Operations Research and Fuzziology 运筹与模糊学, 2024, 14(2), 857-869
Published Online April 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.142186

Hans汉斯

面向企业模糊型技术需求的信息融合方法研究

陶泽奎[1]，张志清[1,2]

[1]武汉科技大学管理学院，湖北 武汉
[2]武汉科技大学服务科学与工程研究中心，湖北 武汉

收稿日期：2024年2月20日；录用日期：2024年3月11日；发布日期：2024年4月24日

摘 要

[目的/意义]：实现需求与供给的精准匹配是产教融合的关键所在，对科技成果转化、技术创新等具有重要意义。然而，企业在发布技术需求时，因商业机密保护等原因，导致技术需求文本描述常具有模糊性，难以与技术供给进行精准匹配。[方法/过程]：设计了面向企业模糊型技术需求的两阶段模型，首先使用BERT + BiLSTM + CRF算法提取企业发布项目中的需求实体，然后通过挖掘企业基本信息，结合需求实体作为检索条件爬取相关专利成果，通过TF-IDF算法获取专利中关键词作为补充信息，精准化技术需求描述，并识别出技术的聚焦方向与发展前沿。[结果/结论]：以机械制造领域内某企业真实需求作为案例，展示了该方法的完整处理过程，挖掘与TBM相关的施工管理、工件改进、系统研发等关键技术点；通过跨时间维度识别技术需求聚焦方向，TBM与人工智能、大数据、机器学习、探测等技术相结合将是未来发展的趋势。

关键词

技术需求，多源化信息，实体识别，BERT，关键词提取

Research on Information Fusion Method for Enterprise Fuzzy Technical Requirements

Zekui Tao[1], Zhiqing Zhang[1,2]

[1]School of Management, Wuhan University of Science and Technology, Wuhan Hubei
[2]Institute of Service Science and Engineering, Wuhan University of Science and Technology, Wuhan Hubei

Received: Feb. 20th, 2024; accepted: Mar. 11th, 2024; published: Apr. 24th, 2024

Abstract

[Purpose/Significance]: Achieving the accurate matching of demand and supply is the key to the

文章引用: 陶泽奎, 张志清. 面向企业模糊型技术需求的信息融合方法研究[J]. 运筹与模糊学, 2024, 14(2): 857-869.
DOI: 10.12677/orf.2024.142186

integration of industry and education, which is of great significance to the transformation of scientific and technological achievements and technological innovation. However, when enterprises release technical requirements, due to trade secret protection and other reasons, the text description of technical requirements is often ambiguous, and it is difficult to accurately match the technical supply. [Method/Process]: A two-stage model for the fuzzy technical needs of enterprises was designed. Firstly, the BERT + BiLSTM + CRF algorithm is used to extract the demand entities in the enterprise release projects, and then the relevant patent achievements are crawled by mining the basic information of the enterprise, combined with the demand entities as the search conditions, and the keywords in the patents are obtained as supplementary information through the TF-IDF algorithm, so as to accurately describe the technical requirements and identify the focus direction and development frontier of the technology. [Result/Conclusion]: Taking the real needs of an enterprise in the field of machinery manufacturing as a case, the complete processing process of the method is demonstrated, and the key technical points related to TBM such as construction management, workpiece improvement, and system research and development are excavated. By identifying the focus direction of technology demand across time dimensions, the combination of TBM with artificial intelligence, big data, machine learning, detection and other technologies would be the trend of future development.

Keywords

Technical Requirement, Multi-Source Information, Entity Recognition, BERT, Keyword Extraction

1. 引言

产教融合有助于推动企业技术进步和产业转型升级，提高企业竞争力。但由于企业在提出技术需求时并未形成统一的需求文本发布规则，导致企业技术需求的多样化特点，有清晰性、模糊型、概念型以及无方向型等多种类型。对于清晰型需求，企业会对技术使用的材料、工艺以及需要达到的效果做出明确要求，技术供给方可以依据相应指标做出准确研判，资源匹配效率和质量较高。因技术保密等原因，模糊型技术需求也极其普遍，为技术资源的合理分配增加了难度。本文就是基于此背景，拟对模糊性技术需求进行研究，在基本需求文本挖掘与分析基础上，结合多源信息对技术需求进行补充和完善。

近年来有部分学者对企业技术需求的识别方法进行了研究。如史敏等[1]基于竞争情报方法开发了一套有效识别企业技术需求的方法；杨德林等[2]分析技术供需文本间的语言差异，并提出了一种供需自动匹配方案。也有学者对技术需求分析相关的自然语言处理[3] [4] [5]、数据挖掘[6] [7] [8] [9]、统计机器学习[10] [11] [12] [13]等进行了研究。命名实体识别作为自然语言处理中的一项关键技术受到学者的关注。例如，张召武等[14]使用改进的向量表示层，使用 BiGRU 和 CRF 分别作为模型的序列建模层和标准层进行中文命名实体识别；Peng 等[15]提出了一种融合 LSTM 和 CRF 的混合模型训练命名实体识别和分词任务；Wu 等[16]提出 CNN-LSTM-CRF 模型用于中文命名实体识别。

以上学者的研究为产教融合背景下技术需求与供给的精准匹配提供了一定的技术与方法支持，但针对因文本模糊性导致的供需不匹配问题尚未得到有效解决。基于此，本文利用 BERT + BiLSTM + CRF 算法提取项目文本中的技术需求实体，在技术需求表述不完整的情况下，结合企业基本信息爬取相同领域的专利数据作为辅助数据，通过 TF-IDF 算法挖掘专利文本中关键词作为补充，丰富并完善技术的关键

信息，识别出企业的真实需求，以期为技术精准对接提供方法支持。

2. 框架模型设计

2.1. 研究框架

本文重点针对产教融合过程中因保密限制导致技术需求描述具有模糊性的问题展开研究，通过数据挖掘提取文本中关键信息，完成企业技术需求的精准分析与描述。由于文本内容短、信息含量少、表述口语化等特点，传统基于统计机器学习的关键词提取算法不适用于该类数据，且仅将单一项目文本作为数据源并不能有效识别出技术在供需双方的发展状况。因此，本文设计一种两阶段模型，如图 1 所示。第一阶段研究聚焦于从企业需求文本中提取技术实体。研究通过技术交易网站收集企业需求文本，采用 BERT + BiLSTM + CRF 模型提取文本中的关键实体，包括需求文本中提到的技术、产品等相关信息，以达到初步了解企业技术需求的目的。第二阶段进行多源化信息的完善。因为专利作为企业、高校、科研机构等主体发布技术创新成果的重要载体，能够获取到更多关键信息，可用于分析技术的演化与发展[17]，故这一过程将以专利成果作为数据源。此外，企业基本信息中包含所处行业、主营产品等信息，能够辅助完成企业技术需求的精准定位。基于此，将第一阶段中提取的关键实体与企业基本信息中关键词作为检索条件，爬取相关专利成果，通过 TF-IDF 算法获取专利成果中关键词，用于补充技术信息，有效完成

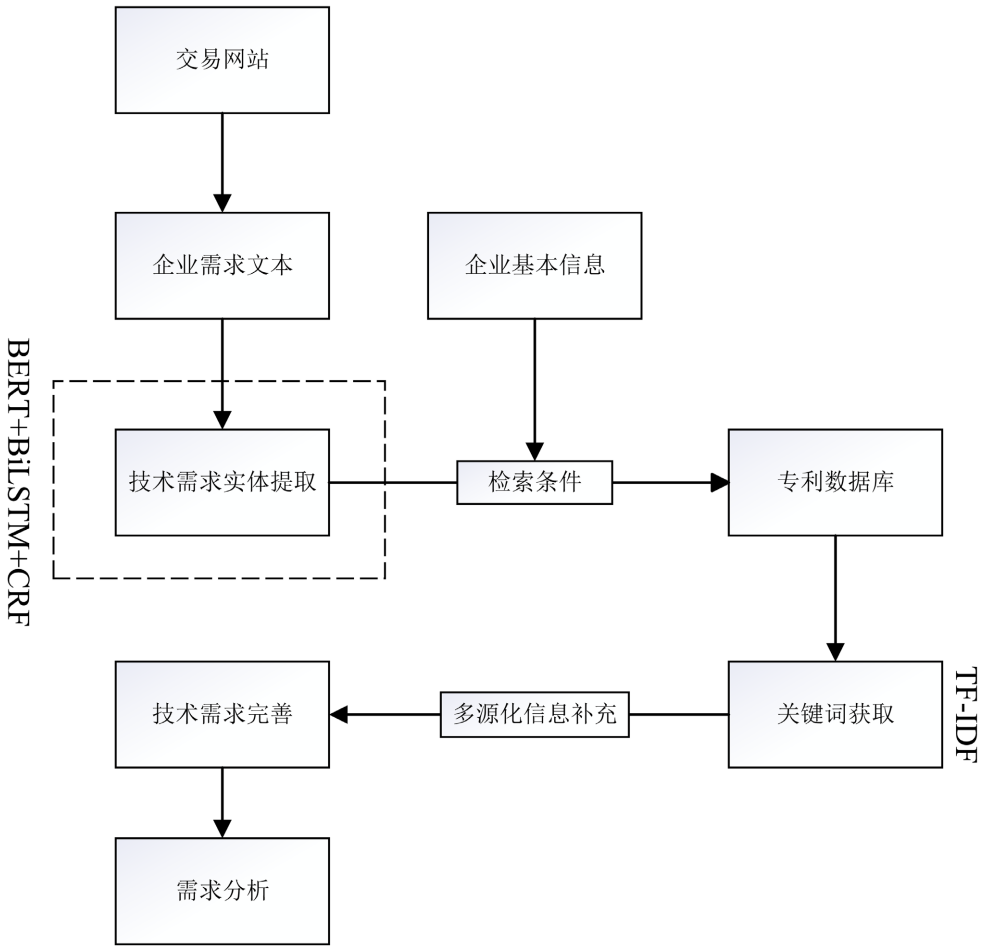

Figure 1. Research framework
图 1. 研究框架

技术需求分析。通过对不同年份的专利成果进行跨时间维度的关键词提取分析，可进一步挖掘技术的发展路径，识别出企业完成创新所需要的核心关键技术，为模糊型需求提供正确的技术指导与创新方向。

2.2. BERT + BiLSTM + CRF 模型

本文采用的 BERT + BiLSTM + CRF 模型如图 2 所示。该模型由三个模块组成。首先，使用 BERT 作为预训练模型，理解文本中的上下文信息，并对输入文本进行编码，将非结构化文本转换为向量，方便计算机理解文本信息；其次，利用 BiLSTM 模块进行建模以捕捉更加全面的上下文信息，增强模型的理解能力；最后，通过 CRF 模块进行标签解码，对未根据自然语言规则所输出的错误预测标签做出约束，得到最优的实体提取结果。

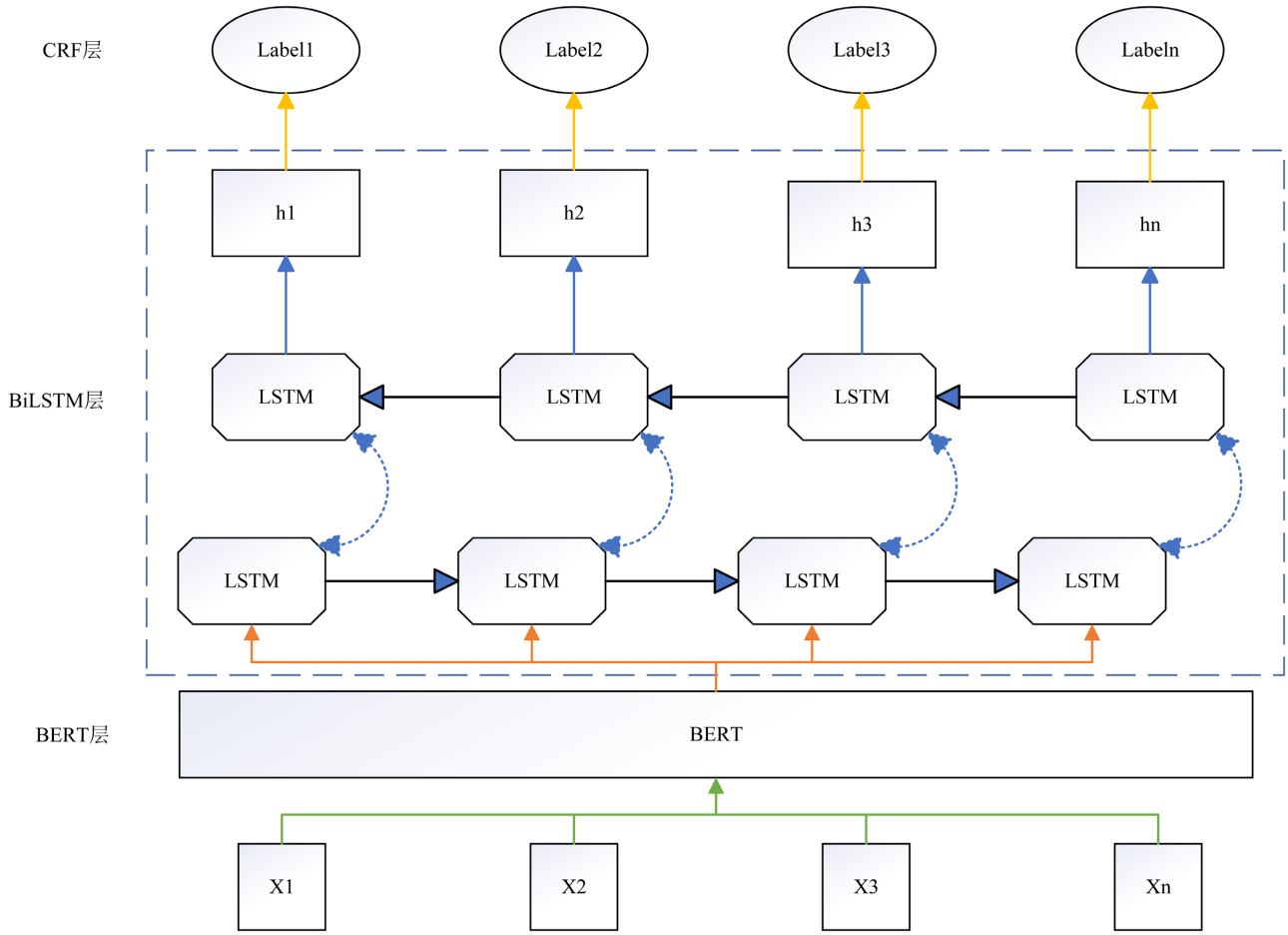

Figure 2. BERT + BiLSTM + CRF model
图 2. BERT + BiLSTM + CRF 模型

2.2.1. BERT 模块

在自然语言处理任务中，将语言信息处理为计算机能识别的知识一直是近些年来的研究热点，较为常见的语言模型有 one-hot、Word2Vec、GPT 等。然而上述模型在处理词嵌入任务时均存在着一些问题。例如，使用 one-hot 编码词向量时每个词被表示为一个高维向量，其中只有一个元素为 1，其余元素都为 0，这样会导致词向量发生维度爆炸且向量十分稀疏，另一方面如果遇到未登录词时它将无法获得有效的表示；Word2Vec 训练的词向量为静态向量，无法解决一词多义的问题；GPT 为单向语言模型，无法捕

获词蕴含的上下文信息。

2018 年，Devlin 等[18]提出了 BERT 预训练语言模型，能够有效处理上述模型中存在的问题，并在 11 项自然语言处理任务中证实了模型的优越性能。预训练阶段，BERT 通过掩码语言模型(Masked Language Model, MLM)和下句预测(Next Sentence Prediction, NSP)这两个任务共同完成。MLM 任务中，BERT 随机掩盖输入句子中的一些词汇，并尝试预测这些被掩盖的词汇，从而促使模型学会理解上下文的语义信息。NSP 任务中，BERT 接收一对句子作为输入，判断两个句子是否在原文中相邻，以此训练模型对句子级别的关联性进行建模。接下来便是对下游任务采用 Fine-Tuning 的微调处理模式，通过在任务特定数据上进行训练。BERT 可以根据具体任务进行调整，并获得出色的性能，其输入向量包含字向量(token embedding)、句向量(segment embedding)和位置向量，其中[CLS]标记表示句子的开始，[SEP]标记表示为句子间的间隔或者句尾，如图 3 所示。

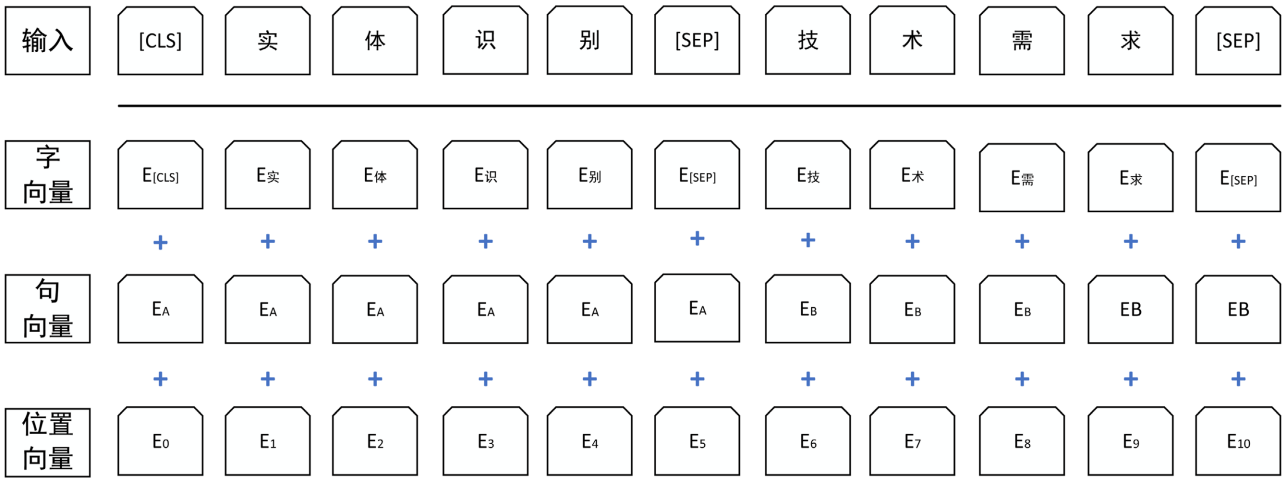

Figure 3. BERT input representation
图 3. BERT 输入表示

2.2.2. BiLSTM 模块

长短时记忆网络(Long Short Term Memory, LSTM)是循环神经网络(Rerrent Neural Network, RNN)的一种变体，更利于处理序列中的依赖关系。该网络结构包括输入门、遗忘门、输出门三个控制单元和一个记忆单元组成，输入门控制新输入对记忆单元的影响，遗忘门控制着记忆细胞中信息被保留的程度，记忆细胞决定当前时间步的 LSTM 单元的输出，记忆细胞通过输入门和遗忘门来更新和调整其内容，并通过输出门控制其输出。LSTM 模型的详细计算过程如式(1)~(5)所示。

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \tag{1}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \tag{2}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o\right) \tag{3}$$

$$C_t = f_t C_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{4}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{5}$$

公式(1)~(5)中，σ 为 sigmoid 激活函数；W 和 b 是模型学到的权重矩阵和偏置；i_t、f_t、o_t、C_t、h_t 分别是 t 时刻输入门、遗忘门、输出门、记忆细胞和隐藏门的输出。

然而，单向的 LSTM 不能同时处理上下文信息，无法挖掘出文本中每个词语与前后词之间的潜在联系，但是在自然语言中词与词之间往往存在着一定的关联性。为了解决这一类问题，Graves A 等[19]在语音处理领域首次对 LSTM 进行了改进，提出了双向长短时记忆网络(Bidirectional Long-Short Term Memory, BiLSTM)，其基本思想是通过引入反向的 LSTM，使网络能够同时考虑到输入序列的前向和后向信息，并将两个方向的信息进行合并输出。BiLTSM 的网络结构如图 4 所示。

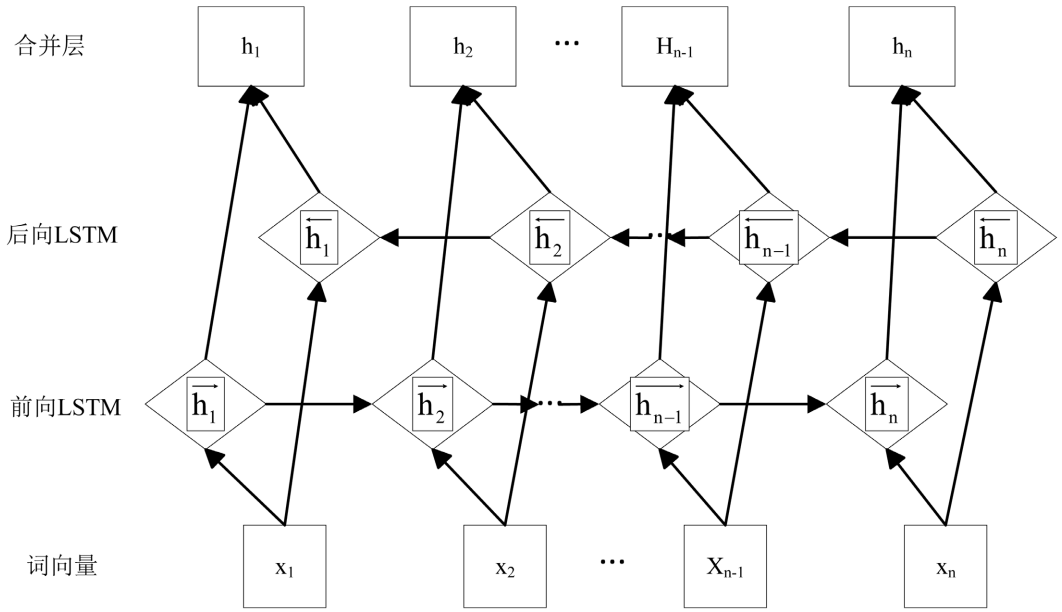

Figure 4. BiLSTM network structure
图 4. BiLSTM 网络结构

2.2.3. CRF 模块

CRF 模型以给定一组输入序列为条件，同时得到另一组输出序列的条件概率分布，它属于判别式概率无向图模型，主要用于自然语言处理中的序列标注。在自然语言中通常存在着某些规则，相邻的字或词之间具有一定的约束，比如在 BIO 标记中，I 标签只能位于 B 标签的前面，I-X 只能位于 B-X 之后而不能位于 B-Y 之后。CRF 模型的主要任务便是根据自然语言的规则对预测标签进行有效的约束，对标签序列进行建模，从而获得最优预测结果。

对于任意一个序列 $X=(x_1,x_2,\cdots,x_n)$，假定 P 是 BiLSTM 的输出得分矩阵，$P\in R^{n*k}$，其中 n 为词序列的长度，k 为标签数量，P_{ij} 为第 i 个词的第 j 个标签得分。对于预测序列 $Y=(y_1,y_2,\cdots,y_n)$ 而言，它的分数函数如式(6)所示：

$$s(X,Y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{6}$$

式(6)中 A 为转移分数矩阵，A_{ij} 表示标签 i 转移到标签 j 的分数，$A\in R^{(k+2)*(k+2)}$，对所有可能的序列路径进行归一化，产生关于输出序列 Y 的概率分布，如式(7)所示：

$$P(Y|X)=\frac{e^{s(X,Y)}}{\sum_{\tilde{Y}\in Y_X}s(X,\tilde{Y})} \tag{7}$$

两边同时取对数便可以得到预测序列的似然函数，如式(8)所示。

$$\log\left(p\left(Y|X\right)\right)=s\left(X,Y\right)-\ln\left(\sum_{\tilde{Y}\in Y_X}s\left(X,\tilde{Y}\right)\right) \tag{8}$$

式(7)、(8)中，$\tilde{Y}$ 表示真实的标注序列，Y_X 表示所有可能的标注序列。

最后经过解码输出最大分数的标注序列，如式(9)所示：

$$Y^*=\text{argmax}_{\tilde{Y}\in Y_X}s\left(X,\tilde{Y}\right) \tag{9}$$

3. 技术需求实体提取方法

3.1. 数据集

为了验证模型的有效性，本文采用爬虫技术从技 E 网(https://www.ctex.cn)、Innomatch (https://www.innomatch.net)、高校官网(https://www.wust.edu.cn)等技术交易平台爬取产学研项目，考虑到本文研究以提取企业技术需求为主，经过数据清洗与缺失值处理后，最终保留先进制造、能源环境、新材料、生物医药及电子信息等五个领域的 7258 份需求文本，具体内容如表 1 所示。

Table 1. Industry university research project requirement description text
表 1. 产学研项目需求描述文本

企业领域	项目名称	具体需求
能源环境	锂电池筛选过程中充电和分容的电力综合利用	锂电池组装前要进行分容，分容过程中锂电池中的电能通常就消耗掉，通过一个能量回收装置储存电能，可用于锂电池组装完成后进行充电，减少生产过程中的电能消耗。
先进制造	干湿两用吸尘器噪音过大难题	“干湿两用吸尘器”与“干式吸尘器”不同，电机噪音过大，通过风扇排水时噪声大，极度影响用户的使用体验。
新材料	硅气凝胶生产关键技术及应用	硅气凝胶粉体、硅气凝胶毡、硅气凝胶复合材料等新材料的研制
新材料	聚合物防水涂料制备方法研究	将聚合物乳液、水等进行混合，再和粉料混合，得到防水涂料
…	…	…

3.2. 实体标注

进行实体识别之前，对文本数据进行标注是最为关键的一步，标注结果将直接影响到识别效果。基于本文研究需要，定义技术需求、主体需求、主体成分需求三类实体并采用 BIO 标注方式，具体定义如表 2 所示。

Table 2. Explanation of the project requirement text entity
表 2. 项目需求文本实体解释

实体名称	实体定义	例句
技术需求	项目实施需要具备的技术	{激光测距}{传感}关键技术
主体需求	项目需要对其进行研究或开发的主体	{鬼臼毒素}的制备技术
主体成分需求	项目中对主体的零部件提出的具体需求	完成{管柄}和{毛头}的自动化装配。

本文采用精确率、召回率和 F1 值作为评估指标，如式(10)~(12)所示：

$$P=\frac{T_P}{T_P+F_P}\times 100\% \tag{10}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{11}$$

$$F1 = \frac{2PR}{P + R} \times 100\% \tag{12}$$

式(10)~(12)中，T_P 表示为正确将正样本预测为正；F_P 表示为错误将负样本预测为正；F_N 表示为错误将正样本预测为负。

3.3. 实验结果分析

本文实验基于 Pytorch1.10.0 框架进行搭建，Windows11 操作系统，Python 版本为 3.9.7，实验参数具体设置见表 3。

Table 3. Experimental parameter settings
表 3. 实验参数设置

参数	数值
词典大小	3000
字向量维度	768
句子最大长度	512
学习率	0.0001
迭代周期	100
隐层维数	256
Transformer 层	12

实验将 7258 条产学研项目文本按照 8:2 的比率分为训练集和测试集，得出模型实体识别的评估结果，并与其他主流深度学习模型进行对比，如表 4、表 5 所示。

Table 4. Result of the project requirement named entity recognition
表 4. 项目需求命名实体识别结果

实体名称	P	R	F1
主体需求	0.79	0.85	0.82
技术需求	0.83	0.73	0.77
主题成分需求	0.69	0.64	0.67

Table 5. Comparison of recognition effects of different models
表 5.不同模型识别效果对比

模型	P	R	F1
BiLSTM-CRF	0.692	0.706	0.699
BERT-LSTM-CRF	0.726	0.732	0.729
BiGRU-CRF	0.704	0.711	0.707
BERT-BiLSTM-CRF	**0.758**	**0.765**	**0.761**

实验结果表明，本文使用的 BERT-BiLSTM-CRF 模型在产学研项目数据集中实体识别的准确率、召回率和 F1 值都达到了最优，分别是 75.81%、76.54%和 76.17%。相比于传统的 BiLSTM-CRF 模型，在加入 BERT 后模型识别效果体现出明显优势，F1 值提升了 6.3%。与 BERT-LSTM-CRF 和 BiGRU-CRF 相比，本文模型 F1 值也分别提升了 3.24%和 5.43%，该模型通过引入 BERT 与 BiLSTM 更加全面地理解文本上下文信息，提高了语义特征提取能力，因此相比于其他模型实体识别效果较好，可以更好地运用于产学研项目命名实体识别。表 6 中展示为部分测试集数据的实体提取结果，从结果中可以看出本文模型能够将测试集中技术需求信息正确识别出来，证实了该模型对于产学研项目实体识别的有效性。

Table 6. Entity recognition results of some industry university research projects
表 6. 部分产学研项目实体识别结果

编号	内容	识别结果
1	如何利用{智能机械手}解决{无人化生产}.拟实现的主要技术目标：实现{无人化生产}	主体需求、技术需求
2	目前{锂电池}在低温下难以{储存}，特别是在北方寒冷气温下容易出现电芯损坏问题。	主体需求、技术需求
3	通过{生物科技}升级改造原材料，使其达到在湿度较大环境下，较长时间存放不发生霉变	技术需求
4	对基于粘结剂喷射技术的{三维成型}工艺进行研究	技术需求
5	制备{色谱纯化}应用技术。使用硅胶填料制备{色谱纯化}为核心的技术。	技术需求
6	电子车间生产线{自动化改造}技术方案	技术需求
7	打造休闲用品{自动化生产}线，建成{数字化控制}的{智能工厂}。	主体需求、技术需求
8	公司主要制造{汽车半轴}、{铸钢件}，在黑色金属材料{热加工}({铸造}、{锻造})、{热处理}等方面需要技术支持。	主体需求、技术需求
9	因公司发展需要，急需具备{3D 打印}技术的专家人才	技术需求
10	解决{甘草泄心汤}药效不稳定导致的血压上升影响	主体需求
…	…	…

4. 基于专利挖掘的需求信息完善

通过对技术需求进行实体提取可以发现，项目文本中含有较少的技术信息，并不能对某一项技术进行全面的分析与理解。以本文收集的企业需求数据集为例，随机挑选一份项目文本，利用第二节中所提模型进行实体识别，文本内容及识别结果如表 7 所示。TBM (Tunnel Boring Machine)中文全称为全断面隧道掘进机，是一种工厂化流水线隧道施工装备。经分析，该项目属于典型的模糊型项目，从需求内容中只得知企业需要具备 TBM、自动化、智能化等技术的人才，描述较为抽象，需要结合多源化信息对需求进行深挖。而专利知识作为技术创新的一大载体，文本中包含较多的技术信息，据中国知识产权局公开数据显示，我国专利申请量每年超百万且逐年增加，如此庞大的数据量为技术信息的补充与完善提供了可能。同时，通过访问专业技术经理人得知，所处行业、主营产品等基本信息能够从侧面反映出企业生产或服务过程中使用的技术，对需求信息的分析具有一定的参考价值。

4.1. 多源化数据获取

通过企查查、天眼查、企业官网等平台查询企业相关信息。经查询，凯**工有限公司所处行业为机

械制造，主营产品为机械设备。接下来，以智慧芽作为专利检索平台，检索与自动化、机械化、智能化、掘进机、TBM、机械制造、机械设备等词条相关的中国专利，检索时间为2010~2023年。经筛选与去重后，共检索相关专利315份，专利成果年度发布信息如表5所示。从表5中相关专利成果年度发布情况可以看出国内从2012年才开始进行机械制造领域TBM设备智能化的相关研究，在此之前TBM主要以“引进设备，自主施工”为主。2012年，铁建重工与中铁联合牵头，开展了国家高技术研究计划(863计划)“大直径TBM关键技术及应用”研究，自此国内开始进行TBM相关研发。研究初期，因技术受限等原因导致专利成果并不高产。直至2019年，国际隧道协会技术委员会更新了TBM的改造指南，为这一领域提供了研发思路，这一年国内TBM专利发表量也以80篇的成绩达到十余年来最高的一次。2019年之后，考虑到疫情等原因的影响，TBM专利发表量有所下降，但总体水平较为稳定，年均37篇，说明企业、高校、科研机构等正不断尝试新的方向，坚持TBM的研发与改造。

Table 7. Entity identification of enterprise demand information
表7. 企业需求信息实体识别

企业名称	项目名称	需求内容	提取关键词
凯**工有限公司	矿用TBM关键技术研究及应用	国家安监总局提出“自动化减人”的目标，煤矿掘进设备需要完成自动化、智能化。寻找矿用全断面硬岩掘进机关键技术研究及应用。	自动化、智能化、掘进机、TBM

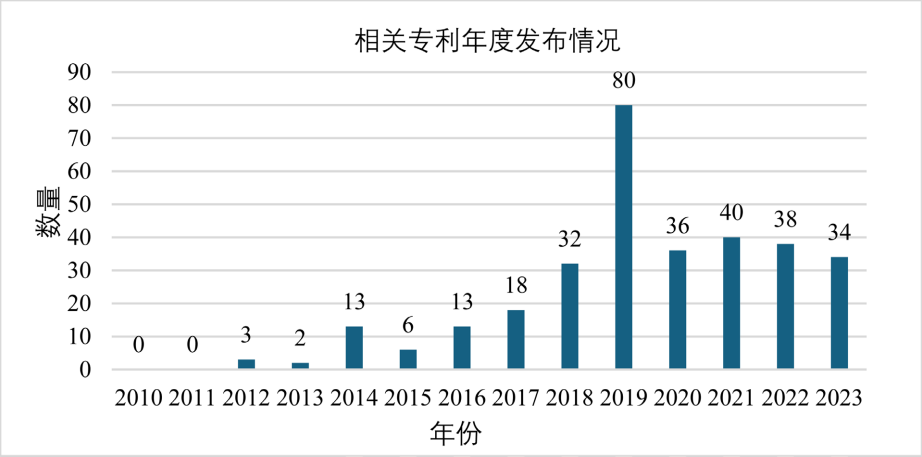

Figure 5. Annual publication statistics of TBM patents
图5. TBM专利年均发布统计

4.2. 技术需求分析

技术需求的分析将基于词频统计及可视化展示的结果。本文使用python的jieba库对文本进行分词，采用哈工大停用词表去除停用词，同时人为去除“本发明”、“技术”、“提出”、“进行”、“方法”、“实现”等无意义词汇后，绘制词云图，如图6所示。词云图中词的大小表示关键词出现频率的高低。从图中可以看出，“TBM”、“掘进”、“施工”等词汇出现频率较高，与项目“矿用TBM关键技术研究与应用”高度相关，进一步证明了本文融合专利数据对项目技术需求进行补充的可靠性。“滚刀”、“刀盘”、“设备”、“油缸”等词提供了TBM中相关工件，要想具备TBM研发技术就需要熟悉这些部件在设备中的原理与作用；“出渣”、“围岩”、“地质”、“断面”等词与工程技术高度相关，说明相关人才需具有一定的施工管理技术才能更好地理解TBM技术创新的目的；“系统”、“参数”、“模

型”、“自动化”、“效率”等词涉及到 TBM 整体研发与改进，该类型的技术发展难度大，对专业人才要求较高，不仅需要熟悉 TBM 工件与系统的工作原理与改进技术，还需掌握计算机、传感、模型开发设计等软件开发技术，这样才能有效推进 TBM 的创新与发展。

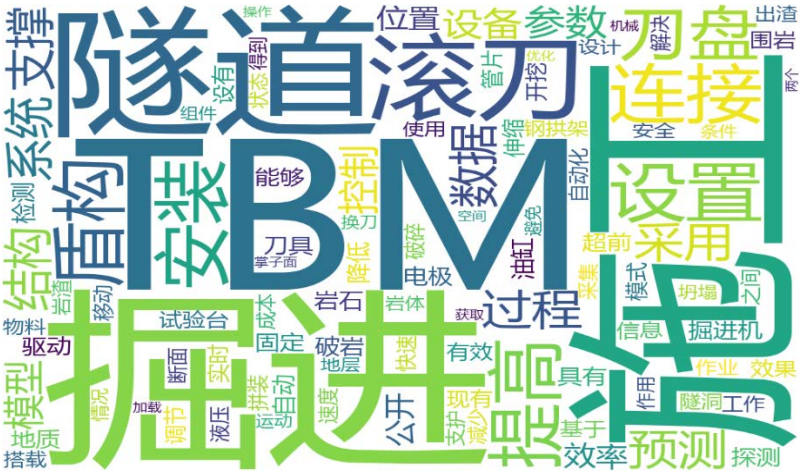

Figure 6. TBM technical word clouds
图 6. TBM 技术词云图

4.3. 跨时间维度的技术需求识别

为了更好地理解 TBM 技术需求的发展路径与研究热点，进一步以时间为维度提取各个年份中专利成果的关键词。为了避免产生词频较高但对本文研究没有作用的无意义词汇，本文采用词频－逆文档频率算法提取关键词，将 TF-IDF 值较高的词汇作为专利成果的关键词。TF-IDF 值的计算过程如式(13)~(15)所示。

$$TF-IDF=TF\left(\omega,d\right)\times IDF\left(\omega\right) \tag{13}$$

$$TF\left(\omega,d\right)=\frac{n_{\omega,d}}{\sum_k n_{k,d}} \tag{14}$$

$$IDF\left(\omega\right)=\log\left(\frac{\left|M\right|+1}{df\left(\omega\right)}\right) \tag{15}$$

式(13)~(15)中，$TF\left(\omega,d\right)$表示词语 ω 在文件 d 中出现的频率：$n_{\omega,d}$ 为词语在文件 d 中出现的次数；$\sum_k n_{k,d}$ 表示文件 d 中 k 个词出现次数的总和。$IDF\left(\omega\right)$ 表示词语 ω 的逆向文件频率：$\left|M\right|$ 是语料库中文件总数；$df\left(\omega\right)$ 是包含词语 ω 的文档数。

对文本进行筛选之后，本文决定以 2014 年为起始年份，分别提取 2014~2023 年专利数据中 TF-IDF 值排名前十的词作为关键词，并统计每个词在语料库的出现的频率，结果如图 7 所示。通过跨时间维度识别技术需求聚焦方向，将有助于确定技术的优先级和重点领域，掌握 TBM 关键核心技术，帮助企业走向技术成熟的蜕变。

通过图 7 可以看出，在 2018 年以前，国内针对 TBM 的研究主要聚焦于施工工艺与掘进机内部构造的研究，将工程技术与 TBM 刀具改进技术等相结合进行研发，解决 TBM 组装拆卸、除尘、适用岩体等与隧道施工相关的技术问题。随着 TBM 应用领域的不断推广，现场施工环境愈发复杂，逐渐发现 TBM 的场景适用性不足，但此时国内 TBM 领域专家的机型设计技术已无法解决这个问题。2018 年，

相关专利中首次出现“自动”、“机械手”等关键词，TBM开启了与人工智能技术相结合的研发思路，由局部改进转向为整体研发，完成了国内掘进机发展由浅入深的大跨越。自此之后，相关专家开始针对 TBM 的驱动系统、数值仿真、模型参数等进行研究，结合传感、机器学习、地质预测、实时监控等技术开启了 TBM 的系统性设计。随着 TBM 运用场景所面临的更加复杂严峻的工况，TBM 与管理、人工智能、大数据、探测、施工设计等技术相结合将是未来发展的方向，以感知为基础，以管理为辅助，以信息集成为中心，以智能掘进为目标，完成 TBM 关键技术的研发与应用，实现 TBM 掘进的自动化与智能化。

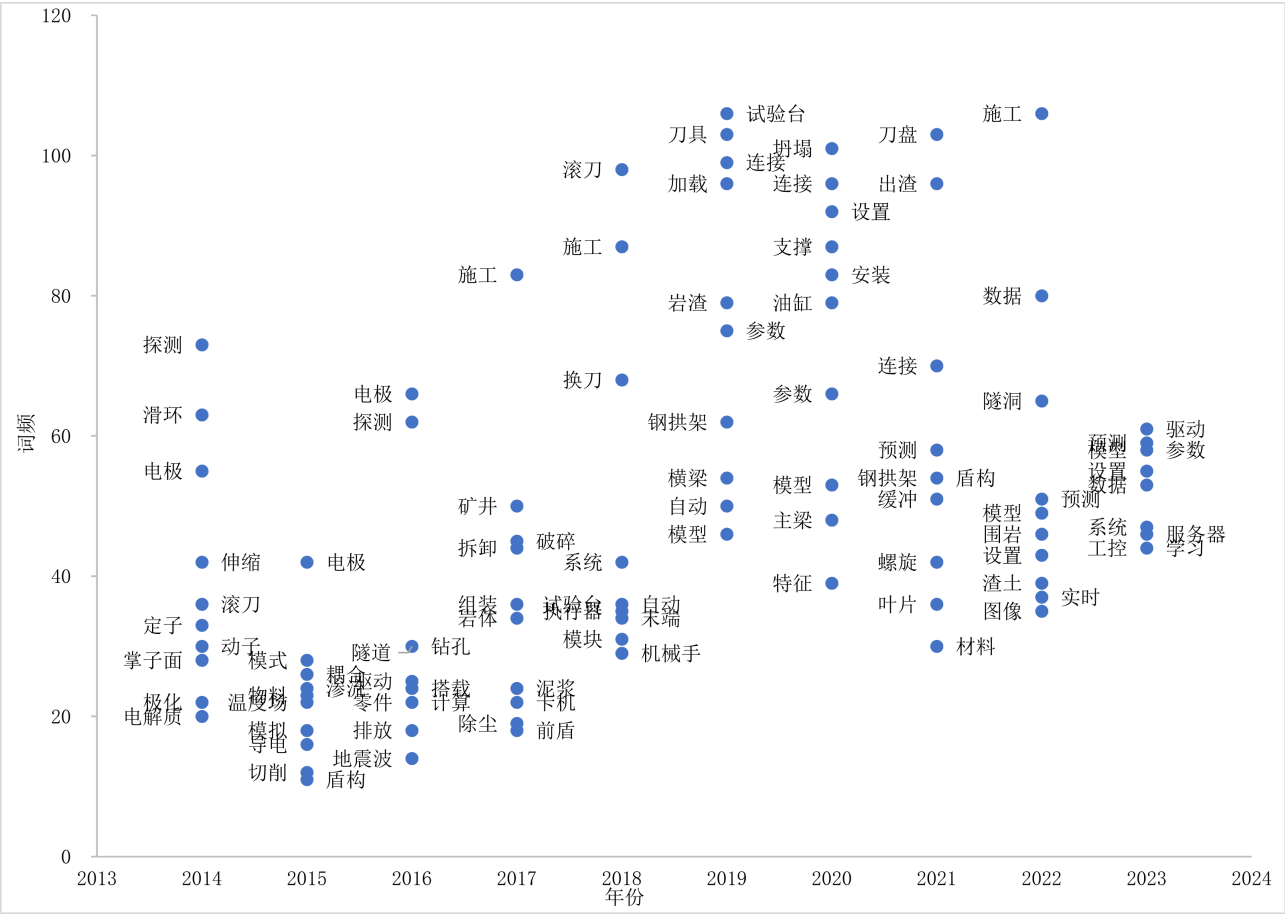

Figure 7. TBM technical analysis across time dimensions
图 7. 跨时间维度的 TBM 技术分析

5. 结论与不足

本文针对企业发布需求具有模糊性的问题，构建了一种两阶段模型。第一阶段提出一种 BERT + BiLSTM + CRF 模型，用于满足企业需求的关键技术实体提取；为了更加精确的理解技术需求，在第二阶段中本文融合了多源信息，以需求实体与企业基本信息作为检索条件爬取相关专利，运用 TF-IDF 算法挖掘专利中关键词作为技术信息的补充，并以时间为维度分析技术的聚焦方向，识别需求的研发重点与核心关键技术点，本研究为模糊型企业需求识别提供了一种可行的思路。后续将在信息来源方面进行进一步扩展，提取技术描述中的研究单位及研究者等实体，从而进行技术需求的精准刻画和推荐。

基金项目

“十四五”湖北省优势特色学科(群)项目：“数字化转型背景下数据驱动的敏捷协同创新理论与方法研究”(项目编号：2023D0402)；湖北省教育厅人文社科重点项目“面向产教精准对接的智能化信息服务与长效机制研究”(项目编号：W201805)。

参考文献

[1] 史敏, 罗建, 周斌. 基于竞争情报的企业技术需求识别 MTS 方法的研究与应用[J]. 图书情报知识, 2018(3): 95-102.

[2] 杨德林, 夏青青, 马晨光. 在线技术转移平台的供需匹配效率分析[J]. 管理科学, 2017, 30(6): 104-112.

[3] Chen, W.L., Zhang, Y.J. and Hitoshi, I. (2006) Chinese Named Entity Recognition with Conditional Random Fields. *Proceeding of the* 15*th Sighan Workshop on Chinese Language Processing*, Sydney, Australia, July 2006, 118-121.

[4] Soomro, P.D., Kumar, S., Banbhrani, Shaikh, A.A. and Raj, H. (2017) Bio-NER: Biomedical Named Entity Recognition Using Rule-Based and Statistical Learners. *International Journal of Advanced Computer Science and Applications*, **8**, 163-170. https://doi.org/10.14569/IJACSA.2017.081220

[5] 包振山, 宋秉彦, 张文博, 等. 基于半监督学习和规则相结合的中医古籍命名实体识别研究[J]. 中文信息学报, 2022, 36(6): 90-100.

[6] Ma, S.C., Xu, J.H. and Fan, Y. (2022) Characteristics and Key Trends of Global Electric Vehicle Technology Development: A Multi-Method Patent Analysis. *Journal of Cleaner Production*, **338**, 1-15. https://doi.org/10.1016/j.jclepro.2022.130502

[7] Cao, Q., Cai, H.J., Wang, J.J., *et al*. (2021) A Scientometric Study of Technological Trend Based on Patent. *Journal of Integrated Design and Process Science*, **23**, 5-28. https://doi.org/10.3233/JID190010

[8] 王京安, 校姜文, 牛建, 汤月. 基于专利分析的技术发展趋势预测研究——以液晶材料技术为例[J]. 科技管理研究, 2019, 39(8): 141-149.

[9] 罗建, 蔡丽君, 史敏. 基于专利的两阶段新兴技术识别研究——以图像识别技术为例[J]. 情报科学, 2019, 37(12): 57 -62.

[10] Sarkar, K. and Shaw, K.S. (2017) A Memory-Based Learning Approach for Named Entity Recognition in Hindi. *Journal of Intelligent Systems*, **26**, 301-321. https://doi.org/10.1515/jisys-2015-0010

[11] 黄水清, 王东波, 何琳. 基于先秦语料库的古汉语地名自动识别模型构建研究[J]. 图书情报工作, 2015, 59(12): 135-140.

[12] Nie, H. (2012) Person-Specific Named Entity Recognition Using SVM with Rich Feature Sets. *Chinese Journal of Library and Information Science*, **5**, 27-46.

[13] Zhang, S., Zhu, H., Xu, H., *et al*. (2022) A Named Entity Recognition Method towards Product Reviews Based on BiLSTM-Attention-CRF. *International Journal of Computational Science and Engineering*, **25**, 479-489. https://doi.org/10.1504/IJCSE.2022.126251

[14] 张召武, 徐彬, 高克宁, 等. 面向教育领域的基于 SVR-BiGRU-CRF 中文命名实体识别方法[J]. 中文信息学报, 2022, 36(7): 114-122.

[15] Peng, N. and Dredze, M. (2016) Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistice*, Berlin, Germany, August 2016, 149-155. https://doi.org/10.18653/v1/P16-2025

[16] Wu, F.Z., Liu, J.X., Wu, C.H., *et al*. (2019) Neural Chinese Named Entity Recognition via CNN-LSTM-CRF and Joint Training with Word Segmentation. *Proceedings of the World Wide Web Conference*, New York NY, 13-17 May 2019, 3342-3348. https://doi.org/10.1145/3308558.3313743

[17] 翟东升, 李梦洋, 何喜军, 徐硕. 非线性技术演化条件下的专利研发投资决策研究[J]. 中国管理科学, 2021, 29(12): 168-178.

[18] Devlin, J., Chang, M.W., Lee, K., *et al*. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. 4171-4186. https://arxiv.org/abs/1810.04805

[19] Graves, A. and Schmidhuber, J. (2005) Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Network*, **18**, 602-610. https://doi.org/10.1016/j.neunet.2005.06.042