DOI：10.16667/j.issn.2095-1302.2022.07.040

一种软件需求依赖关系自动提取方法

关　慧，刘萍萍，盛靖媛
（沈阳化工大学，辽宁 沈阳 110000）

摘　要：软件需求依赖关系提取是众多需求工程问题中的关键问题之一。以往的研究表明，需求依赖提取是由需求工程师根据专家的判断进行的，这不仅需要专业的需求知识，而且会消耗大量的人力。随着深度学习技术的兴起，依赖关系提取的研究也从基于规则的方法向基于深度学习方法转变，一些研究人员开始使用深度学习技术来解决人工识别需求依赖关系所存在的问题。因此，基于深度学习技术，提出了一种系统的需求依赖提取方法。该方法通过构建实体识别模型 BiLSTM-CRF 和实体关系提取模型 Word2Vec-CNN 提取需求实体关系，然后根据需求实体关系提取需求语句的依赖关系；同时系统地评估了该方法，并与其他实验方法进行了比较。实验证明，该方法能够高效、高质量地识别需求依赖关系。

关键词：需求依赖；机器学习；深度学习；命名实体识别；实体关系提取；自然语言处理
中图分类号：TP311　　**文献标识码：**A　　**文章编号：**2095-1302（2022）07-0130-04

0　引　言

相关研究表明，大约 80% 的需求是相互依赖的忽略需求依赖，对项目的成功有不利的影响 [1-2]。近年来，许多研究探索了基于机器学习的需求依赖提取 [3-6]。Deshpande 等人 [4, 7] 使用 NLP 和 ML 方法提取依赖关系。Samer 等人 [6] 分析了小型工业数据集，并使用潜在语义分析提取了依赖类型。Priyadi 等人 [8] 提出了一种针对软件需求规格说明文档的需求依赖图建模方法，以及判断需求之间的相似性、精细化和约束关系的方法。Arora 等人 [9] 应用 NLP 自动识别需求语句的组成短语，计算短语之间的相似度得分，输出语法和语义相似度函数来判断需求之间的关系。

本文基于深度学习技术，提出了一种基于自然语言编写的需求文本的系统依赖关系提取方法。首先，构建需求命名实体识别模型，将最能表达需求语义的词提取为实体；其次，构建需求实体关系提取模型，提取需求实体之间的关系；然后根据需求实体关系提取需求依赖关系；最后利用多个数据集对该方法进行了评估。

1　实验方法

近年来，使用机器学习方法进行需求依赖关系提取时，通常是基于大量高质量的需求依赖关系标注文本进行的。这就导致在缺少标注文本时进行需求依赖关系提取的难度增加。因此本文提出了一个可以在缺少需求依赖关系文本的情况下进行需求依赖关系自动提取的方法，总体流程如图 1 所示。

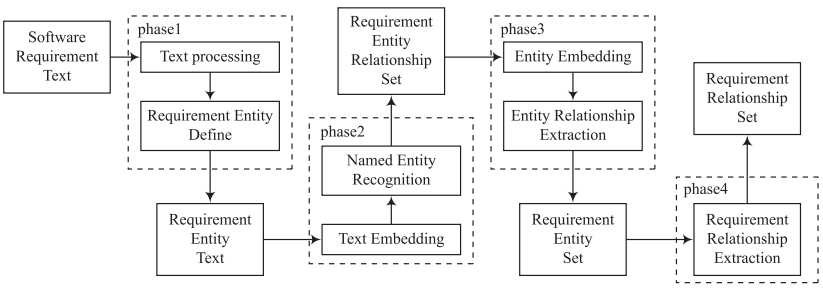

图 1　需求依赖实体关系的识别和提取过程

1.1　需求文本预处理

首先，对需求文本进行预处理。使用自然语言处理技术将需求文本中的语句按照“。”进行分割；然后，用特殊符号 <num> 替换句子的数字，并将句子中的所有字母转换为小写；最后，根据需求语句的句法结构和语义结构特点，选择句子中的主语、谓语和宾语作为目标实体类型。可以使用 BIO 编码方案为每个标记分配一个标识符来标识实体参数（开始位置和结束位置）及其类型，其中 B 表示实体的开始，

收稿日期：2021-09-27　　**修回日期：**2021-10-25

I 表示实体的内部，O 表示非实体。实体标记见表 1 所列。

表 1　命名实体标记

实体类型	实体起始	实体内部
主语	B-Sub	I-Sub
谓语	B-Pre	I-Pre
宾语	B-Obj	I-Obj
非实体	O	O

1.2　需求实体提取

BiLSTM-CRF[10] 模型是由双向长短期记忆网络（BiLSTM）和条件随机场（CRF）组成的命名实体识别模型。BiLSTM 能够有效表达输入向量特征在上下文中的意义，并预测相应的标签概率。CRF 层可以学习序列标签的约束，并通过传递特征来考虑输出标签之间的顺序，以保证预测结果的有效性。该方法使用 Glove 方法将需求文本进行向量表示，然后输入到 BiLSTM-CRF 模型中进行预测，形成需求实体集，用于后续的需求实体关系提取。需求实体处理过程如图 2 所示。

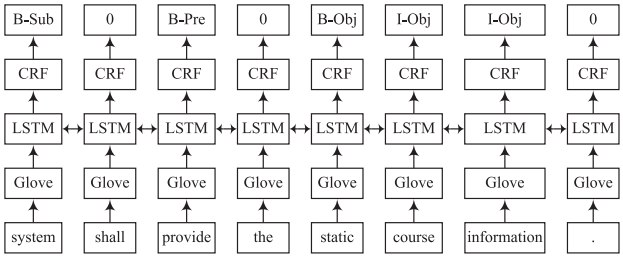

图 2　需求实体处理流程

经过需求实体提取后，形成了需求文本实体集 entity ={Subject, Predicate, Object}，Subject={S_1, S_2, …, S_n}，Predicate={P_1, P_2, …, P_n}，Object={O_1, O_2, …, O_n}，见表 2 所列。

表 2　需求实体

主　语	谓　语	宾　语
CMS 系统	提供信息	学生
CMS 系统	保存信息	学生
摄像系统	显示图片	用户
摄像系统	保存图片	用户
…	…	…

1.3　需求实体关系提取

1.3.1　需求依赖关系

需求之间存在各种类型的依赖关系，如需要、相似性、影响、冲突、业务相关性、演化、价值、成本、细化、包含、部分细化和不相关性 [9, 11-12]。目前基于机器学习的依赖提取的研究大多是研究需求的基本依赖关系，即需求是依赖关系还是独立关系。在此基础上，一些学者对需求语句的相似度、精细化、约束关系等进行了研究。本文自动提取了相似、细化、需要、调用和冲突五种需求关系。以上关系的非正式定义如下：

（1）相似：如果需求 R_1 和 R_2 需要完成相同的行动或目标，那么 R_1 和 R_2 存在相似关系。

（2）细化：如果需求 R_2 是对需求 R_1 的补充或是详细说明，那么 R_1 和 R_2 存在细化关系。

（3）需要：如果需求 R_2 需要在 R_1 实现的情况下才能实现，那么 R_1 和 R_2 存在需要关系。

（4）调用：如果需求 R_2 需要在 R_1 之后实现，那么 R_1 和 R_2 存在调用关系。

（5）冲突：如果需求 R_1 和 R_2 不能同时实现，则 R_1 和 R_2 有冲突关系。

1.3.2　需求实体关系

需求依赖可以由谓语触发，因此谓语可以很好地表达需求语句之间的依赖关系。但是仅仅依靠谓语关系来判断需求相关性，将会得到许多错误的结果，从而影响研究。因此，本文不仅考虑谓词实体之间的语义关系，还考虑主语、宾语实体之间的语义关系。通过判断需求语句中的主 - 谓 - 宾实体关系，可以识别出需求依赖关系。例如，如果两个需求语句中的两个谓语实体存在相似关系，并且主语实体和宾语实体也具有相似关系，则可以将需求语句判定为相似关系。需求实体存在的关系见表 3 所列。

表 3　需求实体关系

关　系	实体 1	实体 2
相似	主语	主语
相似	主语	宾语
相似	谓语	谓语
相似	主语	宾语
细化	主语	宾语
细化	宾语	宾语
细化	谓语	谓语
细化	主语	宾语
需要	谓语	谓语
调用	谓语	谓语
冲突	谓语	谓语

1.3.3　需求实体关系提取

Word2Vec-CNN 模型由分布式词向量表示（Word2Vec）和卷积神经网络（CNN）组成。模型的输入是实体集，输出是两个实体之间的对应关系。对于输入实体集，Word2Vec 可以用上下文语义的向量表示实体集，提高了 CNN 模型的

泛化能力。CNN 模型可以通过学习数据集中的规则进行训练。然后，对实体进行了对应的预测。本文将需求实体集输入 Word2Vec 模型进行向量表示。最后，将向量输入到 CNN 模型中，输出实体关系。该模型处理的需求实体集可以形成需求实体关系集，关系集可用于后续提取需求实体关系。需求实体关系提取过程如图 3 所示。

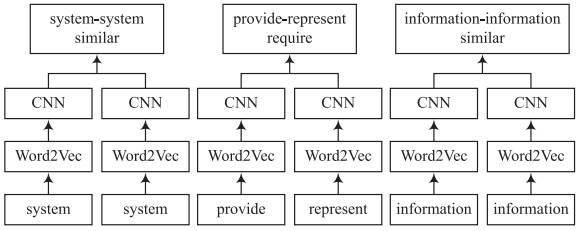

图 3　实体关系提取过程

1.4　需求依赖关系提取

如果两个需求之间存在依赖关系，那么两个需求的谓语之间必须存在某种关系。但是，如果这两个需求之间的关系直接由谓语的关系决定，就会出现一定的语义缺失，导致判断结果不准确。因此，本文通过获取两个需求语句之间的主谓宾实体词之间的关系以及实体之间的关系判断需求依赖关系。关系判断方法如下：

（1）判断谓语实体之间是否存在已定义的关系。如果谓语之间不存在一定的关系，那么需求语句之间不存在依赖关系。

（2）如果谓语之间存在定义的关系，则接下来判断主语之间是否有一定的关系。如果主语之间没有一定的关系，那么辨别主语与宾语之间是否存在一定的关系。如果有，则此需求具有依赖关系；如果没有，则需求之间不存在依赖关系。

（3）如果主语之间存在一定的关系，那么辨别宾语之间是否存在一定的关系。如果没有关系，那么需求是独立的。

（4）在谓语、主语均存在一定关系的情况下如果宾语之间存在一定的关系，则需求之间存在依赖关系。

2　实验和讨论

2.1　数据集

本实验使用的样本数据为德保罗大学 MSC 学生作为学期项目开发的 PROMISE[13] 和来自公共数据集的 PURE[14]，并使用文献 [5] 和文献 [15] 中的需求文本进行对比验证，以证明所提方法的有效性。

2.2　实验

为了证明该方法的可行性，本文采用的评价标准为：准确率 *P*、召回率 *R*、F1-score。

2.2.1　需求实体识别

利用 BiLSTM-CRF 模型进行实验，将需求文本标记为主语、谓语和宾语三种实体类型，然后将标记的数据集输入模型进行训练。实验中使用的 BiLSTM 模型是一个 size 为 100、dropout rate 为 0.5 的单层 BiLSTM。将 dropout rate 设置为 0.5 的原因是 dropout rate 越高，准确率越低；dropout rate 越低，时间效率越低。实验结果见表 4 所列。

表 4　需求实体识别结果

实　体	*P*	*R*	F1
主语	0.86	0.86	0.86
谓语	0.72	0.68	0.70
宾语	0.87	0.87	0.87
平均	0.82	0.80	0.81

从表 4 可以看出，BiLSTM-CRF 模型的平均准确率和召回率都在 80% 以上，证明该方法能够有效区分需求文本中的主谓宾实体。但是，与其他主语实体和宾语实体相比，谓词实体的正确率和召回率都相对较低。这是因为大多数谓语实体是由动词构成的，而动词的时态、人称和上下文会影响其语义，从而对谓语实体的提取产生不利影响。

2.2.2　需求实体关系提取

使用 Word2Vec-CNN 模型进行实验。实验基于需求实体集定义不同的实体关系，并构建实体关系数据集；然后将需求数据输入 Word2Vec-CNN 模型进行训练。在实验中，选择“size=2，step=1”滑动窗口生成句子，即每个句子包含要求文本中的两个句子。将每个句子中出现的实体排列组合为候选实体对，然后对每个样本进行向量化，提取 5 个向量作为模型的输入。实验结果见表 5 所列。

表 5　需求实体关系提取结果

数据集	*P*	*R*	F1
PROMISE[13]	0.91	0.88	0.90
PURE[14]	0.89	0.83	0.86
AVG	0.90	0.86	0.88

从表 5 可以看出，该模型的平均准确率和召回率都在 85% 以上，在 PROMISE[13] 数据集上准确率甚至达到 91%，证明了该模型在提取需求依赖方面的有效性。

为了验证模型的有效性，使用文献 [5] 和文献 [15] 中的数据集进行比较。Deshpande 等人 [15] 利用弱监督学习对未标记数据生成伪标签，解决了所需文本缺乏标记数据集的问题，从而提高了机器学习的准确性，识别了相似、需要、或、异或等需求关系。通过本文方法对文献 [15] 中使用的数据集进行实体识别和关系提取，并根据依赖类型进行依赖关系提取。实验结果见表 6 所列。

表 6　本文方法与文献 [15] 的对比结果

方　法	*P*	*R*	F1
Word2Vec+CNN	0.90	0.84	0.87
WSI+RF[15]	0.85	0.81	0.81
WSI+NB[15]	0.83	0.81	0.81
WSI+SVM[15]	0.88	0.87	0.87

从表 6 可知，平均准确率和召回率均在 85% 以上，说明该方法是有效的。虽然准确率较高，但召回率普遍较低，这说明目前的方法在处理需求依赖方面存在一定的缺陷。主要原因是，在提取需求依赖实体时，Glove 的向量表示方法对一些不熟悉的需求词无效，导致提取实体不准确，出现一些错误。

Atas 等人 [5] 使用 TF-IDF 方法利用 n-gram、POS-tag 和“require”作为三个向量特征；通过网格搜索为每个分类器识别出最适合的特征组合；将特征组合输入到分类器（包括朴素贝叶斯、线性支持向量机、k 近邻和随机森林），以自动识别需求依赖中的“需求”关系。另外，通过本文方法对文献 [16] 中使用的数据集进行实体识别和关系提取。实验结果见表 7 所列。

表 7　本文方法与文献 [5] 的比较结果

方　法	*P*	*R*	F1
Word2Vec+CNN	0.91	0.86	0.88
朴素贝叶斯 [5]	0.82	0.77	0.77
线性 SVM[5]	0.84	0.79	0.79
k 近邻 [5]	0.75	0.70	0.70
随机森林 [5]	0.85	0.81	0.82

从表 7 可知，准确率和召回率均在 85% 以上，说明该方法是有效的。虽然具有较高的准确性，但实验中的召回率较低，主要是因为在实验中使用的文本是从德国英语翻译而来，有一些语义和句子不准确，导致了一些语义错误产生。

3　结　语

本文采用实体识别和实体关系提取的方法，实现了需求依赖关系的自动提取。首先，利用 BiLSTM-CRF 模型识别需求文本的实体；然后，将识别出的需求实体输入到 CNN 模型中进行关系识别，提取需求依赖关系；最后，通过实验系统验证了该方法的有效性。本文提出的方法具有以下优点：（1）该方法在一定程度上解决了需求依赖文本缺乏标注而难以自动提取需求依赖关系的问题；（2）基于命名实体识别方法，解决了自然语言文本的模糊性问题；（3）建立需求实体关系，可以根据需求定义不同的关系，提取不同的需求依赖关系。

本文的方法虽然在一定程度上实现了需求依赖关系的自动提取，但仍存在一些不足。需求实体和实体关系的构建需要一定的人工成本。实验中使用的数据集越多，所需的人力资源就越多。今后将在这方面展开研究，改进关系提取模型，进一步提高实验的准确性和效率。

参　考　文　献

[1] MARTAKIS A，DANEVA M. Handling requirements dependencies in agile projects：a focus group with agile software development practitioners [C]// Proceedings of the 7th International Conference on Research Challenges in Information Science. Paris，France：IEEE，2013.

[2] CARLSHAMRE P，SANDAH K，CARLSHAMRE A F，et al. An industrial survey of requirements interdependencies in software product release planning [C]// Proceedings of Fifth IEEE International Symposium on Requirements Engineering. [S.l.]：IEEE，2001.

[3] GUO J，CHENG J，CLELAND-HUANG J. Semantically enhanced software traceability using deep learning techniques[C]// Proceedings of 2017 IEEE/ACM 39th International Conference on Software Engineering. Buenos Aires，Argentina：IEEE，2017.

[4] DESHPANDE G，ARORA C，RUHE G. Data-driven elicitation and optimization of dependencies between requirements [C]// Proceedings of 2019 IEEE 27th International Requirements Engineering Conference（RE）. Jeju，Korea（South）：IEEE，2019.

[5] ATAS M，SAMER R，FELFERNIG A. Automated identifification of type specifific dependencies between requirements [C]// Proceedings of 2018 IEEE/WIC/ACM International Conference on Web Intelligence（WI）. Santiago，Chile：IEEE，2018.

[6] SAMER R，STETTINGER M，ATAS M，et al. New approaches to the identification of dependencies between requirements [C]// Proceedings of 31st International Conference on Tools with Artifificial Intelligence. Portland，USA：IEEE，2019.

[7] DESHPANDE G. Sreyantra：automated software requirement inter-dependencies elicitation，analysis and learning [C]// 2019 IEEE/ACM 41st International Conference on Software Engineering：Companion Proceedings. Montreal，QC：IEEE，2019.

[8] PRIYADI Y，DJUNAIDY A，SIAHAAN D. Requirements dependency gragh modeling on software requirements specification using text analysis [C]// Proceedings of 2019 1st International Conference on Cybernetics and Intelligent System. Denpasar，Indonesia：IEEE，2019.

[9] ARORA C，SABETZADEH M，GOKNIL A，et al. Change impact analysis for natural language requirements：an NLP approach [C]// Proceedings of International Conference on Requirements Engineering. Ottawa，ON，Canada：IEEE，2015.

[10] 张晨荣 . 基于联合学习的知识库问答研究 [D]. 包头：内蒙古科技大学，2020.

[11] 温赵欣，关慧，贾成真 . 基于需求依赖关系识别横切关注点 [J]. 电子技术与软件工程，2018，7（20）：64-67.

[12] 邵飞 . 基于依赖关系的软件需求建模与优先级评估方法研究 [D]. 武汉：武汉大学，2017.

[13] SAYYAD S S，MENZIES J T. The PROMISE repository of software engineering databases[EB/OL]. [2022-07-20]. http：//promise.site.uottawa.ca/SERepository.

（下转第 138 页）

安全态势评估模型的条件下将本实验结果和文献 [11] 进行比对，把文献 [11] 实验结果化为 1 ～ 12 区间内值，详情如图 6 所示。

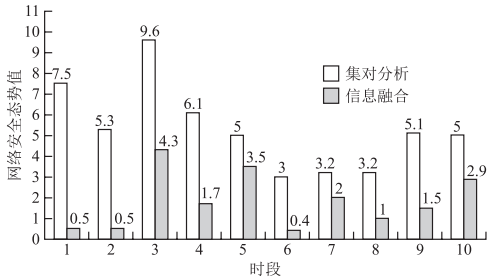

图 6　两种态势评估模型得出的网络安全态势值比较

由图 6 可知，集对分析模型具有一定的优势，文献 [11] 的评估模型获得的网络安全态势值较低，不能引起管理员重视；在时段 3、时段 9 出现攻击时，集对分析方法获得的安全态势值显著升高，能很好地区分 DDoS 等危险性较大攻击和 IP 扫描等危险性较小攻击；在时段 5，对于原始数据未受严重攻击的情况，文献 [11] 的网络安全态势值却显著升高，这表明集对分析模型感知网络安全态势更稳定。

4　结　语

本次研究是在多源数据融合的基础上展开的，由于物联网安全态势评估模型具有较强的功效而对其进行了构建，并得出以下结论：

（1）以多传感器数据融合技术为基础对网络安全态势要素进行划分，可以分为攻击要素、主机要素、共有要素三种。

（2）通过构建网络安全态势评估模型，检测网络各类型的安全信息。此操作是通过传感器来完成的，由此把主机安全态势确定下来。以主机权重为依据确定网络整体安全态势，此过程复杂程度较高，主要由数据采集、态势要素提取、主机安全态势评估、网络安全态势评估四种模块组成。

（3）集对分析模型具在自身特有的优势。在时段 3、时段 9 出现攻击时，集对分析方法获得的安全态势值显著升高，能很好地区分 DDOS 等危险性较大攻击和 IP 扫描等危险性较小的攻击。通过集对分析模型进行研究，可以更加准确地感知网络安全态势。

参　考　文　献

[1] WANG H，CHEN Z F，FENG X，et al. Research on network security situation assessment and quantification method based on analytic hierarchy process [J]. Wireless personal communications，2018，102（2）：1401-1420.

[2] 贾焰，韩伟红，杨行 . 网络安全态势感知研究现状与发展趋势 [J]. 广州大学学报（自然科学版），2019，18（3）：1-10.

[3] 常利伟，田晓雄，张宇青，等 . 基于多源异构数据融合的网络安全态势评估体系 [J]. 智能系统学报，2021，16（1）：38-48.

[4] 文志诚，陈志刚，唐军 . 基于信息融合的网络安全态势量化评估方法 [J]. 北京航空航天大学学报，2016，42（8）：1593-1603.

[5] 张雅琼，张慧，郑欢欢 . 一种物联网感知数据安全传输方案 [J]. 榆林学院学报，2021，31（4）：48-52.

[6] XI R R，YUN X C，HAO Z Y. Framework for risk assessment in cyber situational awareness [J]. IET information security，2019，13（2）：149-156.

[7] 席荣荣，云晓春，张永铮，等 . 一种改进的网络安全态势量化评估方法 [J]. 计算机学报，2015，38（4）：749-758.

[8] 刘效武，王慧强，吕宏武，等 . 网络安全态势认知融合感控模型 [J]. 软件学报，2016，27（8）：2099-2114.

[9] MOUSTAFA N，SLAY J，CREECH G. Novel geometric area analysis technique for anomaly detection using trapezoidal area estimation on large-scale networks [J]. IEEE transactions on big data，2019，5（4）：481-494.

[10] 陈维鹏，敖志刚，郭杰，等 . 基于改进的 BP 神经网络的网络空间态势感知系统安全评估 [J]. 计算机科学，2018，45（11A）：345-347.

[11] 韦勇，连一峰，冯国登 . 基于信息融合的网络安全态势评估模型 [J]. 计算机研究与发展，2009，46（3）：353-362.

作者简介：白　冰（1985—），男，陕西西安人，硕士，陕西警官职业学院讲师，主要研究方向为网络安全、信息安全、虚拟现实技术。

（上接第133页）

[14] FERRARI A，SPAGNOLO G O，GNESI S. PURE：a dataset of public requirements documents [C]// Proceedings of Requirements Engineering Conference. Lisbon，Portugal：IEEE，2017：502-505.

[15] DESHPANDE G，ARORA C，RUHE G. Data-Driven elicitation and optimization of dependencies between requirements [C]// Proceedings of 2019 IEEE 27th International Requirements Engineering Conference（RE）. Jeju，Korea（South）：IEEE，2019.