副作

用名称过滤后挖掘药物的副作用，在人工标注的数据集上识别的准确率为 78.3%，召回率为 69.9%，F 值为 73.9%。

2011 年 Chee 等人[8] 利用 Health & Wellness Yahoo! Groups 上用户对药物的评论信息评估药物的安全性。由于正负例数量不均衡，他们采用 Bootstrapping 方法增加正例，并融合多个分类器对药物进行分类，预测可能将被监管或召回的药物。

2011 年 Yates 等人[9] 从 AskaPatient, Drugs.com 和 DrugRatingz.com 上抓取关于 5 种乳腺癌药物的评论，标注其中部分评论作为训练集与测试集，并利用 UMLS（Unified Medical Language System）数据与训练集构造了一个包含变体的副作用同义词集。作者人工制定模板，结合词典匹配与模板匹配从评论中识别药物副作用，然后依据 SIDER 2 中的数据判别已知与潜在的副作用，并根据支持度与强度对潜在副作用进行过滤。在标注数据集上进行实验取得了良好的结果，证明了本文方法的有效性。

2011 年 Nikfarjam 等人[10] 利用一组语言模式从 DailyStrength 上用户对药物的评论中自动抽取副作用。他们采用关联规则从已标注的评论中挖掘副作用口语化表述的潜在模式，在测试集上对这些模式评估得到的 F 值为 67.96%。

2012 年 Yang 等人[5] 利用用户健康词汇表（Consumer Health Vocabulary, CHV）[11] 构建了一个扩展的副作用名称词典，采用滑动窗口从 MedHelp 上关于药物的帖子中识别副作用名称，并使用关联规则挖掘药物的副作用，对于 5 种指定药物的实验得到较好的效果。

2012 年 Sampathkumar 等人[12] 从 Medications.com 论坛上收集了 7,961 条文本信息，对其中的 100 条文本进行人工标注，剩余的采用字典匹配自动标注。他们利用隐马尔可夫模型对自动标注的文本进行副作用实体识别，10 折交叉验证得到的平均 F 值为 86.4%，在人工标注的文本上进行识别得到的 F 值为 73.2%。

2013 年 Yates 等人[13] 利用滑动窗口初步确定 Breastcancer.org 和 FORCE 论坛帖子中可能的副作用，并采用 CRF 分类器进一步确定是否为副作用，其在分类时利用了句子的依存关系；对识别出的副作用采用 CRF 为其赋予药名，找到对应的药物。在已标注语料上实验，与基准方法相比召回率有损失但准确率取得了较大提高。

2013 年 Yang 等人[14] 利用文本挖掘技术从社交媒体网络的文本中自动提取报告药物不良反应的信息，并采用部分监督学习的方法将其分为 ADR 相关（正例）与 ADR 不相关（负例）两种类型，从而通过社交媒体分析对上市药物进行监测，另外这种方法也同样适用于其他具有大量可用的用户生成内容的领域。

2013 年 Xu 等人[15] 提出了一种方法来从文本中抽取抗癌药物-副作用关系对。他们以抗癌药物词典和药物副作用词典作为辅助，从 MEDLINE 文本语料库中抽取药物-副

作用的共现对，然后滤除其中的药物-适应症对，最后对经过过滤的关系按可信程度进行排序，实验结果显示排名靠前的关系对达到了 77.8%的准确率。

2013 年 Wu 等人[16] 整合了网络上多个讨论药物副作用的文本数据源，得到一个更加全面的数据集。为了从这些文本数据中发现未确认的药物副作用，作者提出生成式和判别式两种方法来对文本中的药物副作用进行挖掘。实验结果表明网络中关于药物副作用的讨论内容可用于未知副作用的监测，生成式模型方法在准确率与召回率两方面均比判别式方法更有效。

1.3 本文工作

由于语言表述的自由性与多样性，人们在表达同一个副作用概念时可能会采用不同的措辞方式，而新药的上市以及用药者的差异性又可能会导致新的副作用出现，故而用户评论中会存在数据库未收录的副作用名称，有的甚至是因拼写错误而造成的不同。因此从评论中挖掘药物的副作用时，识别新的副作用名称并将其映射到统一的副作用概念上是十分重要的，否则将无法发现一些潜在的副作用，或挖掘出的副作用发病率与事实存在偏差。针对这些情况，本文采用条件随机场模型识别用户评论中的药物副作用，可以有效识别出已知的以及新的副作用名称；对于识别出的新的副作用名称，利用本文提出的标准化方法对其进行标准化，将它们映射到已知的副作用概念上，从而对副作用名称有效的聚合与归并，有利于药物副作用的发现。

从用户评论中识别副作用名称是药物副作用发现中基础却关键的步骤，副作用实体的识别效果对药物副作用的发现有直接而重要的影响。由于评论内容语法上的不规范性与副作用名称的多样性，从评论中识别副作用实体具有较大的挑战性。以前工作在对副作用实体进行识别时，得到的识别性能还有较大的提升空间。针对评论中副作用实体的识别问题，本文实现了一个融合不同方法识别结果的副作用实体识别系统。第一种方法中，我们将滑动窗口中的短语与副作用词典中的实体进行词袋匹配，根据匹配结果识别副作用实体；第二种方法中，我们利用条件随机场模型识别副作用实体。在识别过程中，我们通过特征选择找出最佳的特征集合，并对词语上下文特征范围与特征组合方式进行多次尝试，找出效果最好的上下文范围与特征组合方式。最后我们将这两种方法的识别结果进行有效地融合，弥补单一方法在识别中的不足，从而提高副作用实体识别的性能。

1.4 本文结构

本文共分为 4 章，主要阐述了从医疗健康网站上的用户评论中识别副作用实体，对识别出的实体进行处理从而发现药物副作用的相关工作，具体章节内容安排如下：

第 1 章是绪论部分，介绍了从用户评论中挖掘药物副作用的背景和研究现状，分析了该领域目前面临的问题，说明了本文的研究工作和章节安排。

第 2 章阐述了本文用到的文本挖掘相关的技术，包括命名实体识别、条件随机场等，并介绍了本文实验用到的一些外部资源。

第 3 章详细介绍了结合标准化技术进行药物副作用挖掘的方法。首先利用 CRF 模型从文本中识别副作用实体，然后利用标准化技术将识别出的副作用名称映射到标准概念上，并在标注数据集上进行实验验证了方法的有效性。

第 4 章利用两种不同的方法从用户评论中识别副作用实体，并对它们的结果进行有效地融合，从而弥补单一方法在识别实体时的不足，提高副作用实体识别的效果。

结论部分对本文的研究内容与工作进行了总结。

2 相关技术与资源

2.1 命名实体识别

2.1.1 简介

命名实体识别（Named Entity Recognition，NER）是指从非结构化的文本中识别出人名、地名、机构名等实体名称，日期、时间等时间表达式，以及币值、百分比等数字表达式，并对其加以归类[17]，使之成为便于使用的结构化数据。在生物医学文本领域，命名实体识别的对象主要是生物体内相关物质的名称（基因、蛋白质、细胞等）以及疾病、药物、药物副作用等。命名实体识别是一项基础性工作，常见的句法分析、机器翻译、问答系统等的性能都依赖于命名实体识别的效果，因此在自然语言处理（Natural Language Processing，NLP）技术走向实用化的过程中，命名实体识别占据着举足轻重的地位，其识别的性能对后续工作有着直接而重要的影响。

2.1.2 识别方法分类

自从 1991 年 Rau [18] 在第 7 届 IEEE 人工智能应用会议（Seventh IEEE Conference on Artificial Intelligence Applications）提出一个抽取与识别公司名的系统以来，命名实体识别的研究经历了 20 多年的发展，人们提出了各种各样的识别方法。识别实体的方法虽然数量繁多，但基本都可归结为三类，即基于规则的方法、基于统计的方法（或称基于机器学习的方法）以及基于融合的方法。

1. 基于规则的方法

基于规则的方法是命名实体识别研究中最早被提出并应用的，它以字符串和模式匹配为主要手段，依靠命名实体词典和规则来识别名称。基于规则的方法一般把一些已有的实体收入词典作为基础，对于词典中没有的实体，则通过规则进一步识别。为了解决某些情况下的规则冲突问题，这类方法通常会对每个规则赋予权值，以便在规则冲突时选取最高权值对应的规则来识别实体并对其类型进行判别。

基于规则方法中的规则通常与实体的类型、文本的语言以及风格等相关联，因此基于这类方法的系统可移植性一般不好，应用于不同的系统时需要重新制定相应的规则。制定规则的过程耗时耗力，容易产生错误，而且很难涵盖所有的语言现象，另外通常需要建立不同领域知识库作为辅助以提高系统的识别性能，因此这类系统的建设周期一般都较长，开发代价也很大。基于规则的方法通过分析命名实体的内部与外部特征，人工构造规则模板实现命名实体的识别，其选用的特征一般包括统计信息、关键字、指示词、

位置词、标点符号等。通常来说，当方法中使用的规则反映所识别文本的语言现象比较精确且规则覆盖比较全面时，基于规则的方法在性能上并不会比基于统计的方法差，甚至会优于后者。

自从命名实体识别任务在第六届消息理解会议（Message Understanding Conference, MUC）中被提出后，关于命名实体识别的研究开始兴起并逐渐增多。在该研究的初始阶段，大多数系统都采用基于规则的方法实现，这类方法一度占据了主导地位。在参加第六、七届消息理解会议命名实体识别评测任务的系统中，采用基于规则方法的占了绝大多数，其中包括 Black 等人[19] 的 FACILE 系统、Grishman[20] 的 Proteus 系统、Mikheev 等人[21] 的 LTG 系统、Krupka 等人[22] 的 NetOwl 系统等。此外，Farmakiotou 等人[23] 提出一个基于手工建立的词典资源的命名实体识别系统，在希腊语的财政新闻语料库上测试取得了理想的结果。Hanisch 等人[24] 实现了一个基于规则的基因与蛋白质实体识别系统，该系统使用经过预处理的同义词词典来识别生物医学文本中潜在的名称，并将其与数据库中蛋白质、基因的标识符关联。

2. 基于统计的方法

基于统计的方法也称为基于机器学习的方法，这是当下用得较多的一类方法，也是目前研究的主流方法。基于统计的方法在训练语料上训练机器学习的模型，然后将得到的模型用于实体识别中。这类方法的语料标注一般无需专业的语言学和计算语言学知识，但对语料库规模的依赖性较大，通常需要较大规模的标注语料来进行模型训练。基于统计的方法识别命名实体依据的是训练语料上的统计规律，因此它的客观性与可移值性都比较好，在不做或只做小的改动的情况便可移植到新的领域。

基于统计的方法的一个难点是用于学习的特征的选取。为了使机器学习算法得到良好的识别效果，我们需要从文本中选择能够准确反映实体特性的各种特征，剔除可能产生噪音的特征。选取特征时一般通过对训练语料所包含的语言信息进行统计和分析，优先选择贡献度大的特征，挖掘出与实体最相关的特征，此外组合特征也可以提升系统的性能。命名实体识别的常用特征一般包括词语特征（词语长度、拼写、词性等）、词典特征（如在特定词典是否出现以及出现的频率等）、核心词特征、语义特征、句法特征等。另外，当前词语的上下文特征以及相应的组合特征对于提高识别性能也非常有效。

在命名实体识别领域，目前常用的机器学习算法包括：

1）支持向量机（Support Vector Machine，SVM） Kazama 等人[25] 利用 SVM 进行生物医学命名实体的识别，与一个基于最大熵的系统相比，他们识别实体的性能更优越。Isozaki 等人[26] 设计了一个基于 SVM 的命名实体识别系统并提出一种改进方法提高系统的运行效率，实验结果显示其识别性能优于传统的识别系统。Ju 等人[27] 考虑到

SVM 对数据分析与模式识别的有效性与高效性，利用了 SVM 模型来识别生物医学命名实体。

2）隐马尔可夫模型（Hidden Markov Model，HMM） Zhou 等人[28] 提出一个基于 HMM 的组块标注器，并在此基础上构建了一个用于名称、时间以及数字表达式识别与分类的命名实体识别系统。Zhao[29] 在命名实体识别中引入 HMM，并利用词语相似度对其进行平滑，实验显示在拥有大规模未标注语料可用时，基于词语相似度的平滑可以提高性能。

3）条件随机场（Conditional Random Field，CRF） Settles 等人[30] 提出了一种利用 CRF 识别生物医学文献摘要中多种类别实体的方法，实验显示基于 CRF 的模型在只有简单正交特征时就能够获得当前最优系统所具有的性能，而使用语义词典并不能进一步提供性能。Leaman 等人[31] 创建了一个基于 CRF 模型的生物医学命名实体识别系统 BANNER，该系统不使用语义特征或基于规则的处理步骤而使它具有领域独立性，实验显示其性能比已有的基准系统优越很多。McCallum 等人[32] 利用 CRF 模型识别命名实体，识别过程中使用自动特征归纳，并利用 Web 数据对对词典进行扩充，方法在 CoNLL-2003 的 NER 共享任务中取得了良好的结果。

4）最大熵模型（Maximum Entropy，ME） Borthwick 等人[33] 设计了一个基于最大熵理论的命名实体识别系统，该系统采用基于对象的体系结构，能够有效利用各种不同来源的知识来辅助其进行标注。Chieu 等人[34] 提出了一个基于 ME 的命名实体识别系统，该系统利用整篇文档的全局信息且只应用一个分类器对单词进行分类，取得了与其他基于机器学习的最好系统相当的性能。Bender 等人[35] 将 ME 模型应用到命名实体识别任务中，并通过实验证明基于标注训练集的基准系统的性能可以通过添加额外的非标注数据得到提升。

3. 基于融合的方法

总的来说，基于规则的方法依赖于建立的词典与规则，而基于统计的方法则对训练语料以及特征的选取要求较高，这两类方法有着各自的优势，但同时也存在着不足。在单独使用一种方法的识别效果不太理想时，我们可以考虑将两种或多种方法进行有效地融合，使不同方法之间的优势得到互补，克服单一方法的不足之处，从而提高识别的性能。

融合可以在基于规则与基于统计的方法之间进行。Isozaki[36] 提出了一种基于简单的规则生成器与决策树（Decision Tree, DT）学习日语识别系统，实验结果显示该系统与一个基于最大熵最优系统的性能相当，另外该系统在大规模语料上的训练速度更快。

Lin 等人[37] 提出了一种混合的方法用于命名实体识别，该方法以最大熵模型作为底层的机器学习算法，并在后处理中引入基于词典与规则的方法。

融合也可以在多种基于统计的方法之间进行。Florian 等人[38] 将鲁棒线性分类器（Robust Linear Classifier）、最大熵、基于变换的学习（Transformation-based Learning）以及隐马尔可夫模型 4 个分类器在不同状态下进行组合，从而得到一种基于分类器组合的命名实体识别方法。Liu 等人[39] 提出一个结合 K 近邻（K-Nearest Neighbors, KNN）分类器与线性条件随机场模型的半监督学习框架用于识别推特（Twitter）微博中的命名实体，实验结果显示了该方法相对于基准方法的优越性以及 KNN 与半监督学习的有效性。

2.1.3 评价方式

命名实体的识别包括实体边界与实体类别的确定两个方面，只有实体边界与类别均判定无误才算识别正确。英文分词比较简单，其命名实体的形态标志较为明显（如实体中单词的首字母大写），因此英文命名实体边界的确定相对容易些，实体类型的确定是其任务的重点；中文的分词具有一定的难度，而且命名实体不存在明显的形态标志，因此中文命名实体识别相对来说更加复杂与困难。

评价命名实体识别系统性能优劣的指标主要包括三个：准确率、召回率以及 F 值。准确率（Precision，P）是识别出的所有实体中正确实体所占的比例，衡量的是系统的查准率；召回率（Recall，R）是识别出的正确实体占所有正确实体的比例，衡量的是系统的查全率；F 值（F-score，F）是准确率与召回率的加权几何平均值，衡量的是系统的综合性能。三个指标的计算公式如下所示：

$$P = \frac{TP}{TP + FP} \tag{2.1}$$

$$R = \frac{TP}{TP + FN} \tag{2.2}$$

$$F = \frac{(1+\beta^2)\cdot P \cdot R}{\beta^2 \cdot P + R} \tag{2.3}$$

其中 TP 是识别出的正确实体数量，FP 是识别出的错误实体数量，FN 是未识别出的正确实体数量。参数 β 是一个非负实数，用于调节准确率与召回率的相对重要程度：当 β 小于 1 时，结果更注重准确率；当 β 大于 1 时，结果更注重召回率。本文在评价命名实体识别系统的性能时，β 取作 1，即认为准确率与召回率同等重要。

2.2 条件随机场

2.2.1 简介

条件随机场是 Lafferty 等人[40] 在 2001 年提出的一种基于统计的判别模型，其特点是假设输出变量之间的联合概率分布构成概率无向图模型（Probabilistic Undirected Graph Model, UGM）即马尔可夫随机场（Markov Random Field, MRF），这是一种用来标记和切分序列化数据的统计框架模型。CRF 没有隐马尔可夫模型的严格独立性假设条件，因而可以容纳丰富的上下文信息；同时由于 CRF 计算全局最优输出节点的条件概率，因此它克服了最大熵马尔可夫模型（Maximum Entropy Markov Model, MEMM）的长距离依赖和标签偏置的缺点[40]。条件随机场模型中的一种特殊类型是线性链条件随机场（Linear Chain Conditional Random Field），它是一种对数线性模型（Log Linear Model），其重要应用就是序列标注，经常被用于解决词性标注、中文等语言的分词、命名实体识别等问题。

2.2.2 模型描述

在给定随机变量 X 的条件下，随机变量 Y 的马尔可夫随机场称为条件随机场。本文中用于副作用实体识别即标注问题的线性链条件随机场是一种定义在线性链上的特殊的条件随机场，此时在条件概率模型 $P(Y \mid X)$中，输入变量 X 表示需要标注的观测序列，输出变量 Y 表示标记序列（或称状态序列）。在学习时，条件概率模型 $\hat{P}(Y \mid X)$ 是算法利用训练数据集通过极大似然估计或正则化的极大似然估计得到的；在预测时，对于输入序列 x，输出序列 $\hat{y}$ 是最大的条件概率 $\hat{P}(y \mid x)$ 对应的标记序列。

对于随机变量 X 与 Y，设 $P(Y \mid X)$ 是在给定 X 条件下 Y 的概率分布。如果 Y 构成由无向图 $G = (V, E)$ 表示的马尔可夫随机场，即

$$P(Y_v \mid X, Y_w, w \neq v) = P(Y_v \mid X, Y_w, w \sim v) \tag{2.4}$$

对于任意结点 v 成立，则称条件概率分布 $P(Y \mid X)$ 为条件随机场。式中 $w \sim v$ 表示在图 $G = (V, E)$ 中与结点 v 有边连接的所有结点 w，$w \neq v$ 表示结点 v 以外的所有结点 w，Y_v、Y_u 与 Y_w 为结点 v、u 与 w 对应的随机变量。

设随机变量序列 $X = (X_1, X_2, \cdots, X_n)$ 、$Y = (Y_1, Y_2, \cdots, Y_n)$ 均为线性链结构，若在给定随机变量序列 X 的条件下，随机变量序列 Y 的概率分布 $P(Y \mid X)$ 构成条件随机场，即满足马尔可夫性

$$P(Y_i \mid X, Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n,) = P(Y_i \mid X, Y_{i-1}, Y_{i+1}) \tag{2.5}$$

$$i = 1, 2, \cdots, n \quad （在i = 1和n时只考虑单边）$$

则称 $P(Y \mid X)$ 为线性链条件随机场。在标注问题中，X 表示输入的观测序列，Y 表示对应输出的标记序列或状态序列。

图 2.1 显示了线性链条件随机场的结构。在实际应用中，通常假设 X 和 Y 的图结构一致，此时线性链的无向图如图 2.2 所示，即

$$G = (V = \{1, 2, \cdots, n\}, E = \{(i, i+1)\}), \quad i = 1, 2, \cdots, n-1 \tag{2.6}$$

此时 $X = (X_1, X_2, \cdots, X_n)$，$Y = (Y_1, Y_2, \cdots, Y_n)$，最大团是两个相邻结点的集合。

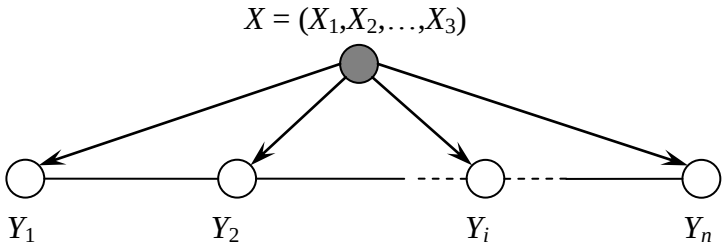

图 2.1　线性链条件随机场的结构

Fig. 2.1　Structure of linear chain conditional random field

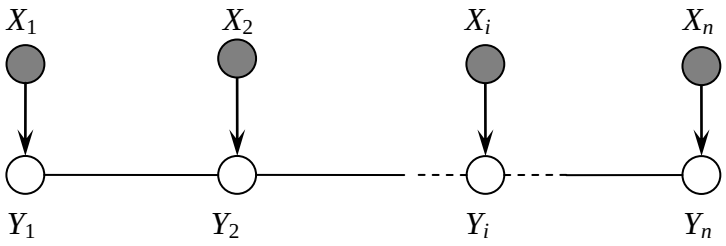

图 2.2　X 和 Y 具有相同图结构的线性链条件随机场

Fig. 2.2　Linear chain conditional random field with X and Y sharing the same graphical structure

2.2.3　参数化形式

设 $P(Y \mid X)$ 构成线性链条件随机场，则在随机变量 X 取值为 x 的条件下，随机变量 Y 取值为 y 的条件概率为

$$P(y \mid x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \tag{2.7}$$

其中

$$Z(x)=\sum_{y}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \tag{2.8}$$

式中 t_k 和 s_l 是特征函数，λ_k 和 μ_l 是对应的权值。$Z(x)$是规范化因子，求和是在所有可能的输出序列上进行的。

式(2.7)和(2.8)是线性链条件随机场模型的基本形式，表示给定输入序列 x，对输出序列 y 预测的条件概率。式(2.7)和(2.8)中 t_k 是定义在边上的特征函数，称为转移特征，依赖于当前和前一个位置；s_l 是定义在结点上的特征函数，称为状态特征，依赖于当前位置。t_k 和 s_l 都依赖于位置，是局部特征函数。通常，特征函数 t_k 和 s_l 取值为 1 或 0，即当满足条件时取 1，否则取 0。条件随机场完全由特征函数 t_k、s_l 以及对应的权值 λ_k、μ_l 确定。

2.3 外部资源

2.3.1 非结构化数据

本文使用的非结构化数据主要是各个医疗保健类社交网站上的用户评论，这些内容包含了用户用自然语言表述的用药反应的相关信息，我们采用文本挖掘技术从这些评论中发现药物的副作用信息，从而将其转化成方便利用的结构化数据。这类医疗保健类社交网站主要有 DailyStrength、AskaPatient、PatientsLikeMe 等。

DailyStrength 是一个以支持小组（Support Group）为特色的医疗社交网站，网站的用户在这里阐述他们的用药、精神以及生活方面的状态，讨论他们正在经历的与疾病抗争与治疗的情况，通过分享自己的治疗与用药经历为他人提供支持与帮助。DailyStrength 上的用户评论按药物或治疗方法分门别类，从网站上获取的评论内容所对应的药物是确定的，无需我们再通过文本挖掘等方法进行判定。每条评论的信息包括评论的文本内容、发表该评论的用户、用户所要治疗的病症以及利用该药物或治疗方法是否产生效果等。

AskaPatient 作为一个病人之间分享与对比用药经历的论坛，包含了自 2000 年开始的可用于比较与参考的药物与医疗信息，这些信息涉及 4,000 多种化学制备的处方药物以及一些流行的生物药品，参与其内容创作的人员包括图书馆员、医疗保健作家、健康专家以及分享自己经历的病人等。用户在这里可以按需搜索药物评论，撰写药物评论，

联系病友或回答他人的提问，关注他人的用药过程以便观察药物副作用的变化，对一些有争议的医疗问题进行投票以及获取多种来源的药物新闻等。

PatientsLikeMe 是一个病人之间沟通的网络，病友们可以通过这样一个在线社区分享他们治疗某种疾病的个人经历。该网站提供了一个在线资讯分享在平台，病人们可以在此处获取来自其他人的第一手用药与治疗信息以及临床试验的结果，寻找与自己相似的病友并互相了解、汲取经验，另外该网站也为科学研究提供了有价值的数据。

表 2.1 列举了一些比较知名的医疗保健方面的社交网站，这些网站为用户提供了药物治疗相关经历的分享平台，同时也是研究人员可挖掘利用的数据来源。

表 2.1　部分医疗保健类的社交网站

Tab. 2.1　Part of healthcare social networks

网站名称	网址
DailyStrength	http://www.dailystrength.org.
AskaPatient	http://www.askapatient.com/
PatientsLikeMe	http://www.patientslikeme.com/
WebMD	http://www.webmd.com/
FORCE(Facing Our Risk of Cancer)	http://www.facingourrisk.org/index.php
Breast Cancer Discussion Forums	http://community.breastcancer.org/
MedHelp	http://www.medhelp.org/
Health & Wellness - Yahoo Groups	https://groups.yahoo.com/neo/dir/1600060813
Medications.com - Prescription Drugs, Conditions, Interactions & Side Effects	http://www.medications.com/
Drugs.com | Prescription Drug Information, Interactions & Side Effects	http://www.drugs.com/
Find a Drug - Drug Reviews and Ratings - DrugRatingz.com	http://www.drugratingz.com/

2.3.2　结构化数据

在从用户评论中挖掘药物副作用时，我们通常需要借助一些副作用名称词典、药物副作用信息等数据，这些数据一般以结构化的形式进行存储。下面列举了与药物副作用相关的一些常用的数据库与术语集。

SIDER (Side Effect Resource) 数据库[41]：包含了上市药物及其不良反应的相关信息，这些信息来自于公开发表的文件与说明书。据 SIDER 主页显示的统计信息，该数据库收录了 996 种药物与 4,192 种副作用概念，共包括 99,423 个药物-副作用对信息，其中带有频率数据的药物-副作用对占 40.8%。随着数据的更新，SIDER 数据库收录的数据量在不断增加，其版本也从原来的 SIDER 1 升级到现在的 SIDER 2。

加拿大药物不良反应数据库 MedEffect[42]：其不良反应报告由消费者和医疗专业人士自愿提交，或由药品制造商和分销商按要求提交；用户可以从数据库获得药物和其他保健品的新安全信息。该数据库共收录了自 1965 年以来的 10,192 种药物与 3,297 种不良反应之间的关联信息。

COSTART（Coding Symbols for a Thesaurus of Adverse Reaction Terms）词汇库[43]：由美国食品药品监督管理局（U.S. Food and Drug Administration, FDA）为药物的不良反应监督管理而开发，用于上市药物不良反应报告的编码、归档和检索，共包含 3,787 个概念。COSTART 为处理提交给 FDA 的不良反应报告中的词汇变化问题提供解决方案，它使得这些报告中的用词规范化并保持一致。目前 COSTART 的功能已被 MedDRA 体系所替代。

MedDRA（Medical Dictionary for Regulatory Activities）术语集[44]：是在人用药物注册技术要求国际协调会（International Conference on Harmonisation, ICH）的主办下编制的国际医学术语集，其目的在于使国际医学术语集标准化，从而便于药事管理交流。MedDRA 中的术语按层级进行组织，共分为 5 级，分别为系统器官分类（System Organ Class, SOC）、高位组语（High Level Group Term, HLGT）、高位术语（High Level Term, HET）、首选术语（Preferred Term, PT）和低水平术语（Low Level Term, LLT）。MedDRA 术语集可用于上市药物不良反应的监测、药物不良反应的报告以及相应的数据分析等。

UMLS（Unified Medical Language System，一体化医学语言系统）[45]：由美国国家医学图书馆（U.S. National Library of Medicine）创建，包含多种文档与软件，聚合了不同的健康与生物医学词汇库与相关标准，是一个医学词典集合，同时也是一个展现术语（term）之间关系的语义网络。

本文在对药物副作用名称进行标准化过程中用到了 WordNet[46] 中的数据。WordNet 是一个由普林斯顿大学的心理学系发起，目前设在该大学计算机科学系的工程。它是一个庞大的英文数据库，它将英文中的名称、动词、形容词以及副词划分成一个个认知同义词集合，每个集合表示一个不同的概念。同义词集合之间通过概念语义与词汇关系相互联系，从而构成词汇与概念的网络。WordNet 可以免费、公开下载并利用，在计算语言学以及自然语言处理方面是一个非常有用的工具。

3 结合标准化技术的药物副作用挖掘

3.1 问题引出

由于语言表述的自由性与多样性，人们在表达同一个副作用概念时可能会采用不同的措辞方式，而新药的上市以及用药者的差异性又可能会导致新的副作用出现，故而用户评论中会存在数据库未收录的副作用名称，有的甚至是因拼写错误而造成的不同。因此从评论中挖掘药物的副作用时，识别新的副作用名称并将其映射到统一的副作用概念上是十分重要的，否则将无法发现一些潜在的副作用，或挖掘出的副作用发病率与事实存在偏差。以前的工作在对药物副作用进行识别时，或者利用了滑动窗口与词袋模型，或者通过副作用口语化表述的模式来识别，这些方法对新表述形式的识别程度有限，效果往往不够理想；另外他们对识别出的新副作用名称的后续处理也很有限，影响药物副作用的发现。

针对这些方法的不足，本文采用条件随机场模型识别药物副作用，可以有效识别出已知的以及新的副作用名称；对于识别得到的新的副作用名称，我们对其标准化并映射到已知的副作用概念上，使其得到有效的聚合与归并。

3.2 系统流程

本文从用户评论中挖掘药物副作用的整个流程分为数据准备、实体识别与过滤、实体标准化、药物副作用发现四个部分，如图 3.1 所示。我们以 DailyStrength 网站上用户对药物的评论作为语料，利用条件随机场模型识别出其中的副作用实体，然后使用本文提出的副作用实体标准化方法对识别出的副作用名称进行标准化，将其映射到统一的副作用概念上，最后统计每种药物评论下副作用概念的发生频率进而挖掘出药物的副作用。

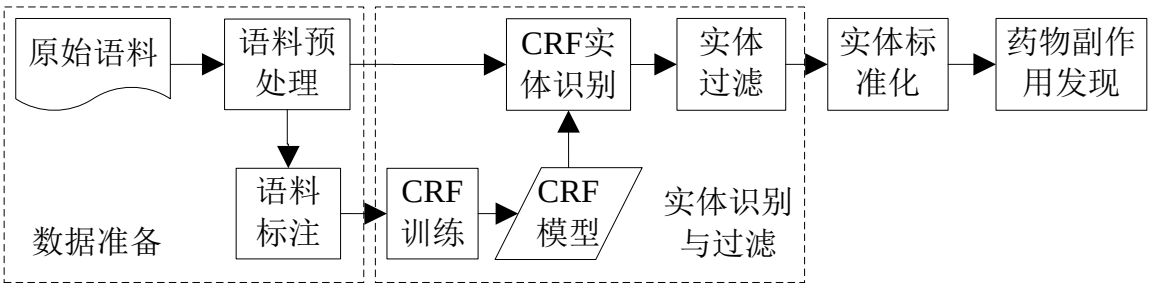

图 3.1 药物副作用挖掘系统架构图

Fig. 3.1 Architecture of the ADR mining system

3.2.1 数据准备

我们从 DailyStrength 网站上抓取了 SIDER 数据库中存在记录的 870 种药物在 2013 年 3 月 24 日之前的用户评论（英文）。DailyStrength 上的户评论是按药物分组的，即每个评论对应的药物是确定的，无需再对它们的关系进行判别。用户在撰写评论时具有一定的随意性，导致语料中存在一些不规范的语言现象。为了减少它们对后续处理造成的影响，我们对评论语料进行了一些预处理：将句首单词的首字母大写；在需要的评论语句后面加上表示结束的句号；修正一些不规则的写法（如!!! -> !, isnt -> is not, im -> I am, ive -> I have）等。

预处理之后，我们对评论内容进行句子划分，共得到 213,466 个不同的句子。从中随机抽取句子，以药物副作用词典中的副作用名称作为参考，对其进行人工标注。我们将句子中的实体分成两类，第一类是药物的副作用，第二类是药物治疗的病症或是从评论中无法判定为药物副作用的其他情况。因此我们的标注工作包括：注明实体在句子中的起止位置；注明实体在句子中属于第一类还是第二类。我们随机抽取并标注了 1,500 个含有实体（第一或第二类）的句子，将其作为后续实体识别的训练集；另外随机抽取出 500 个句子进行标注，将其作为实体识别的测试集（这些句子中有的包含实体，有的不包含，其分布情况与整体语料相同）。从用户评论中挖掘药物副作用这一领域，目前还没有一个权威、公开的标注数据集可以用来测试副作用实体识别方法的性能，因此我们利用自己标注的数据集来对本文方法识别药物副作用的效果进行测试。

3.2.2 副作用实体识别与过滤

我们采用 CRF 模型从用户评论语料中识别副作用实体，具体使用了开源的 CRF++ 工具包[47]。识别时利用了词语的两类特征：词语拼写本身与词语的词性，其中词性特征是采用 Stanford POS Tagger 工具包[48] 中的 english-left3words-distsim.tagger 模型对评论语句标注得到的。在利用 CRF 模型进行识别时，对于每个词语，我们考虑的特征包括当前词语与前两个、后两个词语的拼写与词性特征。

在训练集上对 CRF 模型训练完成后，我们在测试集以及所有的评论语料上进行实体识别。我们利用 CRF 识别出句子中的所有实体，识别完成后再进一步确定实体属于哪一类。对于实体的类别，我们参考 Leaman 等人[6] 的做法并略作改变，根据实体所在子句（Clause）是否含有某些特定词语来确定其类别。第二类实体所在子句中通常含有 ease、work for、help with 等表示治疗、缓解等意义的动词，为此我们收集了这样一个动

词表，并根据实体所在子句内是否含有这些动词确定其为第一或第二类实体。此外，为了提高实体识别的准确性，我们还检测了子句中的否定词，并据此进一步滤除非药物副作用的实体。

3.2.3 副作用实体标准化

在药物的副作用标准化中，药物的每种副作用均被视为一种副作用概念，它对应着一个或多个表述形式即副作用名称。实体标准化就是通过一定的手段将实体映射到对应的标准概念上，一般可分为精确匹配（Exact Matching）和近似匹配（Approximate Matching）两种方法。在本文中，对于从评论里识别出来的副作用名称，若词典中存在该名称，则直接通过精确匹配得到对应的标准化概念；否则进行近似匹配，即利用本文所述的近似匹配方法将其映射到标准化概念上，或者该实体在词典中无法找到对应的概念，而属于一种新的副作用概念。

1. 方法流程

对于一个待标准化实体（副作用名称），如果精确匹配成功，则直接得到标准化概念；否则我们通过近似匹配从副作用词典中寻找与之最相关的副作用名称，并将该名称对应的概念作为标准化概念。本文的近似匹配部分由 3 个模块组成，这 3 个模块分别基于常规检索、扩展语义检索以及编辑距离进行标准化。本文提出的药物副作用标准化方法的流程如图 3.2 所示。

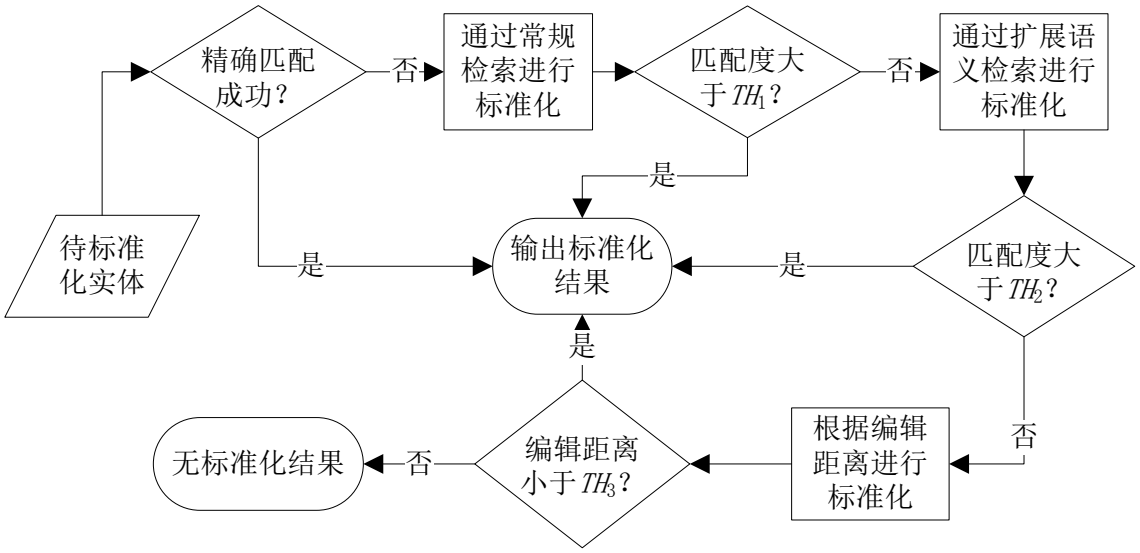

图 3.2 药物副作用名称标准化方法的流程图

Fig. 3.2 Flowchart of the ADR entity normalization

在本文提出的标准化方法中，首先通过常规检索进行标准化，若得到的匹配度大于设定的阈值 TH_1，则将相应的概念作为标准化结果；否则通过扩展语义检索进行标准化，若得到的匹配度大于设定的阈值 TH_2，则将相应的概念作为标准化结果；否则根据编辑距离进行标准化，若得到的最短编辑距离小于设定的阈值 TH_3，则将相应的概念作为标准化结果；否则词典中没有任何概念与当前待标准化实体匹配，该实体可能属于一种新的副作用概念。

2. 近似匹配模块

本文近似匹配 3 个模块中，前两个模块利用信息检索中的 TF-IDF 思想初步限定候选概念范围，然后通过计算副作用名称之间的匹配度进一步确定标准化概念；第 3 个模块则根据最短编辑距离寻找最佳的标准化概念。以下是本文近似匹配中用到的 3 个模块：

模块 1：通过常规检索进行标准化

将每个副作用概念视为一篇文档，概念下的所有副作用名称作为文档内容，对其分词、去停用词以及词干化处理后建立索引（Index）。将待标准化实体作为查询（Query）在索引中检索，返回的前 N_1 个概念作为候选概念（最多 N_1 个，若查询返回结果太少则少于 N_1 个）。然后利用匹配度函数（见下文）计算待标准化实体与候选概念中每种副作用名称之间的匹配度，并将匹配度最大的名称对应的概念作为该实体的标准化概念。若检索结果为空，则该模块的标准化结果为空。

模块 2：通过扩展语义检索进行标准化

若待标准化实体所含词语在其他副作用名称中很少出现，这时直接将其作为查询进行检索的效果可能不好。因此我们利用 WordNet 数据对待标准化实体中的词语进行语义扩展，使每个词语都扩展得到最多 5 个同义词（有些词语在 WordNet 中的同义词数量可能少于 5 个）。与模块 1 建立相同的索引，然后将扩展后的词语集合作为查询在索引中检索，得到（最多）前 N_2 个最相关的候选概念。我们把扩展后的词语集合视为待标准化实体，同样利用匹配度函数计算它与各个候选概念的副作用名称的匹配度，并将匹配度最大的名称对应的概念作为它的标准化概念。若检索结果为空，则该模块的标准化结果为空。

模块 3：根据编辑距离进行标准化

副作用名称中存在以下现象：① 两个词语的意义相同但拼写却存在一定差别（如“病毒血症”的两种拼写 viremia 与 viraemia）；② 某一词语为另一短语的缩写形式（如概念 C0079773 的副作用名称有 CTCL 和 cutaneous T cell lymphoma，前者为后者的

首字母缩写）。待标准化实体中若含有这些词语，则检索结果可能不理想，无法命中正确的概念。这种情况下我们根据字符串之间的编辑距离寻找与之最匹配的名称。

编辑距离（Edit Distance）用来衡量两个字符串字面上的相异性，它是指从其中一个字符串转换成另一字符串所需要的最小代价（一般指操作的步数）。本文在利用编辑距离量化两个副作用名称的相异程度时，将名称短语视作词袋。对于词袋中的每个词语，为其在另一词袋中找到与之编辑距离最短的词语作为匹配，从而将两个词袋中的词语两两配对，最终两个短语的相异程度为这些配对词语的编辑距离之和。通过该方法计算得到与待标准化实体编辑距离最短即字面上最相似的副作用名称，并将其对应的概念作为该实体的标准化概念。

在计算两个副作用名称之间编辑距离时，我们考虑了其中某个词语为缩写词的情况。一般来说，缩写词通常为某个短语单词首字母的缩写，或单词中前缀首字母加上剩余部分首字母的缩写（为此我们收集了一个英文前缀表）。因此在计算两个短语之间的编辑距离时，若某词语为另一短语（或其子串）的缩写词，则其编辑距离为 0。

3. 匹配度函数

模块 1 和 2 中利用匹配度函数计算待标准化实体 Ent_1 与另一副作用名称 Ent_2 之间的匹配度 *MD* (Match Degree)，其具体的计算过程为：

① 将 Ent_1 与 Ent_2 进行分词、去停用词、词干化处理，分别得到词袋 *A* 和 *B*。

② 对于词袋 *A* 中的每个单词 a_m （$m = 1,2,\cdots,p$，p 为 *A* 中单词数），遍历词袋 *B*，由如下公式计算出 a_m 与 *B* 中每个单词 b_n 的字面相似度（Literal Similarity）$LS(a_m,b_n)$：

$$LS(a_m,b_n) = 2 \cdot N_{cc} / (L_a + L_b) \tag{3.1}$$

其中 N_{cc} 是 a_m 与 b_n 的公共子串长度，L_a 为 a_m 的字符数，L_b 为 b_n 的字符数。若 N_{cc} 小于设定的阈值，则认为不具有表征 a_m 与 b_n 内在相关性的作用，并且可能会作为噪音影响计算结果，将其置为 0。

从 *B* 的所有单词与 a_m 的字面相似度中找出最大值作为单词 a_m 最终的字面相似度 LS_m，并将与之匹配的单词从 *B* 中删除，使之不重复作为 *A* 中其他单词的最佳匹配。这样，实体 Ent_1 与 Ent_2 之间的字面相似度 *LS* 为

$$LS = \sum_{m=1}^{p} LS_m / (L_a + |L_a - L_b|) \tag{3.2}$$

③ 计算 Ent_2 对应的概念 Con 涵盖 Ent_1 中单词的程度 *WC* (Word Coverage)。Con 的全词袋 $C = \cup B_i$（B_i 为 Con 的第 i 个实体对应的词袋，$i = 1,2,\cdots,q$，q 为 Con 中实体数），词袋 *A* 与 *C* 之间的相同单词集合为 $|A \cap C|$，则 Con 涵盖 Ent_1 中单词的程度 $WC = |A \cap C| / |A|$。

④ Ent_1 与 Ent_2 之间最终的匹配度 $MD = LS + r \cdot WC$（r 为 0~1 之间的比例系数）。

3.2.4 药物副作用发现

根据从评论语料中发现的所有副作用名称，并参照标准化结果，我们得到每种药物的评论中出现的副作用概念及包含此概念的评论所占的比例即发生频率。对于发现的药物已知的副作用，我们将其与已有的数据进行对比，验证本文挖掘方法的有效性；对于数据库中未记录的药物副作用，我们按其在评论中的发生频率由高到低排序，得到一个药物潜在副作用列表。

3.3 实验结果与分析

3.3.1 副作用实体识别

1. 实体识别效果测试

我们在标注好的 1,500 个评论语句上训练 CRF 模型，然后利用该模型对测试集中的 500 个评论语句进行副作用实体识别，将识别出的实体进行过滤后，得到实体识别的准确率为 87.5%，召回率为 58.7%，F 值为 70.3%。对错误识别的样例进行分析，我们发现 CRF 模型不能识别由分散的词语构成的实体而造成识别错误，如无法从“… major swelling in my ankles and …”识别出副作用 ankles swelling；另外部分错误是由于识别出的实体与标准答案不完全相同造成的，例如 general feeling of illness 与 illness、frequent headaches 与 headaches 等（前者为标准答案，后者为识别结果）。

2. 从评论中识别副作用实体

我们从 870 种药物的 408,318 条评论中识别实体并过滤后，得到了 729 个词典中存在的副作用名称与 3,143 个新的副作用名称，表 3.1 显示了本文挖掘出的词典中已有名称与新名称的统计情况（括号中为对应数值占总体的百分比），表 3.2 显示了识别出的出现频率最高的前 10 个新的副作用名称。从结果可以看到，利用 CRF 模型不但识别出了已知的副作用名称，而且能够识别出潜在的新副作用名称。由表可知，新名称出现的总次数占总体的 18.0%，而平均出现次数相对于已知名称却少得多，说明用户在评论中使用新的、不同的副作用表述方式是很普遍的，因此进行标准化是很有必要的。

表 3.1 药物副作用实体识别结果统计

Tab. 3.1 Statistical results of ADR entity recognition

	数量	出现总次数	每个名称平均出现次数
已知的副作用名称	729 (18.8%)	58810 (82.0%)	80.7
新的副作用名称	3143 (81.2%)	12932 (18.0%)	4.1
总计	3872	71742	--

表 3.2　识别出的频率最高的前 10 个新的副作用名称

Tab. 3.2　Ten most frequent new mined ADR names

排名	副作用名称	出现次数	排名	副作用名称	出现次数
1	gained weight	814	6	stomach pains	94
2	heart race	153	7	shakes	86
3	gaining weight	131	8	stomach hurt	78
4	shaking	125	9	suicidal	74
5	painful	123	10	stomach aches	71

3.3.2　副作用实体标准化

我们利用 SIDER 数据库中的药物副作用数据创建了一个副作用词典，其中共包含 5,719 个副作用概念。每个副作用概念拥有一个统一医学语言系统（Unified Medical Language System, UMLS）的概念编号 CUI (UMLS Concept Id)，并由含义相同的一种或多种副作用名称构成。例如，CUI 为 C0239739 的概念有[sore gums, gum pain, gingival pain, gum tenderness] 4 种意义相同的副作用名称。

为了验证提出的标准化方法的有效性，本文首先对标准化的准确率进行了测试。在测试时，我们从副作用词典中随机抽取满足要求（即该副作用名称在词典中须有属于同一概念的其他副作用名称）的副作用名称作为待标准化实体，同时将该名称从词典中删除，并对删除该名称后的词典建立索引。利用上述的标准化方法得到该实体的标准概念，并与正确的标准概念对比，从而得到标准化的准确率。在测试该标准化方法时，我们对其中的 3 个阈值 TH_1、TH_2、TH_3 调优，并将最优的阈值用于从评论中识别出的药物副作用的标准化中。

1. 检索返回候选概念的数量

为了合理设置检索返回候选概念的数量，我们对 500 个待标准化实体进行常规检索并统计返回的前 n 个候选概念中包含正确概念的比例，结果如图 3.3 所示。可以看出随着返回候选概念数量的增加，结果中包含正确概念的比例逐渐变大，当返回候选概念数量 n 为 20 时该比例已达 82.0%；但增速却逐渐变缓，当 n 为 30 时该比例为 83.2%，仅增加了 1.2%，最终很难达到理想的 100%。造成这种现象的一个可能原因是有些待标准化实体为某些生僻词或缩写词，索引中几乎没有与其拼写相同词语，从而无法通过常规检索返回正确的概念。这也是需要利用扩展语义检索与编辑距离进行标准化的原因。

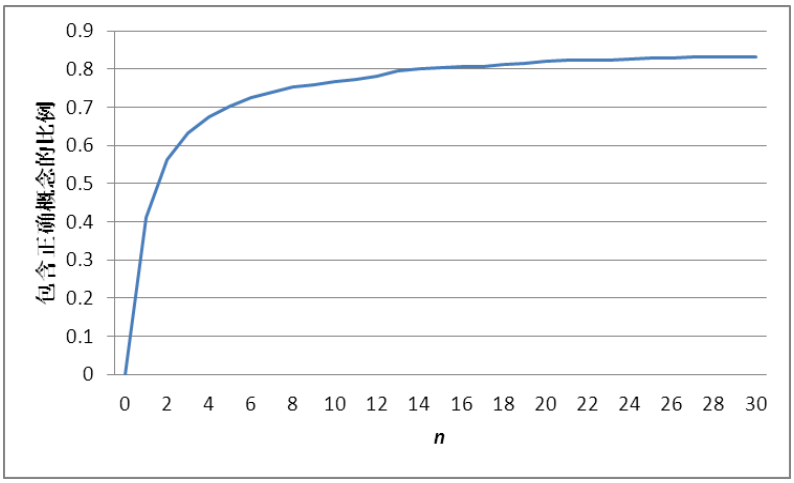

图 3.3　常规检索返回的前 n 个候选概念中包含正确概念的统计概率

Fig. 3.3　Statistical probabilities of inclusion of right concept in the first n concepts returned by normal retrieval

综合考虑检索结果包含正确概念的比例以及检索的效率，我们在实验中将常规检索返回的候选概念数量 N_1 设置为 25；而扩展语义后查询词语得到扩充，与之相关的候选概念数量也会相应地增多，因而我们将扩展语义检索返回的候选概念数量 N_2 设置为 40。

2. 标准化方法测试与分析

为了测试各个模块对标准化准确率的提升作用，我们分别采用“模块 1”、“模块 1 + 2”、“模块 1 + 2 + 3” 3 种组合方式对副作用名称标准化。每种组合方式进行 10 次实验，每次从词典的 10,498 个副作用名称中随机抽取 500 个用于测试，并根据标准化结果计算其准确率。标准化方法测试的结果如表 3.3 所示。

表 3.3　本文标准化方法的测试结果

Tab. 3.3　Test result of the proposed normalization method

组合方式	准　确　率		
	最大值	最小值	平均值
模块 1	68.8%	65.2%	66.7%
模块 1 + 2	70.2%	67.0%	69.1%
模块 1 + 2 +3	72.8%	69.8%	71.4%

由实验结果可以看出，我们的匹配度函数确实在一定程度上反映了副作用名称之间的内在联系，使得大部分待标准化实体映射到了其正确的概念上。模块 2 的加入使标准化的准确率有了提升，说明将待标准化实体进行语义扩展，通过同义词语寻找正确概念的做法在涉及一些低频率词语时具有益处。添加模块 3 后标准化的准确率进一步提升，说明副作用名称中包含一定数量的缩写词以及意义相同、词形相近的词语，此时根据编辑距离进行匹配具有较好的效果，同时也是对前两个模块功能的补充。

分析标准化结果中错误的实例，我们发现了以下几种导致标准化错误的情况。

1）有些形式十分接近的副作用名称属于不同的概念，在对其中某个名称标准化时会错误映射至另一名称对应的概念。例如概念 C0018772 下的 impaired hearing 与概念 C1384666 下 hearing impairment 在词干化并忽略词序后完全匹配，但它们却属于不同的概念。

2）利用 WordNet 数据对副作用名称扩展语义时，由于 WordNet 本身的局限性，有时并不能将合适的词语扩充进来。例如在对概念 C0549448 下的 elevated hemoglobin 标准化时，WordNet 并不能将 increase 扩充为 elevate 的同义词，从而无法匹配到同概念的 increased hemoglobin。

3）副作用名称中的专业词汇常常无法得到扩展，而专业词汇与同概念下的其他名称在词形上的关联又很弱，从而导致标准化错误。例如 cholelithiasis 属于概念 C0008350，而此概念下的所有名称为[gall stone, gallstones, cholelithiasis, biliary calculi]。

3. 对识别出的副作用实体进行标准化

对 3,143 个新的副作用名称进行标准化处理，其中 2,337 个新名称映射到了 974 个概念上，平均每个概念约对应 2.4 个新名称；剩余的 806 个新名称无法对应到词典中已有的概念上，可能属于新的副作用概念。图 3.4 显示了副作用概念 C0043094 在词典中已有的名称以及本文从评论中挖掘的新名称，其中实线框中的是词典中已有的名称，虚线框中的是挖掘出的新名称（有些拼写是错误的）。可以看出，通过对新名称进行标准化，我们可以将用户对同一概念的不同表述形式（包括评论中常见的因拼写错误而产生的不同形式）映射到其真正所指的概念上，实现副作用名称的有效聚合与归并，使副作用概念在评论中的发生频率更接近其在用药者中发生的真实概率，从而有利于药物潜在副作用的发现。

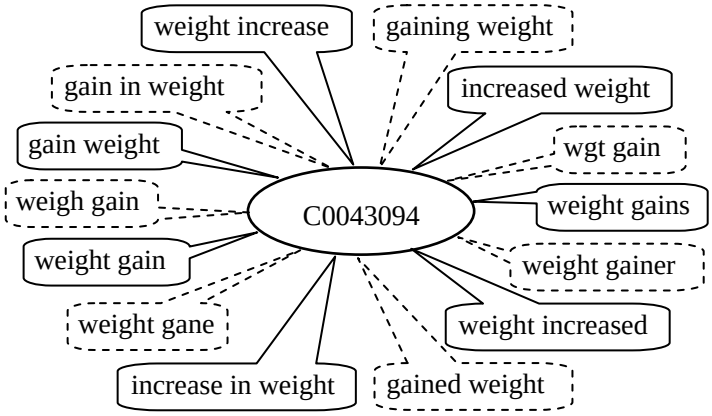

图 3.4　概念 C0043094 在词典中已有的副作用名称及本文挖掘的新名称

Fig. 3.4　Existing names in the dictionary and new mined names for concept C0043094

3.3.3　药物副作用发现

通过对识别出的副作用名称进行标准化，我们将不同的表述形式映射到了其所指的概念上，从而可以统计出副作用概念在每种药物评论中的发生频率。为了避免偶然现象而使结果更具统计意义，我们选择评论数量大于 50 的药物，将挖掘出的副作用概念按照在一种药物评论中的发生频率由高到低排序，得到药物副作用的列表。

对于药物已知的副作用，我们将挖掘出的发生频率与 SIDER 数据库中记录的发生频率进行了对比。表 3.4 显示了挖掘得到的发生频率最高的前 10 种已知的药物-副作用对与数据库中记录的相应数据，其中“postmarketing”表示副作用在药物上市后得到确认，“potential”表示可能为药物的副作用。从表中可以看出，我们从评论中挖掘出的具有较高发生频率的药物副作用，其在数据库中记录的发生率一般也相应地较高，两种来源的药物副作用发生频率具有一定的相似性与对应性，说明本文的药物副作用挖掘方法的有效性。

表 3.4　挖掘的发生频率最高的 10 种已知的药物–副作用对与数据库记录之间的对比

Tab. 3.4　Mined frequencies versus database records of top 10 most frequent known drug-ADR pairs discovered by us

排名	药物	副作用概念	挖掘得到的发生频率	数据库记录的发生频率
1	Nitrostat (Nitroglycerin)	C0018681[headache, …]	18.5%	18%
2	Mirena (Provera)	C0030193[unspecified pain, …]	13.7%	potential
3	Gleevec (Imatinib)	C0027497[nausea, …]	13.6%	9%
4	Zyprexa (Olanzapine)	C0043094[weight gain, …]	12.0%	5%
5	Oxytrol (Oxybutynin)	C0043352[dry mouth, …]	11.6%	8.33%

6	Indocin (Indomethacin)	C0038354[gastric disorder, …]	11.5%	potential
7	Lupron (Leuprolide)	C0600142[hot flushes, …]	10.7%	postmarketing
8	Doxorubicin	C0027497[nausea, …]	10.1%	18.2%
9	Danazol	C0085633[mood swings, …]	10.0%	potential
10	Parlodel (Bromocriptine)	C0012833[dizziness, …]	9.3%	postmarketing

对于发现的数据库中未记录的药物副作用，可以认为副作用概念在某种药物的评论中的发生频率越高，则其为该药物潜在副作用的可能性越大。因此，我们按挖掘到的发生频率由高到低排序，得到了一个可能性由大到小排列的药物潜在副作用列表。表 3.5 显示了本文挖掘到的前 10 个最有可能的潜在药物-副作用对。对于挖掘出的具有较高发生频率的药物副作用，可以作为药物潜在的副作用以备参考。

表 3.5　挖掘的发生频率最高的前 10 种潜在药物–副作用对

Tab. 3.5　Top 10 most frequent potential drug-ADR pairs mined in this paper

排名	药物	副作用概念	挖掘得到的发生频率
1	Arthrotec	C0038354[gastric disorder, …]	15.7%
2	Piroxicam	C0038354[gastric disorder, …]	11.3%
3	Robaxin	C0234450[sleepy]	7.7%
4	Tussionex	C0030193[unspecified pain, …]	7.6%
5	Nitrostat	C0008031[chest pain, …]	7.4%
6	Carboplatin	C0015672[fatigue, …]	7.2%
7	Seasonique	C0030193[unspecified pain, …]	7.2%
8	Fosavance	C0038354[gastric disorder, …]	6.8%
9	Mirena	C0026821[cramps, …]	6.8%
10	Danazol	C0149931[migraine, …]	6.7%

3.4　本章总结

从社交网络的用户评论中提取药物副作用信息是一种快捷、有效的渠道，而评论中含有大量数据库未收录的副作用名称，识别这些新名称并标准化对药物副作用的挖掘十分重要。针对前人工作中对新副作用名称的识别效果不佳以及对识别出的新名称后续处理不足的问题，本文利用 CRF 模型识别评论中的副作用，可以识别出已知的与新的名称。将副作用名称标准化可以对其进行有效的聚合与归并，有利于药物副作用的发现。我们通过将挖掘出的药物已知的副作用与数据库记录进行对比验证本文方法的有效性，对挖掘出的数据库未记录的药物副作用按其在评论中的发生频率排序，得到了一个可能性由大到小排列的药物潜在副作用列表。

在未来工作中，1）鉴于 WordNet 数据存在的局限性，在标准化时可以考虑引入生物医学领域的专业词典，或是借助语义相似度数据来衡量词语之间的关联程度，提高标准化方法的准确率；2）对于挖掘出的新的副作用名称，如果无法映射到现有的副作用概念上，则考虑通过它们之间的关联度将其进行聚类与归并，更好地发现药物潜在的副作用以及新的副作用概念。

4 基于方法融合的副作用实体识别

4.1 问题引出

从医学社交网站的用户评论中识别副作用实 4F53 是药物副作用挖掘的关键步骤，实体识别的效果对药物副作用的发现有重要的影响。与其他文本中的实体识别相比，从用户评论内容中识别副作用名称存在以下几个难点：

1）用户评论通常都是一些口语化的表达，因此构成副作用的部分词语的专业、领域特征不明显，较多词语是非专业普通用词（如 anxious、extreme fatigue），经典英文命名实体（人名、地名、机构名等）的一些特征（如单词首字母大写）在副作用名称中也不明显。

2）构成副作用实体的词语不一定相邻分布，可能由分散的词语组合形成。如“There is pain in my head.”存在由分离的 pain 与 head 构成的副作用名称 head pain。

3）不同的用户对同一种副作用采取的表述形式可能不同，评论中的副作用表述形式具有多样性。如对于副作用“增重（gain weight）”，我们在用户评论中发现的就有 weight gain、gain in weight、weight increased、increase in weight、wgt gain、gain pounds 等多种表述形式。

4）由于撰写存在一定的随意性，用户评论中的单词会存在拼写上的错误，如在评论中我们发现“增重”一词存在 weight gane、weigh gain 等错误的拼写。

5）用户评论中存在一定数量语法上不太规范的语句（如“Made me sick loss of appetite nose bleeds dizzier than usual”中全句都未加标点），以及一些语法结构不明显的短句（这类短句子通常直接由若干副作用名称直接构成，如“Anxiety, mood swings, nausea, vomitting .”）。

由于上述几个因素，从用户评论中识别副作用名称具有一定的挑战性，目前已知的用于此任务的识别方法的性能尚不够理想，具有较大的提升空间。例如，Leaman 等人[6]在人工标注的 DailyStrength 评论数据集上识别的 F 值为 73.9%，Nikfarjam 等人[10] 在人工标注的 DailyStrength 评论数据集上识别的 F 值为 67.96%，Sampathkumar 等人[12] 在人工标注的 Medications.com 上的评论数据上识别得到的 F 值为 73.2%。因此本章工作对用户评论中的副作用实体识别进行专门的研究，利用基于词袋（Bag of Words）匹配与基于 CRF 模型两种方法识别副作用，并通过将它们的识别结果进行融合来提高最终的识别效果，在测试数据集上进行实验取得了良好的结果，证明了本文方法的有效性。

4.2 算法设计

精确的词典匹配可以识别出与词典项完全相同的、由连续词语构成的副作用名称，而使用滑动窗口中的词袋匹配则可以进一步识别出由分散词语构成的副作用名称，如在句子“There is pain in my head.”中，若设置的滑动窗口大小不小于 4，则可以识别出副作用名称“head pain”。然而，基于词典匹配的方法无法识别出词典中不包含的新副作用名称，其识别效果依赖于所用词典的完备性，而用户评论中副作用表述形式的多样性普遍存在，因此仅利用词典匹配的方法无法完全识别出评论中的所有副作用名称。

利用条件随机场模型识别评论中的副作用实体时，它不但可以识别出训练集中已有的实体，而且还能够识别出训练集中不包含的新副作用名称。但是由于评论撰写的随意性，其内容中存在一定数量的不太规范的句子以及无明显语法结构的短句子（这类短句子通常直接由若干副作用名称构成），由于条件随机场模型在识别时需要考虑词语的上下文特征，因此这类句子中实体的识别对条件随机场模型来说会存在一定的困难。

基于上述观察，本文提出了一种融合滑动窗口中的词袋匹配与基于条件随机场模型识别两种手段的副作用实体识别方法，通过融合克服了单一实体识别方法存在的不足，使方法之间的优势得到互补，从而提高副作用实体识别的效果。

4.2.1 基于词袋匹配的识别

本文利用 SIDER 数据库中的药物副作用数据构建了一个用于匹配的副作用词典，共包含隶属于 5,719 个副作用概念的 10,498 个副作用名称。对用户评论语料进行分句，得到独立的句子，然后利用基于词袋匹配的方法识别出句子中的副作用实体。

1. 匹配算法

基于词袋匹配方法的主要思想是将滑动窗口（Sliding Window）中的词语与词典中的实体均视作词袋，根据词袋之间的包含关系识别出滑动窗口中的实体。考虑到用户评论中存在单词拼写错误的情况，本文在确定实体对应的词袋是否被滑动窗口对应的词袋包含时，并非只是严格意义上的包含，而允许前者中的单词与后者中匹配的单词存在一定的编辑距离。只要词袋中所有单词的编辑距离总和不超过设定的阈值，则认为两个词典存在包含关系。因此，对于一个待识别实体的句子，本文基于词袋匹配的识别流程如下：

1）对输入的待识别句子进行分词，得到一个由单词和标点符号组成的 token 数组。

2）让大小为 N_{SW} 的滑动窗口在 token 数组上从左向右逐词移动，每次取出滑动窗口中的单词（或标点），并对其进行词干化处理，若是停用词则舍弃，从而得到由此时窗口内的单词构成的词袋 A。

3）对副作用词典中的实体进行同样的分词、去停用词以及词干化处理，得到该实体的词袋 B。

4）若词袋 A 包含词袋 B 中的所有元素，则将词袋 B 对应的词典中的实体添加到该句子已识别出的实体集合 S 中；否则转 5）。

5）对于词袋 B 中的每个单词，找出词袋 A 中与之编辑距离（Edit Distance）最短的单词，并将它们的编辑距离加入词袋 B 相对于 A 的编辑距离总和 D 中。如果词袋 B 相对于 A 的编辑距离总和 D 小于设定的阈值 TH_{ED}，则认为词袋 A 包含词袋 B，此时将词袋 B 对应的词典中的实体添加到该句子已识别出的实体集合 S 中；否则此时滑动窗口中不包含词袋 B 对应的实体。

6）当滑动窗口从左至右移动完毕后，返回集合 S 中包含的所有识别出的实体。

2. 识别结果的过滤

在识别出的副作用实体中，可能存在两个实体包含句中同一个词语的情况。考虑到含有较少词语的名称通常为含有较多词语名称的一部分，而含较多词语的名称通常能够更精确地表示句子中提到的副作用，因此这种情况下我们只保留包含词语最多的副作用名称而舍弃其他名称，从而可以在几乎不损失召回率的同时增加识别的准确率。例如当窗口大小 $N = 5$ 时，从句子“Right foot turn red with swelling and much pain last week.”中通过滑动窗口可以匹配得到两个副作用名称 swelling 和 foot swelling，而 foot swelling 显然更精确地表明了副作用，因此只保留 foot swelling 而舍弃 swelling。

4.2.2 基于条件随机场的识别

这里使用开源的 CRF++工具包[47] 进行条件随机场模型的训练与预测。在应用条件随机场模型时，本文对其做了词语内部特征的选择与相关参数设置，其中的参数设置是指词语的上下文特征范围与特征组合方式的确定。本文利用条件随机场识别时所做工作对 CRF 模型是通用的，与具体采用何种 CRF 工具包实现无关。

1. 词语内部特征的选取

利用条件随机场模型进行副作用名称识别时，本文考虑每个词语（英文单词）的候选内部特征包括：拼写，词性，在副作用词典中的频率，在副作用词典中作为副作用实体首词语的频率，在副作用词典中作为副作用实体尾词语的频率，如表 4.1 所示。由于药物副作用名称没有明显的首字母大写等特征，因此获取单词的拼写特征时，一律取单词的小写形式。单词的词性是采用 Stanford POS Tagger 工具包[48] 中的 english-left3words-distsim.tagger 模型对评论语句标注得到的。在查询某个词语在副作用

词典中的频率时，我们将查询词以及词典中的所有词语进行相同的词干化处理，然后统计词语在词典中相应的频率。

表 4.1　词语的候选内部特征

Tab. 4.1　Candidate internal features of a word

特征编号	特征含义
F1	拼写
F2	词性
F3	在副作用词典中的频率
F4	在副作用词典中作为实体首词语的频率
F5	在副作用词典中作为实体尾词语的频率

在得到词语候选的内部特征后，本文采用向前选择法（Forward Selection）[49] 对候选内部特征进行选择，即从空集开始，逐个添加候选特征，每次选择提高识别性能最多的特征加入进来，直到进一步添加不会提高识别的性能为止。通过对词语内部特征的选择，可以剔除对提升识别效果帮助不大的内部特征，并得到用于 CRF 识别的最佳的词语内部特征集。

2. 上下文范围与特征组合方式的确定

在确定一个词语是否为副作用实体或其一部分的时候，仅仅依据当前词语的内部特征进行识别的效果并不会理想，因为一个词语是否为副作用实体或其一部分还依赖于其所在的上下文，上下文特征对实体的识别有重要影响。词语的上下文特征是指其左右若干位置上的词语特征，直观上看，考虑其左右词语的数量越多，利用的信息也就越多，识别的效果就会更好。然而如果考虑上下文的词语过多，系统运行的效率会受到严重影响，同时可能形成过拟合，给实体识别带来噪音，反而降低识别的效果。

在本文工作中，我们考虑了不同上下文范围以及词语内部特征的组合方式对 CRF 识别实体效果的影响，以便获取 CRF 识别性能最好时的上下文范围与内部特征组合方式。试验过程中，以当前词语作为起点，将上下文范围分别向左右两边延伸。在每个上下文范围之下，进一步测试该范围内词语特征的组合方式对识别效果的影响。以上下文范围“左边第 1 个词语 + 当前词语 + 右边第 1 个词语”（此时窗口大小为 3）为例，此时所有可能的 1-gram、2-gram、3-gram 特征如表 4.2 所示。我们选择这些特征的组合作为 CRF 模型的特征进行试验，测试此时模型的识别效果，并将最佳识别效果对应的内部特征组合方式作为后续实验中 CRF 模型的配置。

表 4.2　一个窗口大小为 3 的上下文中所有特征及其组合

Tab. 4.2　Features and their combinations based on a context of current word (window size = 3)

特征类别	特征含义
1-gram 特征	当前词语的内部特征 左边第 1 个词语的内部特征 右边第 1 个词语的内部特征
2-gram 特征	当前词语与左边第 1 个词语组合对应的内部特征 当前词语与右边第 1 个词语组合对应的内部特征
3-gram 特征	当前词语、左边第 1 个词语与右边第 1 个词语组合对应的内部特征

3. 语料的标注与识别结果的后处理

在对评论中的实体进行标注以便用于 CRF 的训练与测试时，本文采用的标注记号包括 B、I、E、W、O 五个[50]，每个记号的具体含义如表 4.3 所示。表 4.4 显示了使用这套标注记号对句子进行标注的实例。

表 4.3　CRF 数据中的标注记号

Tab. 4.3　Annotation tokens in the data for CRF

标注记号	含义
B	多个单词实体的首单词
I	多个单词实体的中间单词
E	多个单词实体的尾单词
W	一个单词的实体
O	非实体单词

表 4.4　CRF 数据标注实例

Tab. 4.4　Annotation examples in the data for CRF

句子 1	标注	句子 2	标注	句子 3	标注
Made	O	Had	O	Had	O
me	O	weight	B	one	O
too	O	gain	E	episode	O
sleepy	W	side	O	of	O
the	O	effects	O	shortness	B
next	O	.		of	I
day	O			breath	E
.	O			.	O

在标注语料时，对于句子中的 pain in the head、lost 30 pounds、gain a lot of weight、loss of memory and concentration 等短语，虽然它们所含的副作用实体应为 head pain、lost pounds、gain weight、loss of memory 和 loss of concentration，但是鉴于 CRF 的标注方式不便表示由分散的词语构成的实体，我们将这些短语整体标注为一个实体用于 CRF 的

训练。相应地，在对于 CRF 的识别结果中，若识别出的实体中存在“and”或“,”，则将该实体从此处分开，使其成为两个实体。对于识别结果为 pain in the head、gain a lot of weight 这种情况，我们认为它们与标准答案 head pain、gain weight 吻合，识别结果正确。

4.2.3 非药物副作用实体的过滤

在评论内容中，用户提到的一些副作用实体在上下文中并非为该药物的副作用，而可能是指药物所治疗的病症，或用户表示担心但并未发生的情况，或否定当前药物具有某种副作用等情况。因此在识别出句子中的实体后，我们需要进一步确定实体在当前的评论上下文中是否为该药物的副作用。本文参考并改进了 Leaman 等人[6] 的做法，采用一种启发式方法确定实体是否为评论所属药物的副作用（该方法与第 3 章中所述相似，但在一些步骤与细节上进行了改进）。

观察包含实体但并非为药物副作用的句子，可以发现这种句子中通常含有一些指示性的词语，如 ease、work for、help with 等表示治疗、缓解等意义的动词，或 worry、afraid 等表示用户担心但实际并未发生的情况，以及 no、not 等表示否定含义的词语。为此本文收集了一个非药物副作用指示词表，并根据实体某个范围的上下文中是否含有这些指示词来确定该实体在句子中是否为药物的副作用。为了确定与实体类别密切相关的上下文范围，我们从实体所在位置出发，分别向左右两边搜索，在遇到“,”“.”等标点符号或“but”“however”等表示转折意义的词语或句子起始处时停止，左右停止位置之间的部分即为所需要的实体上下文范围。

为了得到非药物副作用的指示词列表，本文使用了一种自动获取加人工筛选的方法。具体做法是对于训练集含有非药物副作用实体的句子，首先确定与每个实体类别相关的上下文范围，然后统计所有上下文范围内词语的频率，并按频率由高到低对词语进行排序，然后通过人工观察并结合对词语的理解，取出排序靠前、具有强指示意义的词语加入指示词列表，同时剔除那些频率虽高但指示意义不明显的词语。另外，我们还通过对句子的直接观察向非药物副作用指示词列表中补充了一些具有指示意义词组。

4.2.4 识别结果的融合

对于同一个句子，通过两种方法分别识别其中的副作用实体后，本文将它们的识别结果进行融合，得到该句子所有识别出的实体。在融合后的结果中，若存在两个实体包含句中同一个词语，考虑到含有较少词语的实体通常为含有较多词语实体的一部分，而含较多词语的实体通常能够更精确地指示句子中提到的副作用，因此我们只保留包含词语较多的副作用实体而舍弃另一个实体，从而可以在几乎不损失召回率的同时提高实体识别的准确率。

4.3 实验结果与分析

4.3.1 数据集

本章实验使用的数据与第 3 章工作中的相同，语料是从 DailyStrength 网站上抓取的 SIDER 数据库中存在记录的 870 种药物在 2013 年 3 月 24 日之前的用户评论（英文），并做了相同的预处理。

与第 3 章工作中的做法相同，我们将句子中的实体分成两类，第一类是药物的副作用，第二类是药物治疗的病症或是从评论中无法判定为药物副作用的其他情况。副作用实体识别实验的训练集共有 1,500 个已标注的含有实体（第一或第二类）的句子，它们是从所有句子中随机抽取并保留包含实体的句子而得到的；测试集共有 500 个已标注的句子，这些句子中有的包含实体，有的则不包含，由于是从所有句子中随机抽取的，故其分布情况与整体语料相同。

4.3.2 基于词袋匹配的识别

对于滑动窗口的大小 N_{SW}，若设置得太小，则无法识别一些较长的副作用名称；若设置过大，则可能引入一些由相距较远词语构成的错误的副作用名称。为了获得最佳的窗口大小，本文在不同的 N_{SW} 下进行实验，测试实体识别的效果。图 4.1 显示了在不同的滑动窗口大小下实体识别的准确率、召回率与 F 值。其中，对于由滑动窗口内的单词构成的词袋与词典中的实体对应的词袋之间的编辑距离总和应小于的阈值 TH_{ED}，本文通过试验取经验值 2。

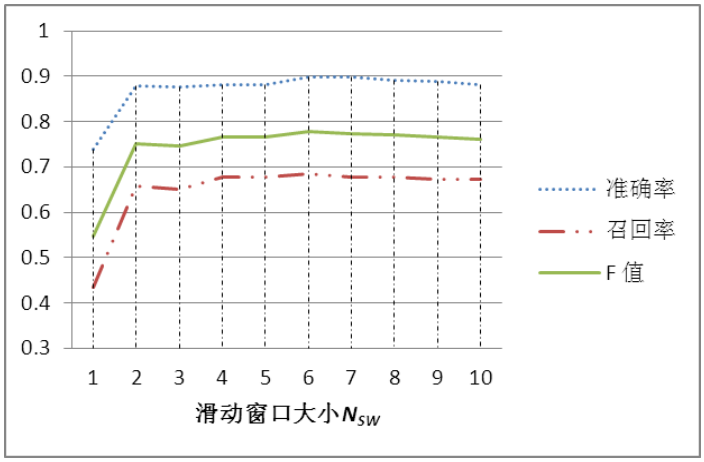

图 4.1 不同窗口大小 N_{SW} 时词袋匹配的实体识别效果

Fig. 4.1 Entity recognition results of word bag matching with different size N_{SW} of the sliding window

从图中可以看出，当滑动窗口大小 N_{SW} 由小变大时，实体识别的准确率、召回率以及 F 值呈现出先上升后逐渐达到稳定的趋势。当滑动窗口大小超过 8 以后，识别性能开始出现略微的降低。当 $N_{SW}=6$ 时识别的效果最好，此时的准确率、召回率、F 值分别达到最高的 89.9%、68.5%、77.8%。在后面的实验中，本文取滑动窗口大小为 6。

分析滑动窗口中词袋匹配识别实体的结果，我们发现制约识别效果的主要原因在于所用副作用词典的完备性有限，未收录评论中所有的副作用名称，如测试集中未识别出的 memory loss、puffy face、sick stomach 等。

4.3.3 基于条件随机场的识别

为了确定最佳的上下文范围及特征组合的方式，本文先不对词语的内部特征进行选择，即让每个词语拥有表 4.1 所示的 5 个内部特征。我们对上下文窗口大小从 1 至 5 变化时内部特征的各种组合方式进行了试验，最后得到当上下文为“左边第 1 个词语 + 当前词语 + 右边第 1 个词语”、特征为这三个词语的 1-gram 特征时，CRF 模型的识别效果最好，其准确率、召回率、F 值分别达到了 89.1%、74.1%、80.9%。在后面的实验中，本文将此组合作为 CRF 模型的特征组合方式。

在确定上下文范围及特征组合之后，我们测试了各种单一内部特征用于 CRF 识别的效果，实验结果如图 4.2 所示。从综合评价指标 F 值来看，单一内部特征 F1 即单词拼写特征用于 CRF 识别的效果最好，其 F 值达到了 70.0%，比其他单一内部特征要好很多；其次为特征 F4 即该单词在副作用词典中作为实体首词语的频率，其 F 值为 51.5%，说明实体首单词的相关信息对副作用实体识别比较重要。

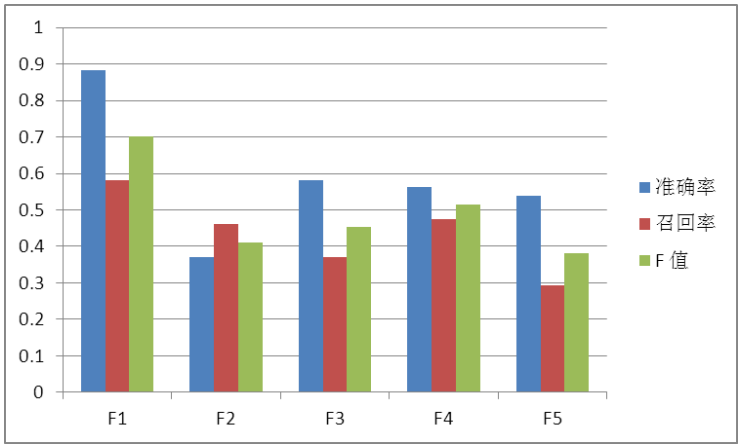

图 4.2 单一内部特征的 CRF 识别效果

Fig. 4.2 CRF recognition results with single internal feature

根据向前特征选择法，每次选择欲添加的特征时，我们检测所有剩余的候选特征加入后 CRF 模型识别性能的变化情况，找到使性能提升最大的特征并将其添加进来。当剩余的候选内部特征不能使识别性能得到提升时选择停止，此时已添加的特征构成了 CRF 识别实体所需要的最佳内部特征集合。向前选择内部特征的实验结果如表 4.5 所示。从表中数据看出，内部特征 F1、F4、F2、F5 的依次加入使得 CRF 模型识别实体的性能逐渐升高，特别是在模型中只有特征 F1 时，加入特征 F4 后识别的 F 值提高了 10.2%。特征选择完成后，最终 CRF 模型识别副作用实体的准确率、召回率、F 值分别达到了最高的 90.0%、75.5%、82.1%。在特征选择过程中，内部特征 F3 即词语在副作用词典中的频率对 CRF 识别实体的性能并无提升作用，因此我们将其从 CRF 模型所采用的内部特征集合中剔除。

表 4.5　向前特征选择过程中 CRF 的识别效果

Tab. 4.5　CRF recognition results during forward selection of features

实验序号	叠加的特征	准确率	召回率	F 值
1	F1	88.3%	58.0%	70.0%
2	F1+F4	90.4%	72.0%	80.2%
3	F1+F4+F2	89.1%	74.1%	80.9%
4	F1+F4+F2+F5	90.0%	75.5%	82.1%

根据实验结果可知，当 CRF 模型的特征设置为当前词语、左边第一个词语、右边第一个词语的 1-gram 特征，且每个词语的内部特征为“F1（拼写）+ F4（在副作用词典中作为实体首词语的频率）+ F2（词性）+ F5（在副作用词典中作为实体尾词语的频率）”时，CFR 识别副作用实体的效果最好，因此在最终的融合中， CRF 模型的特征为此特征组合。

本文第 3 章工作中同样利用 CRF 模型识别副作用实体，在与本实验相同的数据上进行训练与测试，得到实体识别的准确率为 87.5%，召回率为 58.7%，F 值为 70.3%。与之前的识别效果相比，本章工作中 CRF 的识别性能有了明显的提升。对比二者之间的不同，主要在于本章工作的 CRF 模型采用了更加丰富的词语内部特征，并对词语的上下文特征范围与特征组合方式进行了选择，最终模型识别性能的提升说明这些做法确实产生了作用，合理的特征选择与参数设置能够使 CRF 模型的性能得到有效提高。

4.3.4　识别结果的融合

将基于词袋匹配识别的最佳结果与基于条件随机场识别的最佳结果进行融合，得到最终副作用实体识别的准确率为 90.5%，召回率为 79.7%，F 值为 84.8%。从 F 值上看，融合后的结果比基于词袋匹配的最佳结果提高了 7.0%，比基于条件随机场的最佳结果

提高了 2.7%，证明了将不同方法的识别结果进行融合的有效性，说明通过融合确实能够弥补单一方法在副作用实体识别时存在的不足，从而使实体识别的效果得到提升。

同样以 DailyStrength 上的用户评论作为实验语料，Leaman 等人[6] 在人工标注的数据集上识别的准确率为 78.3%，召回率为 69.9%，F 值为 73.9%；Nikfarjam 等人[10] 在人工标注的数据集上识别的准确率为 70.01%，召回率为 66.32%，F 值为 67.96%。由于文献[6] [10] 未公开所用的标注数据集，我们无法利用它们的数据集对本文方法进行测试；而因为不能获取文献[6][10] 方法中所用的一些辅助数据，我们也难以将它们的方法在本文数据集上进行重现。虽然无法将本文方法与文献[6][10] 方法的识别效果在更加合理的前提下进行比较，但从识别的准确率、召回率以及 F 值来看，本文方法的识别效果与文献[6][10] 的具有可比性，甚至可能优于它们的识别效果。

4.4 本章总结

由于医疗社交网站用户评论与药物副作用名称自身存在的特点，从评论中识别副作用实体具有较大的难度与挑战性，前人工作中识别副作用实体的性能还有较大的提升空间。针对用户评论中的副作用实体的识别，本文实现了一个不同方法识别结果相融合的副作用实体识别系统。

第一种方法基于词袋匹配进行识别，即把滑动窗口中的短语与副作用词典中的名称进行词袋匹配，根据匹配的结果识别实体。在词袋匹配的过程中，我们考虑了单词之间的编辑距离，这样可以识别出用户评论中存在拼写错误的单词构成的实体，从而使识别结果更加完备，提高实体识别的召回率。第二种方法利用条件随机场模型进行识别，其中采用了向前特征选择法选出用于副作用识别的最佳内部特征集合，并对不同上下文范围内的特征组合方式进行试验，找出识别效果最好的上下文窗口与特征组合方式。与本文第 3 章中利用条件随机场模型识别的结果相比，此处基于条件随机场的识别性能有了明显提升，说明采用更加丰富的词语内部特征，并对词语的上下文特征范围与特征组合方式进行选择的做法发挥了作用，对 CRF 模型进行合理的特征选择与参数设置能够使其性能得到提升。

本文将两种方法的识别结果进行有效地融合，得到的最终结果要好于两种方法各自的识别结果，说明通过融合确实可以弥补单一方法在实体识别中的不足，并使各自的优势得到互补，从而提高副作用实体识别的性能。与文献[6][10] 中副作用实体识别方法相比，本文方法的识别效果不逊色甚至可能优于他们。

结　论

随着生活水平的提高，人们越来越重视医疗健康问题，因此也更加关注药物的安全性。而市场上药物种类的极大丰富给药物的安全管理带来了巨大的挑战，药物的副作用问题变得越来越突出，如何发现药物的副作用具有重大的理论与实用价值。人们在面对诸多药物进行选择也非易事，不知哪种药物适合自己而没有副作用。另一方面，随着 Web 2.0 技术的发展，互联网上出现不少的医疗健康类社交网站，人们在这里分享用药经历，交流用药体验。互联网的普及使得越来越多的人们倾向于从网上搜索感兴趣药物的相关信息，辅助自己进行药物的选择，同时也乐于将自己的用药体验与对药物的评论分享给他人作为参考。这些医疗健康类的社交网站在人们安全用药方面带来很大的帮助，扮演的角色也越来越重要。

而随着人们越来越多地参与到医疗健康类社交网站上用药问题的分享与交流中，这些网站中的相关数据也变得越发地丰富，其中蕴含的药物相关知识开始受到研究人员的关注，他们借助生物信息学的相关技术与手段来从大量数据中进行药物安全性的评估，并逐渐形成从这些用户评论中进行文本挖掘进而找出药物副作用这样一种快捷、有效的药物副作用发现机制。

由于人们可能采用不同的表述方式来描述副作用，而新药的上市与用药者的差异性会造成新的副作用出现，因此在挖掘药物副作用时，从用户评论中识别新的副作用名称并进行标准化十分重要。在本文第 3 章中，我们利用条件随机场模型识别评论中的副作用，在识别得到的所有副作用名称中，数据库未收录的新名称占到 18.8%，说明我们采用的方法确实能够有效识别出评论中的新副作用名称。在识别出副作用名称后，我们利用本文提出的标准化方法对它们进行标准化，将其映射到已有的标准概念上。通过标准化，我们将用户对同一概念的不同表述形式映射到其所指的概念上，实现副作用名称的有效聚合与归并，使副作用概念在评论中的发生频率更接近其在用药者中发生的真实概率，从而有利于药物潜在副作用的发现。将挖掘出的药物已知的副作用与数据库记录进行对比，发现两种来源的药物副作用发生频率具有一定的相似性与对应性，从而说明本文的药物副作用挖掘方法的有效性。最后，对于发现的数据库中未记录的药物副作用，我们按挖掘到的发生频率由高到低排序，得到了一个可能性由大到小排列的药物潜在副作用列表。

从用户评论中识别副作用名称是药物副作用发现中基础却关键的步骤，副作用实体的识别效果对药物副作用的发现有直接而重要的影响。由于用户评论内容在语法上的不

规范性以及副作用名称的多样性，从评论中识别副作用实体具有较大的挑战性，前人方法识别副作用实体的性能还有较大的提升空间。针对该问题，本文第 4 章实现了一个融合不同方法识别结果的副作用实体识别系统。第一种方法基于词袋匹配进行识别，即把滑动窗口中的短语与副作用词典中的实体进行词袋匹配，根据匹配结果识别副作用实体，匹配过程中考虑了单词的编辑距离。第二种方法利用条件随机场模型进行识别。在识别过程中我们采用向前特征选择法选出用于副作用识别的最佳内部特征集合，并对不同上下文范围内的特征组合方式进行尝试，找出识别效果最好的上下文窗口与特征组合方式。与本文第 3 章中利用条件随机场模型识别的结果相比，这次的识别性能有了明显提升，说明采用更加丰富的词语内部特征，并对词语的上下文特征范围与特征组合方式进行选择的做法发挥了作用，对 CRF 模型进行合理的特征选择与参数设置能够使其性能得到提升。我们将两种方法的识别结果进行有效地融合，得到的融合后结果要好于两种方法各自的识别结果，说明通过融合确实可以弥补单一方法在实体识别中的不足，并使各自的优势得到互补，从而提高副作用实体识别的性能。与其他文献中的副作用实体识别方法相比，本文方法的识别效果不逊色甚至可能优于他们，证明了本文提出的融合方法的有效性。

参 考 文 献

[1] MEDLINE Fact Sheet[OL]. 2014-5-6. http://www.nlm.nih.gov/pubs/factsheets/medline.html.

[2] Campillos M, Kuhn M, Gavin A C, et al. Drug target identification using side-effect similarity[J]. Science, 2008, 321(5886): 263-266.

[3] Lounkine E, Keiser M J, Whitebread S, et al. Large-scale prediction and testing of drug activity on side-effect targets[J]. Nature, 2012, 486(7403): 361-367.

[4] Adverse drug reactions[M]. Pharmaceutical Press, 2006.

[5] Yang C C, Jiang L, Yang H, et al. Detecting Signals of Adverse Drug Reactions from Health Consumer Contributed Content in Social Media[C]//Proceedings of ACM SIGKDD Workshop on Health Informatics (August 12, 2012). 2012.

[6] Leaman R, Wojtulewicz L, Sullivan R, et al. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks[C]//Proceedings of the 2010 workshop on biomedical natural language processing. Association for Computational Linguistics, 2010: 117-125.

[7] Fox S, Duggan M. Health Online 2013[R/OL]. Washington, D.C.: Pew Research Center, 2013. http://www.pewinternet.org/2013/01/15/health-online-2013/.

[8] Chee B W, Berlin R, Schatz B. Predicting adverse drug events from personal health messages[C]//AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2011, 2011: 217-226.

[9] Yates A, Goharian N. ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites[M]//Advances in Information Retrieval. Springer Berlin Heidelberg, 2013: 816-819.

[10] Nikfarjam A, Gonzalez G H. Pattern mining for extraction of mentions of adverse drug reactions from user comments[C]//AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2011, 2011: 1019-1026.

[11] Zeng Q T, Tse T. Exploring and developing consumer health vocabularies[J]. Journal of the American Medical Informatics Association, 2006, 13(1): 24-29.

[12] Sampathkumar H, Luo B, Chen X. Mining Adverse Drug Side-Effects from Online Medical Forums[C]//Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on. IEEE, 2012: 150-150.

[13] Yates A, Goharian N, Frieder O. Extracting Adverse Drug Reactions from Forum Posts and Linking them to Drugs[C]//Proceedings of the 2013 ACM SIGIR Workshop on Health Search and Discovery. 2013.

[14] Yang M, Wang X, Kiang M. Identification of Consumer Adverse Drug Reaction Messages on Social Media[J]. 2013.

[15] Xu R, Wang Q Q. Toward creation of a cancer drug toxicity knowledge base: automatically extracting cancer drug-side effect relationships from the literature[J]. Journal of the American Medical Informatics Association, 2014, 21(1): 90-96.

[16] Wu H, Fang H, Stanhope S J. Exploiting online discussions to discover unrecognized drug side effects[J]. Methods Inf Med, 2013, 52(2): 152-159.

[17] Chinchor N, Robinson P. MUC-7 named entity task definition[C]//Proceedings of the 7th Conference on Message Understanding. 1997.

[18] Rau L F. Extracting company names from text[C]//Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on. IEEE, 1991, 1: 29-32.

[19] Black W J, Rinaldi F, Mowatt D. FACILE: Description of the NE System Used for MUC-7[C]//Proceedings of the 7th Message Understanding Conference. 1998.

[20] Grishman R. The NYU System for MUC-6 or Where's the Syntax?[C]//Proceedings of the 6th conference on Message understanding. Association for Computational Linguistics, 1995: 167-175.

[21] Mikheev A, Grover C, Moens M. Description of the LTG system used for MUC-7[C]//Proceedings of 7th Message Understanding Conference (MUC-7). Fairfax, VA, 1998.

[22] Krupka G R, Hausman K. IsoQuest Inc.: Description of the NetOwl (TM) Extractor System as Used for MUC-7[C]//Proceedings of MUC. 1998, 7.

[23] Farmakiotou D, Karkaletsis V, Koutsias J, et al. Rule-based named entity recognition for Greek financial texts[C]//Proc. of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000). 2000: 75-78.

[24] Hanisch D, Fundel K, Mevissen H T, et al. ProMiner: rule-based protein and gene entity recognition[J]. BMC bioinformatics, 2005, 6(Suppl 1): S14.

[25] Kazama J, Makino T, Ohta Y, et al. Tuning support vector machines for biomedical named entity recognition[C]//Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3. Association for Computational Linguistics, 2002: 1-8.

[26] Isozaki H, Kazawa H. Efficient support vector classifiers for named entity recognition[C]//Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002: 1-7.

[27] Ju Z, Wang J, Zhu F. Named entity recognition from biomedical text using SVM[C]//Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on. IEEE, 2011: 1-4.

[28] Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger[C]//proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002: 473-480.

[29] Zhao S. Named entity recognition in biomedical texts using an HMM model[C]//Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics, 2004: 84-87.

[30] Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets[C]//Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Association for Computational Linguistics, 2004: 104-107.

[31] Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition[C]//Pacific Symposium on Biocomputing. 2008, 13: 652-663.

[32] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]//Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003: 188-191.

[33] Borthwick A, Sterling J, Agichtein E, et al. NYU: Description of the MENE named entity system as used in MUC-7[C]//In Proceedings of the Seventh Message Understanding Conference (MUC-7. 1998.

[34] Chieu H L, Ng H T. Named entity recognition: a maximum entropy approach using global information[C]//Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002: 1-7.

[35] Bender O, Och F J, Ney H. Maximum entropy models for named entity recognition[C]//Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003: 148-151.

[36] Isozaki H. Japanese named entity recognition based on a simple rule generator and decision tree learning[C]//Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2001: 314-321.

[37] Lin Y F, Tsai T H, Chou W C, et al. A maximum entropy approach to biomedical named entity recognition[C]//BIOKDD. 2004: 56-61.

[38] Florian R, Ittycheriah A, Jing H, et al. Named entity recognition through classifier combination[C]//Proceedings of the seventh conference on Natural

language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003: 168-171.

[39] Liu X, Zhang S, Wei F, et al. Recognizing named entities in tweets[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 359-367.

[40] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.

[41] Kuhn M, Campillos M, Letunic I, et al. A side effect resource to capture phenotypic effects of drugs[J]. Molecular systems biology, 2010, 6(1).

[42] MedEffect Canada - Health Canada[OL]. 2014-5-6. http://www.hc-sc.gc.ca/dhp-mps/medeff/index-eng.php.

[43] Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART) Source Information[OL]. 2014-5-6. http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CST/.

[44] Brown E G, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA)[J]. Drug Safety, 1999, 20(2): 109-117.

[45] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology[J]. Nucleic acids research, 2004, 32(suppl 1): D267-D270.

[46] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[47] CRF++: Yet Another CRF toolkit[OL]. 2014-5-6. http://crfpp.sourceforge.net/.

[48] Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network[C]//Proceedings of HLT-NAACL 2003, 2003: 173-180.

[49] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. The Journal of Machine Learning Research, 2003, 3: 1157-1182.

[50] Vlachos A. Tackling the BioCreative2 gene mention task with conditional random fields and syntactic parsing[C]//Proceedings of the Second BioCreative Challenge Evaluation Workshop; 23 to 25 April 2007; Madrid, Spain. 2007: 85-87.

攻读硕士学位期间发表学术论文情况

1　Enhancing the Accuracy of Knowledge Discovery: A Supervised Learning Method. **Liangxi Cheng**, Hongfei Lin, Feng Zhou, and Zhihao Yang. 2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)（被推荐到 BMC Genomics 期刊），2013 年，2013: 620. 主办单位：IEEE Computer Society。EI 检索，本文 EI 检索号：14079547。

2　Integrating various resources for gene name normalization. Yuncui Hu, Yanpeng Li, Hongfei Lin, Zhihao Yang, **Liangxi Cheng**. PloS one, 2012, 7(9): e43558. 主办单位：Public Library of Science（PLOS）。SCI 检索期刊，本文 SCI 检索号：000308738500017。

3　基于评论挖掘的药物副作用发现机制. **程亮喜**，林鸿飞.（本硕士学位论文第三章）（在投）

4　基于方法融合的副作用实体识别. **程亮喜**，林鸿飞.（本硕士学位论文第四章）（在投）

致　　谢

三年的硕士研究生生活即将结束，感觉光阴荏苒，青春易逝。研一研二时的情景还依稀如昨，却马上就要告别研三，迎来毕业离校的一刻。回顾这三年的时光，感谢实验室这三年里对我的栽培与磨练，实验室的这一段生活给我留下了很多值得回忆的篇章。实验室打羽毛球的传统让我熟悉并喜欢上了这项体育活动，每年元旦的实验室新年晚会让我感受到了同门的活力与激情，还有那一年一次的徒步大会，不定期举行各种人文讲座等等，都让我感受到了实验室生活的多姿多彩。

首先感谢我的导师林鸿飞教授，在这三年里林老师在学术与生活上都给了我很大的帮助与启发。林老师为人和蔼、热情，不拘小节，对学生的态度包容，常常给学生便利。在学术上林老师思维活跃，经验丰富，在我们的论文工作上给予了关键的指导，并常常在我们感到束手无策时提供灵感与思路，协助我们一块解决问题。另外，林老师坚持体育锻炼，虽然工作繁忙却干练高效，保持着充沛的活力，这也给了我们启发，不能只顾学习或工作而忘了锻炼，毕竟身体是一切活动的基础。

感谢杨志豪老师这三年的指导与栽培。杨老师在学术上态度严谨、工作勤勉，在生活中待人真诚、和蔼，平易近人，是我们应该看齐的榜样。杨老师不但在学术上给我们指导，同时在生活理念与价值取向上也常常给我们阐述一些非常有意义的观点与建议，给予我们人生的启迪，让我们可以更清楚自己想要的生活，不要在纷繁的世界中迷失了自己。

感谢实验室其他所有的老师以及师兄师姐、师弟师妹，实验室全体成员的互帮互助、团结友爱让我感到实验室大家庭的温暖，整个实验室包容、和谐的环境让我可以愉快地生活，安心地进行学术研究；感谢其他所有的朋友以及与我有过交集的人，虽然在此无法一一道出你们的名字，但在这三年的生活中你们的陪伴让我成长，让我一路风风雨雨却不失精彩地过来。

最后感谢家人对我一直以来的关心与爱，你们的支持让我心中感受到坚定的力量，你们的付出让我可以从容淡定地面对与克服生活中的艰难与坎坷，并满怀信心与激情地去迎接明天。

大连理工大学学位论文版权使用授权书

本人完全了解学校有关学位论文知识产权的规定，在校攻读学位期间论文工作的知识产权属于大连理工大学，允许论文被查阅和借阅。学校有权保留论文并向国家有关部门或机构送交论文的复印件和电子版，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、或扫描等复制手段保存和汇编本学位论文。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年____月____日

导 师 签 名 ：_____ 日期：_____年____月____日