文章编号:1671-6833(2020)03-0008-06

基于双维度中文语义分析的食品领域知识库问答

左 敏, 徐泽龙, 张青川, 毕铭文

(北京工商大学 农产品食品安全追溯技术及应用国家工程实验室, 北京 100048)

摘 要: 基于知识库的简单问答是自然语言处理中的一个研究热点,也是实际生活中应用最广泛的一种情况。然而在研究中文方面基于知识库问答的过程中,存在诸如不同处理过程间的错误传播、难以从关系名称表达不明确的问句中抽取关系等问题。在自主构建的食品领域知识库以及食品领域问答语料库的基础上,从中文词义和中文字义两个语义角度出发,利用两个维度进行实体链接,并采用长短时记忆网络进行语义向量建模来抽取关系,提出一种基于双维度的中文语义分析的食品领域知识库问答模型。实验结果表明:所提出的模型在中文食品领域知识库问答上准确率比常用的端到端模型和语义解析模型均高出 5.83%~13.07%,验证了所提出模型的有效性。

关键词: 自然语言处理; 知识库; 问答系统; 实体识别; 关系抽取

中图分类号: TP391.1 **文献标志码:** A **doi:**10.13705/j.issn.1671-6833.2020.02.003

0 引言

问答系统能够自动、准确回答由自然语言组成的问题,在各大知识问答网站以及网络搜索问句中,比例最多的是基于事实的具有单一关系的简单问题,这类问题可以通过查询知识库准确回答。知识库由大量的知识三元组组成,三元组包含实体、属性以及属性值。目前大规模的知识库包括英文的 Freebase、DBpedia 等,以及中文的 OpenKG.CN 和 CN-DBpedia 等。在利用知识库回答简单问题时,可以将知识三元组理解为由实体-关系-答案组成,问答的目的在于通过对问句的语义分析得到知识库中所包含的实体和关系,然后通过知识库进行实体关系匹配,直接获取该实体关系所对应的答案。但在现实中,由于表达方式的原因,存在着两个关键问题:一是如何准确地抽取出关系;二是如何将实体映射到知识库中。

针对上述两个问题开展研究,利用课题来源构建了一定规模的食品领域知识库,作为知识问答系统的底层支撑,并在此基础上利用双维度中文语义分析模型搭建问答系统。模型的核心在于两点:①实体链接阶段,通过字级的相似度以及语义关系将更多识别出的实体名称映射到知识库中;②关系抽取阶段,利用 LSTM 网络对实体关系名称和问句在词级的维度上进行向量建模。在人工构建的食品领域问答语料集上的测试表明,笔者提出的模型获得了 85.66%的准确率。

1 相关工作

对基于知识库的简单问答的研究主要有两个方向:一个是语义解析的方法[1-2],语义解析是将自然语言问句转化为逻辑表达形式,进而可以通过程序执行,从知识库中查询获得答案,但是语义解析的方法往往伴随着大量的人工提取特征,难以在大规模的开放领域上应用;另一个是向量建模的方法[3-4],向量建模将问句和候选知识映射为同一向量空间表示的向量,在向量空间内距离问句最近的候选知识即为正确答案。

在知识库问答中最重要的是将实体映射到知识库中,可以将此过程看作是消除本体异构性的过程。陈淑鑫等[5]利用 WordNet 语义词典库来对本体的不同表达形式进行相似度计算,将实体映射到目标本体之中。张凌宇等[6]利用不同本体之间的多种类型表达进行相似度计算,如名称、内容、属性等,并根据计算结果对本体进行映射。

近些年来,随着深度学习的发展,研究者开始

收稿日期:2019-10-27;**修订日期:**2019-12-12

基金项目:国家重点研发计划项目(2016YFD0401205);北京市自然科学基金资助项目(4202014)

作者简介:左敏(1973—),男,安徽铜陵人,北京工商大学教授,博士,主要从事食品安全大数据研究,E-mail:zuomin1234@163.com。

将传统的方法与神经网络相结合进行试验。Yih 等[7]提出了一个新型的语义解析框架,框架通过阶段性生成查询图,并且采用卷积神经网络寻找关系来提升问答系统的准确率。为了更好地提升问答的准确率,使得字母级的编码能更好地处理字词不在词典中出现的情况[8],Lukovnikov 等[9]通过将英文中的字级向量与字母级向量进行结合,提高了词向量质量,并且使用端到端的模型直接从问句中抽取出实体-关系对,不再使用分开的流水线任务来处理问题,减轻了自然语言处理流水线中的错误传播问题的程度。Hao 等[10]先对问句进行模式抽取和实体链接,然后采用模式修订来减轻错误传播问题的程度。

基于知识库的中文问答研究起步较晚,Lai 等[11]提出了 SPE (subject predicate extraction) 算法,从问句中自动抽取实体-谓词对,然后在知识库中查询获取答案,该算法在 NLPCC-ICCPOL 2016 竞赛中的开放领域知识问答任务中获得了最好成绩。周博通等[12]利用长短时记忆(long short term memory, LSTM)网络在相同的数据集上进行试验,利用注意力机制从关系候选集中选出最相似的关系名称,也取得了不错的结果。

2　食品领域知识库构建

笔者采用独立构建的食品领域知识库(FD-KB),其中实体来自国家食品抽检检测数据,包括食品名称、风险因子以及食品添加剂等食品名称实体,共计 0.7 M。以该实体库为种子,利用网络爬虫从各大权威百科网站获取知识,并以三元组的形式存储于 FD-KB 中。由于实体与关系名称的表达方式多样,不同平台获取的知识无法直接融合,利用关系重合率对知识三元组进行校正和整合。

要重点说明的是,所使用的实体库来自国家食品抽检检测数据,是面向专业领域的实体库,所包含的实体名称大都为专业名称。下面将详细介绍如何利用关系重合率对知识进行校正和整合。

食品领域实体库中的实体是从知识平台获得的关系字典,以键值对的方式表示。当从不同平台获得不同的关系字典时,利用式(1)计算关系重合率:

$$p_r = \frac{n((R_s \cdot keys) \cap (R_b \cdot keys)) + n((R_s \cdot values) \cap (R_b \cdot values))}{2\max\{n(R_s), n(R_b)\}}, \quad (1)$$

式中:$R \cdot keys$ 为所有键的集合;$R \cdot values$ 为所有值的集合;$n(\cdot)$ 为计数函数;$\max(\cdot)$ 为最大取值函数;R_s 和 R_b 表示两个不同的关系字典。

在这里规定:当 $p_r \leqslant 0.5$ 时,两个关系字典属于不同的实体(实体名称可能相同);否则,两个关系字典所提到的是同一个实体,需要对关系字典进行整合。在整合时,利用词袋模型对实体的关系名称进行编码,并计算编码之间的余弦相似度。根据计算结果,对实体的知识三元组进行整合或补充。

获取知识三元组时,对实体进行了清洗,包括去除关系名称中的空白符等无关符号(即非中文、英文和数字的符号),将英文字母统一为小写格式,最终得到包含 6 M 知识三元组的 FD-KB。

3　双维度中文语义分析模型

提出的模型使用了中文字符级和词语级两个语义维度,主要包含 3 个步骤:实体识别、实体链接和关系抽取。图 1 展示了模型的执行过程。

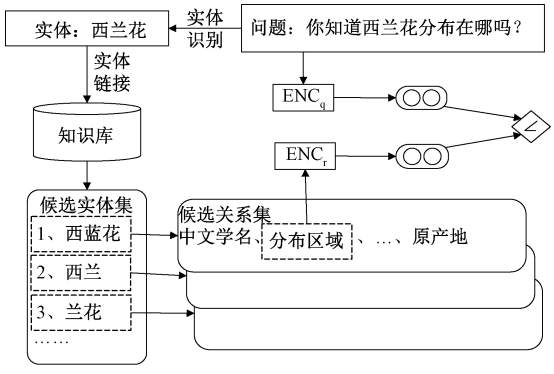

图 1　双维度中文语义分析模型图

Figure 1　The model of two-dimensional Chinese semantic analysis

步骤 1　通过实体识别得到问句中的实体信息。在实体识别阶段利用中文字符级的语义向量编码,字级语义编码能够不受问句中错别字的影响从而能正确标注出实体在问句中的位置。

步骤 2　实体链接阶段利用字、词两个维度的语义相似度获得实体候选集,通过对实体识别结果校正,将更多的识别结果映射到知识库中。

步骤 3　通过实体候选集获得关系候选集,使用 LSTM 网络对问句与关系名称进行词级的语义向量编码,计算得到更接近的关系。

在模型中,错误传播始于实体识别,其他两个步骤的运行依赖于实体识别的结果。实体链接是一个有承上启下作用的关键步骤,通过优化实体链接的算法,能有效减轻错误传播问题的程度。接下来将详细介绍模型的各个阶段。

3.1 实体识别

长短时记忆网络在自然语言处理中发挥了重要的作用,它的记忆单元由3个门构成:输入门、遗忘门和输出门,这种结构能够帮助网络有效控制信息的记忆与遗忘,使其能够比循环神经网络识别更长距离的上下文信息。

使用双向LSTM网络和条件随机场模型(conditional random field,CRF)[13]来识别问句中的实体,这样做的优势在于:①双向LSTM网络可以充分利用句子的正反序列信息;②条件随机场模型可以避免最终结果产生不合理的标签序列。由于模型对词向量的质量好坏依赖较小,并且考虑到问句中的词语可能会出现错别字或者超出词典范畴的情况,从而导致分词错误,进而影响实体识别的结果,因此在训练时采用字级编码作为输入。具体模型如图2所示。

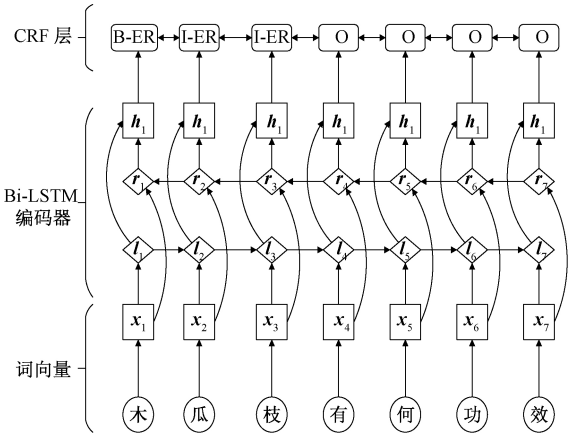

图2 Bi-LSTM-CRF网络模型图

Figure 2 The structure of Bi-LSTM-CRF

Bi-LSTM层:对于一个输入长度为n的问句$Q(w_1,w_2,\cdots,w_n)$,利用one-hot编码得到每个字的向量$\boldsymbol{X}(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$,之后将$\boldsymbol{X}$分别以正序和倒序输入到两个不同的LSTM网络中,最终在时长t获得的状态$\boldsymbol{h}_t$包含此刻的上下文信息。

CRF层:Bi-LSTM层的输出是每一个中文字符w_t被标记为每一个标签的概率,最终的概率矩阵作为CRF层的输入来计算不同标签序列的得分,这种方式能够有效避免不合理的标签序列,例如B-ER,O,I-ER等。

3.2 实体链接

如果将FD-KB中的知识三元组看作由实体-关系-答案组成,那么实体链接阶段的目的就是将第一步的识别结果映射到FD-KB中的实体上,将更多的结果映射到相应的实体上对模型准确率的提升至关重要。

使用中文字符级与词语级语义相结合的链接方法,用C'表示识别结果的中文字符集合,C表示FD-KB中实体名称的中文字符集合。在计算词语级语义时,利用预训练好的中文Word2Vec词向量对词语进行表示[14]。$\boldsymbol{W}'$为识别结果的矩阵向量,$\boldsymbol{W}$为候选实体的矩阵向量。中文字符的语义相似度得分通过计算集合中的重复字符得到,词语级语义相似度则通过计算矩阵的余弦相似度得到,总得分由式(2)计算得到:

$$s=\alpha\cdot\frac{n(C'\cap C)}{n(C)}+(1-\alpha)\cdot\cos(\boldsymbol{W}',\boldsymbol{W}),\quad(2)$$

式中:α为权重系数。当$\alpha=1$时,表示只依赖中文字符级语义;当$\alpha=0$时,表示只使用词语级语义相似度。

3.3 关系抽取

关系抽取更精确地说是关系匹配,即根据问句的描述计算得到最相似的关系名称。实际上,对于中文的简单问句来说,很多关系名称直接包含在问句中,因此可以直接从候选关系集合中得到关系名称。统计显示,大约有52.17%的训练语料符合上述情况。

根据上述情况,首先,在关系抽取中利用正则表达式获取直接包含的关系名称。值得强调的是,通过这种方法得到的关系名称长度必须大于1,否则会出现包含多个关系的情况。其次,对于无法直接抽出关系的情况,则利用不同的LSTM网络来获得候选关系名称和问句的向量表示,然后计算它们的余弦相似度得到最优结果,如图3所示。

步骤1 利用问句编码器$\mathrm{ENC_q}$对去除实体后的问句进行编码。首先对问句进行分词,然后得到词语的向量编码作为LSTM网络的输入,最后获得神经网络的最终状态为问句的向量表示,计算得到:

$$\boldsymbol{r}_q=\mathrm{ENC_q}(\boldsymbol{q}_{w1},\boldsymbol{q}_{w2},\cdots,\boldsymbol{q}_{wt})。\quad(3)$$

步骤2 关系编码器$\mathrm{ENC_r}$与$\mathrm{ENC_q}$结构相似,关系名称的向量表示计算方法如式(4)所示:

$$\boldsymbol{r}_{\mathrm{r}}=\mathrm{ENC}_r(\boldsymbol{r}_{w1},\boldsymbol{r}_{w2},\cdots,\boldsymbol{r}_{wn})。\quad(4)$$

步骤3 利用式(5)计算$\boldsymbol{r}_q$与$\boldsymbol{r}_{\mathrm{r}}$的余弦相似度:

$$Score=\cos(\boldsymbol{r}_q,\boldsymbol{r}_r)=\frac{\boldsymbol{r}_q\cdot\boldsymbol{r}_r}{|\boldsymbol{r}_q|\cdot|\boldsymbol{r}_r|}。\quad(5)$$

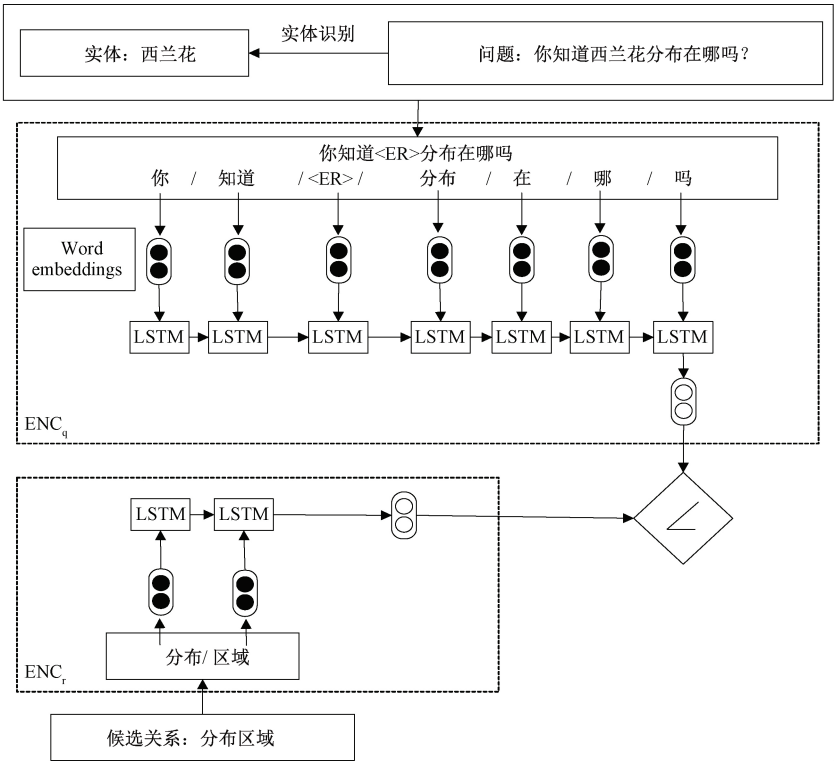

图 3　关系抽取模型图

Figure 3　The model of relation extraction

4　实验

4.1　语料以及评估方法

所使用的问答语料是通过人工的方法利用 FD-KB 构造而成,问句的模式符合实际生活中的用户问句,共有 23 000 条。在构建语料的同时也标注了问句所依赖的知识三元组,语料被随机分为 14 000 条训练语料以及 9 000 条测试语料。

利用准确率来评估模型的效果,准确率的计算方法如式(6)所示:

$$Accuracy = \frac{n_{\text{correct}}}{N} \times 100\%, \qquad (6)$$

式中:N 代表所有的样本总数;n_{correct} 表示得到正确结果的样本数。

4.2　实验步骤

实体识别步骤所使用的语料是依据问答语料标注而成,同样也分为 14 000 条训练语料以及 9 000 条测试语料。实体识别的目的是在问句中标注出最有可能是实体的位置,因此对于一些打印错误的情况,也将其标注为实体。通过上述的标注策略,一些实体的关键信息将不会被遗漏。实体识别的最终准确率为 92.78%。

在实体链接阶段,通过调整 α 的取值进行调优,得到字级语义和词级语义的最佳结合点。通过训练,当 $\alpha=0.63$ 时,能得到最好的实体链接效果,找到实体的准确率提升至 93.38%,如图 4 所示。

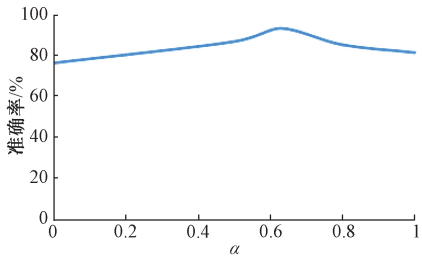

图 4　α 与准确率的关系

Figure 4　Accuracy when α takes different values

关系抽取时,训练 LSTM 神经网络的主要目的是为了最小化损失函数,损失函数的计算方法如式(7)所示:

$$Loss_{q,p,p'} = \max\{0, Score(q,p') - Score(q,p) + \gamma\}, \qquad (7)$$

式中:$Score(q,p')$ 表示负样本的得分;$Score(q,p)$ 表示正样本的得分;γ 表示正样本的得分必须高于负样本得分 γ 分,在本文中 $\gamma = 0.3$。

为了进一步评估模型,利用其他两种常用的知识库问答模型来做对比实验,分别是语义解析模型和端到端模型,实验结果如表 1 所示。对比实验结果显示,提出的模型能够更好地解决中文食品领域内的知识库问答。

表1 对比实验结果

Table 1 The results of the contrast experiments

模型名称	准确率/%
语义解析模型	79.83
端到端模型	72.59
双维度语义模型	85.66

4.3 实验结果分析

根据实验结果可以发现,笔者提出的模型问答准确率高于语义解析模型及端到端模型。在问答模型中,实体识别是影响整个模型准确率的关键,通过对识别错误的实体进行分析能够减轻错误传播问题的程度。

识别错误的情况大致分为3种:①正确的实体名称与问句中的实体名称不一致;②少识别了一些中文字符;③关系名称包含在了识别结果中。针对前两种情况,笔者提出的模型利用字级与词级两个维度的语义信息进行计算,将实体正确地映射到知识库中,而第3种情况会严重影响之后的关系识别结果,该问题仍有待解决。

5 结论

分别从中文问答的字义与词义两个语义维度出发,将中文单词所蕴含的词义表达与字义相结合,有效地提升了问句中实体映射的准确率,提出了一种基于字词双维度中文语义分析的食品领域知识库问答模型。该模型能够有效减轻处理过程中的错误传播问题的程度,并能够提高问句中语义关系提取的准确率。实验中所采用的FD-KB以及问答语料具有中文问答的特点,也符合实际的网络搜索情况,如果将本方法应用到大规模开放领域知识库也会取得不错的效果。

参考文献:

[1] CAI Q Q, YATES A. Large-scale semantic parsing via schema matching and lexicon extension[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria:ACL,2013:423-433.

[2] FADER A, ZETTLEMOYER L, ETZIONI O. Open question answering over curated and extracted knowledge bases[C]//Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. New York: ACM,2014:1156-1165.

[3] BORDES A, WESTON J, USUNIER N. Open question answering with weakly supervised embedding models [C]//The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Nancy, France: ECML/PKDD, 2014: 165-180.

[4] YANG M C, DUAN N, ZHOU M, et al. Joint relational embeddings for knowledge-based question answering[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: EMNLP,2014:645-650.

[5] 陈淑鑫,张凌宇.基于WordNet的本体查询方法研究[J].郑州大学学报(工学版),2016,37(3):16-21.

[6] 张凌宇,马志晟,陈淑鑫.一种基于多种类型匹配器的本体映射方法[J].郑州大学学报(工学版),2015,36(3):106-109.

[7] YIH W T, CHANG M W, HE X D, et al. Semantic parsing via staged query graph generation: question answering with knowledge base[C]//The 53rd Annual Meeting of the Association for Computational Linguistics & the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing: ACL, 2015: 1321-1331.

[8] GOLUB D, HE X D. Character-level question answering with attention[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: EMNLP, 2016: 1598-1607.

[9] LUKOVNIKOV D, FISCHER A, LEHMANN J, et al. Neural network-based question answering over knowledge graphs on word and character level[C]//International World Wide Web Conference Committee. Perth: ACM,2017:1211-1220.

[10] HAO Y C, LIU H, HE S Z, et al. Pattern-revising enhanced simple question answering over knowledge bases[C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: ACM, 2017:3272-3282.

[11] LAI Y X, LIN Y, CHEN J H, et al. Open domain question answering system based on knowledge base [C]//The 5th Conference on Natural Language Processing and Chinese Computing & The 24th International Confer-ence on Computer Processing of Oriental Languages. Kunming: CCF,2016:722-733.

[12] 周博通, 孙承杰, 林磊, 等. 基于LSTM的大规模知识库自动问答[J]. 北京大学学报(自然科学版), 2018, 54(2): 286-292.

[13] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging [EB/OL]. (2015-08-09) [2019-09-30].http://arxiv.org/abs/1508.01991.

[14] LI S, ZHAO Z, HU R F, et al. Analogical reasoning

on chinese morphological and semantic relations[C]// The 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018: 138-143.

A Question Answering Model of Food Domain Knowledge Bases with Two-Dimension Chinese Semantic Analysis

ZUO Min, XU Zelong, ZHANG Qingchuan, BI Mingwen

(National Engineering Laboratory for Agri-product Quality Traceability, Beijing Technology and Business University, Beijing 100048, China)

Abstract: Simple Question Answering over Knowledge Bases (KB-QA) was a hot topic in the field of Natural Language Processing (NLP), and it was also the most widely used case in real life. However, in the field of Chinese KB-QA, there were still many technical challenges such as extracting relations from questions which relation names were ambiguous, and problems such as error propagation between different processes. Based on the self-built food domain knowledge base (FD-KB) and the food field corpus, this paper proposed a new perspective based on two semantic dimensions of Chinese characters and Chinese words to extract relations and mitigate the error propagation. Contrasting experimental results showed that the model of two-dimensional Chinese semantic analysis that proposed here was 5.83%~13.07% higher than the end-to-end model and the semantic parsing model, and verified its rationality and validity.

Key words: natural language processing; knowledge base; question answering; entity recognition; relation extraction

---

(上接第 7 页)

Review of the Analysis Methods of Effective Connectivity Based on Granger Causality

SHANG Zhigang[1,2], SHEN Xiaoyang[1,2], LI Mengmeng[1,2], WAN Hong[1,2]

(1.School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China; 2. Henan Key Laboratory of Brain Science and Brain-Computer Interface Technology, Zhengzhou University, Zhengzhou 450001, China)

Abstract: At present, the effective connectivity analysis methods based on Granger causality was widely used in neural signals analysis of multiple brain regions. First of all, the calculation principle and functional characteristics of representative algorithms commonly used in this kind of method were systematically introduced. Then the key points that should be paid attention to in practical application of this kind of methods were summarized. Finally, the improved algorithm for Generalized Partial Directed Coherence was taken as examples to show the application effect on the actual electroencephalogram data set.

Key words: Granger causality; effective connectivity; neural signals; information flow