中南大學
CENTRAL SOUTH UNIVERSITY

论文翻译

THESIS TRANSLATION

题　　目:	基于动态多池化卷积神经网络的事件抽取
学生姓名:	杨程
指导老师:	邓磊 教授
学　　院:	计算机学院
专业班级:	软件工程 1704

本科生院制

2021 年 3 月

基于动态多池化卷积神经网络的事件抽取

摘要

传统方法的ACE事件提取主要依赖于精心设计的特征和复杂的自然语言处理(NLP)工具。这些传统的方法缺乏普遍性、需要消耗大量的人力资源，故而容易出现错误传播和数据稀疏的问题。因此，本文提出了一种新的事件抽取方法方法，其目的是在不使用复杂的 NLP 工具的情况下自动提取词汇级和句子级特征。本文引入一种单词表示法获取单词有意义的语义规则并采用了一个基于卷积神经网络 (CNN) 的框架来捕获句子级别的线索。然而，CNN 在一个句子中只能捕捉到其最重要的信息，所以在考虑多事件句子时可能会漏掉关键信息。为了解决上述问题，本文提出了一种动态多池化卷积神经网络（Dynamic Multi-Pooling Convolutional Neural Network，DMCNN），它根据事件触发器和参数使用动态多池层来保留更多的关键信息。实验结果表明，本文所提出的方法明显优于其他最先进的方法。

关键词：事件抽取　卷积神经网络　自然语言处理

Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks

ABSTRACT

Traditional approaches to the task of ACE event extraction primarily rely on elaborately designed features and complicated natural language processing (NLP) tools.These traditional approaches lack generalization,take a large amount of human effort and are prone to error propagation and data sparsity problems. This paper proposes a novel event-extraction method, which aims to automatically extract lexical-level and sentence-level features without using complicated NLP tools. We introduce a word-representation model to capture meaningful semantic regularities for words and adopt a framework based on a convolutional neural network (CNN) to capture sentence-level clues.However, CNN can only capture the most important information in a sentence and may miss valuable facts when considering multiple-event sentences. We propose a dynamic multi-pooling convolutional neural network (DMCNN), which uses a dynamic multi-pooling layer according to event triggers and arguments, to reserve more crucial information. The experimental results show that our approach significantly outperforms other state-of-the-art methods.

Key words: Event Extraction Convolutional Neural Networks Natural Language Processing

目录

第 1 章　介绍

事件抽取是信息提取（IE）中的一项重要且具有挑战性的任务，旨在发现具有特定类型及其参数的事件触发器。目前最先进的方法（Li et al，2014; Li et al，2013; Hong et al，2011; Liao and Grishman，2010; Ji and Grishman，2008）经常使用一组精心设计的特征，通过文本分析和语言知识提取。一般来说，我们可以将这些特征分为两类: 词汇特征和语境特征。词汇特征一般包含词性标签（POS），实体信息和形态特征，目的是捕获语义或单词的背景知识。例如，下面这两个例子中就有一个含糊不清的单词 *beats*：

S1: *Obama **beats** McCain.*

S2: *Tyson **beats** his opponent.*

在第一句话中，*beats* 是一个 *Elect* 类型的触发词，而在第二句话中，*beats* 是一个 *Attack* 类型的触发词，由于这是同一个单词，所以传统的方法可能会将第一句话中的 *beats* 错误的标成 *Attack* 类型的触发词。如果我们知道奥巴马和麦凯恩都是总统竞选人，就可以预测 *beats* 是 *Elect* 类型的触发词。我们把这些知识称为**词汇层面的线索**。为了表示这些特征，现有的方法（Hong et al，2011）经常依赖于人工标注，这是一个耗时的过程，且缺乏通用性。此外，先前方法中的传统词汇特征是 one-hot 表示，其可能遭受数据稀疏性问题并且可能无法充分捕获单词的语义（Turian et al，2010）。

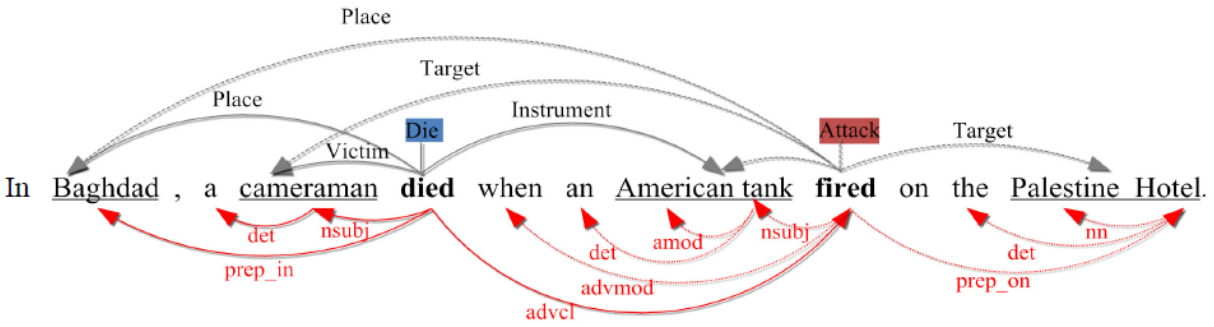


图 1-1　事件提及 S3 的语法分析器结果

上方显示了两个事件提及，它们共享三个参数：*Die* 事件提及，由 “*died*” 触发，而 *Attack* 事件提及由 “*fired*” 触发。下方显示折叠的依赖性结果。

S3: *In Baghdad, a cameraman **died** when an American tank **fired** on the Palestine Hotel.*

为了更准确地识别事件和参数，以前的方法通常会捕获上下文特征，例如句法特征，其目的是从更大的视角了解事实如何联系在一起。例如，在 S3 中，有两个共享三个参数的事件，如图 1-1 所示。从参数 *cameraman* 和触发词 *die* 之间的 *nsubj* 依赖关系，我们可以在 *Die* 事件中向摄影师添加受害者角色。我们称这种信息为**句子级的线索**。但是，参数 *cameraman* 及其触发的单词在不同的子句中，并且它们之间没有直接的依赖路径。因此，使用传统的依赖功能很难在它们之间找到目标角色。此外，提取此类特征在很大程度上取决于现有的 NLP 系统的性能，而 NLP 系统可能会遭受错误传播。

要正确地将 *cameraman* 识别为 *fired* 的 *Target* 参数，我们必须利用整个句子的内部语义，以便使得 *Attack* 事件可以导致 *Die* 事件。对卷积神经网络（CNN）的改进已被证明对于捕获 NLP 任务的句子内的单词之间的句法和语义是有效的。CNN 通常使用最大池化层，其对整个句子的表示应用最大操作以捕获最有用的信息。但是，在事件提取中，一个句子可能包含两个或多个事件，并且这些事件可能共享具有不同角色的参数。传统的 CNN 只使用最重要的信息表示句子，将会错过有价值的线索。例如，S3 中有两个事件，*Die* 事件和 *Attack* 事件。如果我们使用传统的最大汇集层并且只保留最重要的信息来表示句子，我们可能会获得描述“摄影师死亡”的信息但却错过了“美国坦克向巴勒斯坦酒店开火”的信息，这对于预测攻击事件具有重要意义，对于将 *cameraman* 识别为 *fired* 的 *Target* 参数也很有价值。在我们的实验中，我们发现这样的多事件句占我们数据集的 27.3%，这是一个我们不能忽视的现象。

所以，本文提出动态多池化卷积神经网络（DMCNN）来解决上述问题。为了捕捉词汇层面的线索并减少人为干预，我们引入了一个单词表示模型（Mikolov et al，2013b），它已经被证明能够捕获单词的有意义的语义规律（Bengio et al，2003; Erhan et al，2010; Mikolov et al，2013a）。为了在不使用复杂的 NLP 工具的情况下捕获句子级线索，并更全面地保留信息，我们为 CNN 设计了动态多池层，它根据事件触发器和参数返回句子每个部分的最大值。本文的贡献如下：

- 我们提出了一种新的事件提取框架，它可以自动从纯文本中引入词汇级和句子级特征，而无需复杂的 NLP 预处理。
- 我们设计了一个动态多池卷积神经网络（DMCNN），旨在捕获句子中更有价值的信息以进行事件提取。
- 我们对广泛使用的 ACE2005 事件提取数据集进行了实验，实验结果表明我们的方法优于其他最先进的方法。

1.1 二级标题

1.1.1 三级标题

(1) 四级标题

一级标题根据学校提供的 Word 模板要求，三号黑体居中，上下各空一行，章节号空一个汉字，并且每一章节单独起一页，章节号格式应使用阿拉伯数字而非中文汉字。

二级标题为小四号黑体，缩进两个汉字。章节号后空一个汉字。

三级标题小四号楷体 GB2312，字体包含在项目中，同样缩进两个汉字，章节号后空一个汉字。

四级标题参照本科学术论文设计样式，分项采取 (1)、(2)、(3) 的序号。

所有标题样式由 undergraduate.cls 模板文件 `\ctexset` 进行设置。

1.2 字体

正文字体默认使用小四号宋体，英文为小四号 Times New Romen，各段行首缩进两个汉字

中南大学坐落在中国历史文化名城 —— 湖南省长沙市，占地面积 317 万平方米，建筑面积 217 万平方米，跨湘江两岸，依巍巍岳麓，临滔滔湘水，环境幽雅，景色宜人，是求知治学的理想园地。

中南大学由原湖南医科大学、长沙铁道学院与中南工业大学于 2000 年 4 月合并组建而成。原中南工业大学的前身为创建于 1952 年的中南矿冶学院，原长沙铁道学院的前身为创建于 1953 年的中南土木建筑学院，两校的主体学科最早溯源于 1903 年创办的湖南高等实业学堂的矿科和路科。原湖南医科大学的前身为 1914 年创建的湘雅医学专门学校，是我国创办最早的西医高等学校之一。中南大学秉承百年办学积淀，顺应中国高等教育体制改革大势，弘扬以“知行合一、经世致用”为核心的大学精神，力行“向善、求真、唯美、有容”的校风，坚持自身办学特色，服务国家和社会重大需求，团结奋进，改革创新，追求卓越，综合实力和整体水平大幅提升。

英文字体展示如下：

TeX (/t□x, t□k/, see below), stylized within the system as TEX, is a typesetting system (or a ”formatting system”) which was designed and mostly written by Donald Knuth[1] and released in 1978. TeX is a popular means of typesetting complex mathematical formulae; it has been noted as one of the most sophisticated digital typographical systems.

1.2.1 调节字号

可以使用 `\zihao` 命令来调节字号。

`\zihao{3}`	三号字 English
`\zihao{-3}`	小三号 English
`\zihao{4}`	四号字 English
`\zihao{-4}`	小四号 English
`\zihao{5}`	五号字 English
`\zihao{-5}`	小五号 English

1.2.2 调节字体

需要说明的是由于学校写作指导要求的字体部分不可在 Linux 上使用，即便你的写作过程是在 Linux 或者 macOS 上完成的，我们仍**强烈建议**您在 Windows 操作系统上编译最终版论文。

中文字体可以使用如下命令来调节。

\songti 宋体

\heiti **黑体**

1.3 模板主要结构

本项目模板的主要结构, 如下表所示:

csuthesis_main.tex		主文档，可以理解为文章入口。
content 目录	info.tex	作者、文章基本信息
	abstactzh/en.tex	中/英文摘要内容
	subchapters 目录	章节内容
images 目录		用于存放图片文件
csuthesis.cls		模板入口

我们不建议模板使用者更改原有模板的结构，但如果您确实需要，请务必先充分阅读本模板的使用说明并了解相应的 LaTeX 模板设计知识。

第 2 章　图表示例

2.1　图片与布局

2.1.1　插图

图片可以通过\includegraphics 指令插入，我们建议模板使用者将文章所需插入的图片源问卷放置在 images 目录中，另外，矢量图片应使用 PDF 格式，位图照片则应使用 JPG 格式（LaTeX 不支持 TIFF 格式）。具有透明背景的栅格图可以使用 PNG 格式。

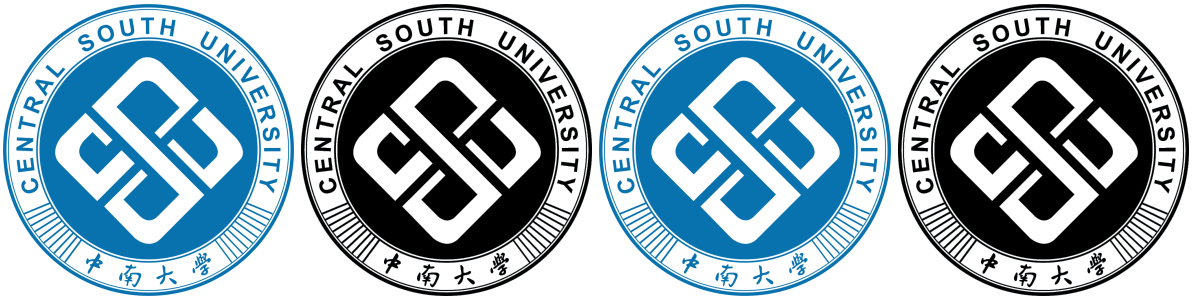
下面是一个简单的插图示例。

图 2-1　插图示例

如果一个图由多个分图（子图）组成，应通过 (a),(b),(c) 进行标识并附注在分图（子图下方）。目前子图标识不居中问题没有解决，预计下个版本修复。

2.1.2　横向布局

模板提供常见的图片布局，比如单图布局2-1，另外还有横排布局如下：



(a) test　(b) test　(c) test　(d) test

图 2-2　图片横排布局示例

2.2　纵向布局

纵向布局如图2-3

(a) test

(b) test

图 2-3　图片纵向布局示例

2.3　竖排多图横排布局

(a) aaa　　(b) bbb

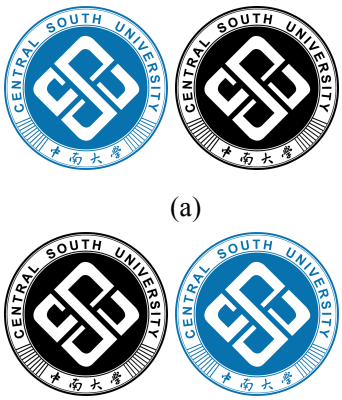
图 2-4　图片竖排多图横排布局

竖排多图横排布局如图2-4所示。注意看 (a)、(b) 编号与图关系

2.4　横排多图竖排布局

中南大学由原湖南医科大学、长沙铁道学院与中南工业大学于 2000 年 4 月合并组建而成。原中南工业大学的前身为创建于 1952 年的中南矿冶学院，原长沙铁道学院的前身为创建于 1953 年的中南土木建筑学院，两校的主体学科最早溯源于 1903 年创办的湖南高等实业学堂的矿科和路科。原湖南医科大学的前身为 1914 年创建的湘雅医学专门学校，是我国创办最早的西医高等学校之一。中南大学秉承百年办学积淀，顺应中国高等教育体制改革大势，弘扬以“知行合一、经世致用”为核心的大学精神，力行“向善、求真、唯美、有容”的校风，坚持自身办学特色，服务国家和社会重大需求，团结奋进，改革创新，追求卓越，综合实力和整体水平大幅提升。

横排多图竖排布局如图2-5所示。注意看 (a)、(b) 编号与图关系。



(a)

(b)

图 2-5　图片横排多图竖排布局

第 3 章　表格插入示例

表 0-1　学校文件里对表格的要求不是很高，不过按照学术论文的一般规范，表格为三线表。

	A	B	C	D	E
1	212	414	4	23	fgw
2	212	414	v	23	fgw
3	212	414	vfwe	23	嗯
4	212	414	4fwe	23	嗯
5	af2	4vx	4	23	fgw
6	af2	4vx	4	23	fgw
7	212	414	4	23	fgw

表格如表0-1所示，latex 表格技巧很多，这里不再详细介绍。

中南大学由原湖南医科大学、长沙铁道学院与中南工业大学于 2000 年 4 月合并组建而成。原中南工业大学的前身为创建于 1952 年的中南矿冶学院，原长沙铁道学院的前身为创建于 1953 年的中南土木建筑学院，两校的主体学科最早溯源于 1903 年创办的湖南高等实业学堂的矿科和路科。原湖南医科大学的前身为 1914 年创建的湘雅医学专门学校，是我国创办最早的西医高等学校之一。中南大学秉承百年办学积淀，顺应中国高等教育体制改革大势，弘扬以“知行合一、经世致用”为核心的大学精神，力行“向善、求真、唯美、有容”的校风，坚持自身办学特色，服务国家和社会重大需求，团结奋进，改革创新，追求卓越，综合实力和整体水平大幅提升。

第 4 章　公式插入示例

中南大学由原湖南医科大学、长沙铁道学院与中南工业大学于 2000 年 4 月合并组建而成。原中南工业大学的前身为创建于 1952 年的中南矿冶学院，原长沙铁道学院的前身为创建于 1953 年的中南土木建筑学院，两校的主体学科最早溯源于 1903 年创办的湖南高等实业学堂的矿科和路科。原湖南医科大学的前身为 1914 年创建的湘雅医学专门学校，是我国创办最早的西医高等学校之一。中南大学秉承百年办学积淀，顺应中国高等教育体制改革大势，弘扬以“知行合一、经世致用”为核心的大学精神，力行“向善、求真、唯美、有容”的校风，坚持自身办学特色，服务国家和社会重大需求，团结奋进，改革创新，追求卓越，综合实力和整体水平大幅提升。

公式插入示例如公式 (4.1) 所示。

$$\gamma_x = \begin{cases} 0, & \text{if } |x| \leq \delta \\ x, & \text{otherwise} \end{cases} \tag{4.1}$$

第 5 章　引用文献标注

文献标注和索引的处理一直是学术写作的一个麻烦事，特别是在 word 环境下。latex 中我们只需要编辑（或直接获取）BibTeX 格式索引文件然后在正文中使用`\cite` `\citet` 等指令进行引用标注就可以。下面介绍在文章中引用指令的具体使用方法。

5.1　顺序编码

根据学校要求，参考文献标注用中括号上标形式进行标注。使用方式与效果如下表所展示

`\cite{knuth1984texbook}`	⇒	[1]
`\citet{knuth1984texbook}`	⇒	Knuth et al.[1]
`\citep{knuth1984texbook}`	⇒	[1]
`\cite{knuth1984texbook,lamport1994latex}`	⇒	[1, 2]

5.2　获取 BibTeX 格式索引

获取参考文献的 BibTeX 格式索引有两种方式

- 通过 Google Scholare 或者百度学术等学术文献搜索引擎获取，自行编辑.bib 文件
- 通过 Zotero 等学术文献整理软件，添加所有的引用文献至库中，导出对应的.bib 文件

编译带参考文献的文章时，我们需要两次编译过程。我们提供了对应的自动化脚本，以及配合 vscode latex 插件的任务流程，帮助模板使用者进行编译。

5.3　参考文献插入示例

LaTeX[2] 插入参考文献最方便的方式是使用 bibliography[3]，大多数出版商的论文页面[2, 3] 都会有导出 bib 格式参考文献的链接，把每个文献的 bib 放入“csuthesis_main.bib”，然后用 bibkey 即可插入参考文献。

中南大学由原湖南医科大学、长沙铁道学院与中南工业大学于 2000 年 4 月合并组建而成。原中南工业大学的前身为创建于 1952 年的中南矿冶学院，原长沙铁道学院的前身为创建于 1953 年的中南土木建筑学院，两校的主体学科最早溯源于 1903 年创办的湖南高等实业学堂的矿科和路科。原湖南医科大学的前身为 1914 年创建的湘雅医学专门学校，是我国创办最早的西医高等学校之一。中南大学秉承百年办学积淀，顺应中国高等教育体制改革大势，弘扬以“知行合一、经世致用”为核心的大学精神，力行“向

善、求真、唯美、有容”的校风，坚持自身办学特色，服务国家和社会重大需求，团结奋进，改革创新，追求卓越，综合实力和整体水平大幅提升。

致谢

感谢最先制作出中南大学博士学位论文 LaTeX 模板的郭大侠 @CSGrandeur。

感谢添加本科学位论文样式支持的 @BlurryLight。

感谢帮助重构项目并进行测试的 @burst-bao 以及为独立使用 LaTeX 进行毕业论文写作提供宝贵经验的 16 级的姜析阅学长。

感谢 CTeX-kit 提供了 LaTeX 的中文支持。

感谢上海交通大学学位论文 LaTeX 模板的维护者们 @sjtug 和清华大学学位论文 LaTeX 模板的维护者们 @tuna 给予的宝贵设计经验。

感谢所有为模板贡献过代码的同学们!

参考文献

[1] KNUTH D E, BIBBY D. The texbook[M]. Addison-Wesley Reading, 1984.

[2] LAMPORT L. LATEX: a document preparation system: user's guide and reference manual[M]. Addison-wesley, 1994.

[3] PRITCHARD A, et al. Statistical bibliography or bibliometrics[J]. Journal of documentation, 1969, 25(4): 348-349.

附录 A　附录代码

附录部分用于存放这里用来存放不适合放置在正文的大篇幅内容、典型如代码、图纸、完整数学证明过程等内容。

A.1　堆溢出检测算法

算法 A.1 堆溢出检测算法

1: **if** $\beta \in \mathbb{N}^* \wedge \Delta_\beta = \Delta_{\beta-1} \wedge \beta < S$ **then**
2:　　正常写入
3: **else if** $\beta \in \mathbb{N}^* \wedge \Delta_\beta \neq \Delta_{\beta-1} \wedge \beta \geq S$ **then**
4:　　发生堆溢出
5: **end if**

A.2　KMP 算法 C++ 描述

```
const int maxn=2e5+5;
int nt[maxn];
int aa[maxn],bb[maxn];
int a[maxn],b[maxn];
int n;
//参数为模板串和next数组
//字符串均从下标0开始
void kmpGetNext(int *s,int *Next)
{
    Next[0]=0;
//    int len=strlen(s);
    for(int i=1,j=0;i<n;i++)
    {
        while(j&&s[i]!=s[j]) j=Next[j];
        if(s[i]==s[j]) j++;
        Next[i+1]=j;
    }
//    Next[len]=0;
}
```

```
int kmp(int *ss,int *s,int *Next)
{
    kmpGetNext(s,Next);
//  调试输出Next数组
//  int len=strlen(s);
//  for(int i=0;i<=n;i++)
//          cout<<Next[i]<<" ";
//  cout<<endl;

//    int ans=0;
//    int len1=strlen(ss);
//    int len2=strlen(s);
    for(int i=0,j=0;i<2*n;i++)   //倍长
    {
        while(j&&ss[i%n]!=s[j])j=Next[j];
        if(ss[i%n]==s[j])  j++;
        if(j==n){
            return 1;
        }

    }
    return 0;
}
int main(void)
{
    while(cin>>n)
    {
        memset(a,0,sizeof(a));
        memset(b,0,sizeof(b));
        rep(i,0,n) cin>>aa[i];
        rep(i,0,n) cin>>bb[i];
        sort(aa,aa+n);
        sort(bb,bb+n);
        rep(i,0,n-1){
            a[i]=aa[i+1]-aa[i];
            b[i]=bb[i+1]-bb[i];
        }
```

```
            a[n-1]=360000+aa[0]-aa[n-1];
//              rep(i,0,n) cout<<a[i]<<" ";
//              cout<<endl;
            b[n-1]=360000+bb[0]-bb[n-1];
//              rep(i,0,n) cout<<b[i]<<" ";
//              cout<<endl;
            if(kmp(a,b,nt))
                cout<<"possible"<<endl;
            else cout<<"impossible"<<endl;
        }
        return 0;
    }
```

附录 B　康托尔辩辞录：数学的自由与制约

（录自康托尔：《一般集合论基础》，1883）

数学在其发展中是完全自由的，它只受下述自明的关注所制约，即它的概念既要内在地不存在矛盾，还要参与确定与此前形成的，已经存在着地和已被证明地概念之关系（借助定义贯串起来）。特别地，在引入新数时，数学只遵循：在给出它们地定义时使之具有某种确定性，并且在某些情况下，使之与老数有某种关系，在特定地场合中这种关系一定会使它们（新数和老数）互相区别开来，只要一个数满足这些条件，数学只能而且必须把它看作是存在的和实在的东西，这正是我……关于为什么必须把有理数、无理数和复数看作与有限正整数一样是实在的所建议的理由。

我相信，没有必要害怕，许多人是害怕，这些原则含有对于科学的危险，一方面，实行造出新数的自由必须服从所设计的条件，但这些条件给任意性留下的活动空间是非常小的。而且，每一数学概念在其自身之中也带有必要的矫正物；如果它没有收获也不合适（它的无用很快就会表明这一点），那么它将由于没有成功而被丢弃。另一方面，在我看来，对于数学研究工作的任何多余的限制只会随之而带来更大的危险，由于实际上并没有任何理由可说明它是由科学的本质推断出来的，它的危险就更大了，而数学的本质恰恰在于它的自由。

如果高斯、柯西、阿贝尔、雅可比、狄利克雷、魏尔斯特拉斯、埃尔米特和黎曼总是被束缚而拿他们的新想法去臣服于形而上学的控制，那么，我们今日就不可能为现代函数论的雄伟建筑而高兴，现代函数论的设计和矗立是完全自由的，毫无短视的瞬间目的……。如果福克斯、庞加莱和其他许多杰出的智者受外来影响所包围和限制，我们就会见不到他们带给微分方程论的巨大的推动，还有，如果枯莫尔不是斗胆地（大有仿效者）把所谓的“理想”数引入数论，我们今天也无从去羡慕钦佩克罗内克和戴德金在代数和算术上十分重要和杰出的工作。

因此，如已说明的，数学是要脱离形而上学的桎梏而完全自由地发展 …