科技文献关键短语的抽取和功能分类（草稿）

金笑缘[1]，陈科锜[1]，李昱勇[1]，化柏林[1]
[1](北京大学信息管理系 北京 100871)

摘要:

关键短语自动提取是文本挖掘任务领域中的基本任务，本文提出了短语抽取和短语功能分类的一种联合抽取模式，能更加高效的同时抽取关键短语并进行分类。提出了基于语义功能角色的 TMP 关键短语功能分类模型，该模型合理且完备地对关键短语进行分类并揭示了关键短语对文章主题的贡献的不同维度。在模型算法上，实现了基于 BERT+Bi-LSTM+CRF 的联合多分类序列标注抽取模型，在提取关键短语的同时对短语进行分类。该功能分类方法能够从文档整体对语义层面上的 TMP 分类模型进行很好地刻画。由于标注训练数据较少，本研究所训练的模型还存在提升空间，且训练方法存在标签不平衡问题。总体而言，基于联合抽取方法的 TMP 功能分类为更细粒度的语义抽取提供了新的可行的思路。

关键词: 关键短语；联合抽取；功能分类；序列标注；科技文献

Keyphrases Extraction and Functional Classification in Scientific Texts（DRAFT）

Jin Xiaoyuan[1], Chen Keqi[1], Li Yuyong[1], Hua Boling[1]
[1](Department of Information Management, Peking University, Beijing 100871, China)

Abstract:

Automatic keyphrase extraction is a fundamental task of text mining. In this paper, we proposed a new model for joint phrase extraction and functional phrase classification, hoping to reflect keyphrases with more granularity concerning their contribution to the topic of academic papers. We proposed a TMP functional classification model based on semantic topic, and the TMP model classifies keyphrases reasonably and completely revealing topics in different dimensions. We implemented a multi-category sequence annotation model to jointly extract keyphrases with functional classification. Experiments proved that our functional classification method could well characterize the TMP classification model at the semantic level. Due to the small amount of labeled training data, there is space for improvement in the models trained in this study, and the training methods do not figure out the label imbalance problem well. Overall, we provide new and feasible ideas for finer-grained semantic extraction.

Keywords: keyphrase; integrated extraction; functional classification; sequence annotation; scientific texts

1 背景介绍

关键词和关键短语是图书情报领域中语言处理的重要基础概念。关键词是指揭示文献主题的、有实质意义的语词。对于文献等常见的长文本类型，关键词一般直接来源于文献的标题、摘要、正文。狭义的关键词即揭示文章主题的单个词语，关键短语是狭义关键词的超集，它由一个或多个揭示主题的词语构成的短语，一般科技文献的作者关键词是指关键短语。

关键词和关键短语的抽取在多个学科领域均有重要的应用。在文本挖掘领域被称为关键词抽取（keyword extraction），在计算语言学领域通常着眼于术语自动识别（automatic term recognition），在信息检索领域就是指自动标引（automatic indexing）。由于关键词是表达文件主题意义的最小单位，因此大部分对非结构化文件的自动处理，如自动标引、自动摘要、知识挖掘、自动问答等，都依赖关键词提取进行后续处理。关键词提取是文档自动处理的基础与核心技术。

在英文文献中，由于英文单词语义表达能力有限，关键短语相比关键词往往是更常见的揭示主题的语言单位。一般而言，作者标引的“关键词”也往往以广义的关键词也即关键短语的形式呈现。

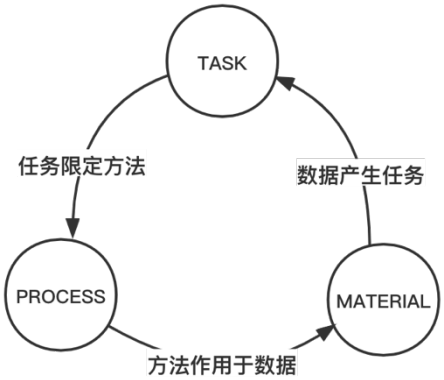
在科学技术领域，存在大量和科技文献在语言模型上基本呈现同分布的未标注关键短语的语料，比如科技新闻、科技政策、专利说明等科技文本，由于这些文本在主题和语言组织方式上和容易获取的大量有标记科技文献具有较高的相似性，基于科技文献的相关自然语言处理模型能够较好地迁移到类似的文本上。因此，尽管几乎所有科技文献都在各大数据库的题录数据中完整地标注了关键短语，但本文研究科技文献语料的已标注特性依然是非常有价值的，基于科技文献的关键短语抽取任务具有重要意义。

除了对关键短语本身进行识别和抽取，对短语的细粒度抽取和分析也具有很高的研究价值。比如对抽取结果进行语义分类、对关键短语间的关系进行识别和分类等，都可以在诸多下游应用场景中起到理解、压缩语意的关键作用。利用细粒度的关键短语信息，可以在科技文本摘要、颠覆性技术识别、研究创新性的抽取等细粒度知识挖掘中提供结构化的端到端信息。

本文重点关注关键短语的分类，在获得短语抽取结果的基础上进一步对科技文本进行细粒度的分析。比较常见的针对语言的细粒度语义分类有词性分类（POS），命名实体识别（NER）等。这些分类或者侧重语法信息，或者关注实体信息，但是并不非常适用关键短语的分类场景，因为关键短语自身具有重要的语义地位信息，而普通的短语分类方法大多缺乏对语义的识别，更不能在语意层面对关键短语的重要语义地位进行细粒度分类。针对这种研究缺失，本文基于科学研究第四范式提出了关键短语功能分类模型 TMP：该模型认为科技文本的关键短语可以完备地分为三类，T 即任务（Task），M 即资料（Material），P 即处理过程（Process），下面的表格对 TMP 模型进行详细的解释。

功能分类	解释
Task	语义上作为驱动研究的任务或者目标
Material	语义上作为支撑研究的资料、数据集等材料
Process	语义上作为推动研究的方法、过程

本文构建了如下的关系图对 TMP 功能模型的三个元素之间的关系及其全局的完备性进行论证。基于数据第四范式的科学研究，首先起源于数据时代产生的大规模数据记录即 Material，在数据驱动下对应的任务被提出即 TASK，任务起源于数据应用过程中的需求或问题，由于数据应用场景具有稳定性和连续性，因此所提出的任务往往具有共性，在给定的技术水平下它们会对应着相似的技术手段即 PROCESS，最终方法手段针对实际应用场景作用于数据并解决任务。该模型对数据驱动的科学研究问题具有普遍性的概括,在科学研究的宏观视角下，它能对研究的整体动机、方法和资料进行完备的概括，而在微观视角下，它同样对数据科学研究中的论证或描述的逻辑链条有完整的概括能力。因此基于 TMP 模型框架对数据科学的主题分析能很好揭示或概括科学文献的主题。



利用 TMP 模型对科技文献中的关键短语进行功能分类，未来的研究人员可以进一步对文章的主题关键短语进行细粒度的分析，比如关系识别和分类，也可以在下游应用中作为文档特征进行文档间的图模型分析等。因此 TMP 模型在科技文本主题分析领域具有广泛的应用前景，特别是基于迁移学习将分类模型推广到普通科技文本的能力。

2 相关工作

2.1 无监督抽取

在深度学习得到广泛应用之前，无监督算法一直是知识抽取类任务的主流算法。文本的无监督抽取方法主要有以下几种：首先是基于统计特征的抽取，比如使用 TF-IDF，这种经典的算法主要基于文本的词频分布特征，由于统计特征的计算复杂度较低，基于统计特征的算法在信息检索领域得以广泛使用，能很好的满足信息检索系统低延迟的用户体验需求。此外还有基于图模型的抽取算法：如基于共现相关关系结构的 TextRank 算法，它以词为结点，词之间的共现关系作为边构造文档的图模型；也有基于主题特征的 LDA 算法，它需要指定主题的个数并以词向量的方式来对主题进行表示。为了获得更好的对语义的表示，还可以基于词嵌入表示进行抽取，比如使用 WordVec 等词嵌入工具，结合聚类或图模型进行计算。最后还有基于语言模型的抽取：比如使用 N-Gram 模型的抽取算法，它主要是通过语言模型的思维来对文档进行词的建模，可以满足短语抽取的需求，常常结合 TF-IDF 等统计方法进一步计算。

2.2 有监督抽取

无监督算法由于不涉及深度学习的计算，因此大多模型具有快速简洁的特点，但同样因为缺少对语言上下文的深度的刻画，早期基于图结构模型或者统计特征的无监督算法，对于语言顺序和语义特征的提取比较有限。而且无监督算法由于模型的静态性不能很好地处理短语的问题，往往只能实现单个关键词或固定长度的关键短语抽取。

有监督算法主要是基于深度学习的算法。自从深度学习在算力和模型上得到关键性的突破和发展之后，关键短语处理领域逐渐出现较多使用深度学习进行有监督抽取的方法。尽管深度学习模型众多、参数各异，但基本都采用了序列标注的经典框架来实现关键短语抽取任务。序列标注是一种灵活的、能够抽取不定长度文本序列的算法，其最本质的特点就是使用位置标签 BIO 来给每一个单词进行标注。这一思路最初来自自然语言处理领域的重要领域命名实体识别（Named Entity Recognition，以下简称 NER），虽然关键短语在训练数据的语言学统计分布上和命名实体存在区别，但是在模型的描述结构上两者具有很高的相似性，特别是短语长度的不确定性，以及分类标签的可扩展性。

序列标注算法模型的结构一般分为两个部分，底层由表示学习模型对输入的语句进行词向量的嵌入，下游使用循环神经网络（Recurrent Neural Network，以下简称 RNN）或者是 RNN 结合条件随机场（Conditional Random Field）对词向量进行序列标注标签的分类。其中最为经典的序列标注模型是百度提出的 Bi-LSTM+CRF 模型，这种序列标注模型结合了 LSTM 的序列处理记忆能力和 CRF 的序列规则限制能力，在序列标注上取得了不错的效果。

2.3 功能分类

关键短语的功能分类是科技文本知识挖掘的重要手段。上述早期的关键短语抽取方法一般不能很好的结合关键短语分类，往往将关键短语的功能分类作为关键短语抽取的派生任务。现有的研究关于关键短语的功能特征提出了不同的理论模型，虽然大多研究致力于通过功能分类使得关键短语对文章主题的概括能力在更细粒度的维度上有所区分，但是对于究竟使用怎样的维度来定义功能分类至今没有达成一致。朱惠从“过程-问题”的视角重构了信息学领域的理论和技术，重点关注解决信息问题的方法。本文中提出的对关键短语的 TMP 功能分类也是对“过程-问题”模式的合理扩展。王晓光构建了 SAO 本体 (Scientific paper Argumentation Ontology)，从文章论证结构的视角揭示了文章的关键内容。其结构同样与 TMP 有共同之处，这也证实了 TMP 结构的合理性。陆伟认为将关键短语分为研究问题和研究方法，这与本文提出的 TMP 模型中的任务和过程的内涵是相似的。

2.4 研究的创新之处

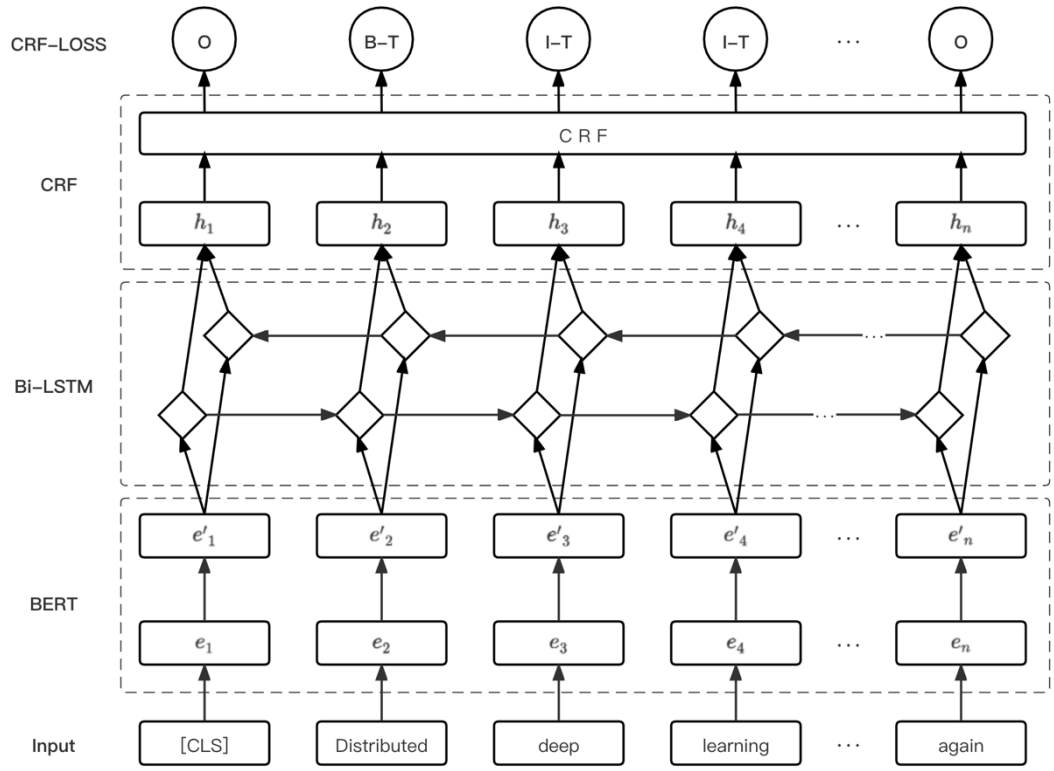
目前的关键短语自动提取技术已经满足了某些领域的基本需求，但对关键短语自动提取的研究仍处于起步阶段，相关理论基础不完善，缺乏明确有效的模型以及可行的统一标准。而且，许多现有的研究缺乏对深层和隐藏语义信息的考虑。

本文的主要贡献有：（1）提出了科技文本关键短语的 TMP 功能分类模型，组织标注了相关的数据集；（2）在上述功能分类数据集上，提出并训练了基于 TMP 分类序列标注的联合抽取-分类模型，基于大量文献进行了预测分析。

3 方法

传统的流水线方法短语的功能分类（Pipeline，即先完成上游任务，然后以上游输出作为下游输入进行其他的阶段），也就是先抽取关键短语，在抽取的基础上进一步尝试短语分类，由于短语抽取和功能分类不能共享参数，分类工作无法基于抽取结果进行。因此需要重新为各个短语构建特征工程并训练分类器（比如使用 PoS 词性标签等特征），经过实验发现这种流水线的独立流程方法效果较差，一方面因为分类器所使用的特征不能像序列标注等循环神经网络结构那样提取上下文的信息，另一方面也因为分阶段的训练而导致误差不断累积。

本文采取了基于 BERT + Bi-LSTM + CRF 的序列标注模型的联合训练方法进行功能分类。通过序列标注的设计将关键短语的抽取和关键短语的分类融合为一个模型并共用参数，模型的架构图如下。



模型的第一层是基于 BERT（Bidirectional Encoder Representations from Transformers）的词嵌入表示。BERT 是 Google 提出的基于大规模语料的预训练模型，使用词掩码和句子对预测的预训练任务进行联合训练，最终生成动态句子和词嵌入的预训练模型。BERT 巨大的参数量和训练数据规模使得基于 BERT 微调的语言处理具有很好的词汇表示能力。本实验使用 Google 发布的预训练模型，作为词嵌入来微调下游任务，该预训练模型的隐藏层维度是 768，注意力头个数是 12 个，编码器层数为 12 层，总计 110M 参数。

模型的第二和第三层是双向长短期记忆循环网络模型（Bi-directional Long Short Term Memory，以下简称 Bi-LSTM）和条件随机场（Conditional Random Field，以下简称 CRF）进行序列标注，这两层模型的融合最初由百度提出，用于序列标注算法。

Bi-LSTM 层由前向 LSTM 与后向 LSTM 组合而成，继承了 LSTM 对序列信息处理的优点，LSTM 通过在朴素循环神经网络中加入遗忘门、记忆门、输出门结构，更好地对输入序列的特征进行选择性的记忆或遗忘。双向的 LSTM 层通过隐藏层输出的叠加实现融合，不同于单向的 LSTM，Bi-LSTM 能够同时关注并建模输入句子在两个方向的序列信息，实现更好的长短期记忆效果。

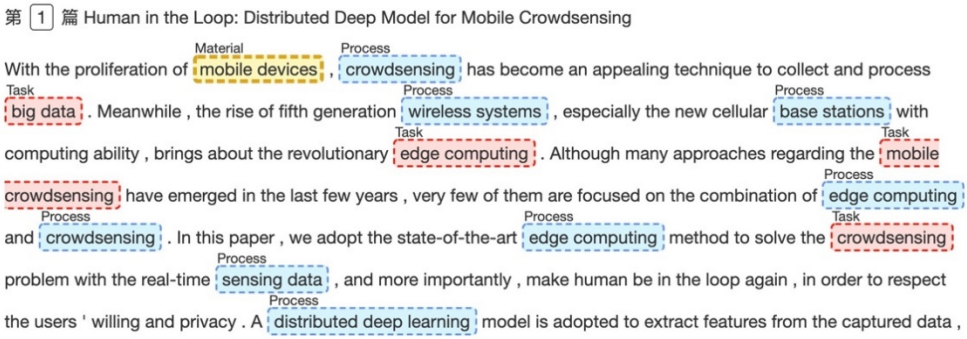
CRF 层作为概率图结构模型，基于条件概率转移矩阵来刻画序列标注的最优路径，能很好对需要学习的序列标注标签特征进行约束性建模。特别地，CRF 的作用主要体现在防止学习得到的模型会预测出不合顺序规范的标注结果，比如 I 标签出现在 B 标签前，或者 B 和 I 不一致的情况。

在模型的预测层，使用带分类信息的序列标注方法来对分类抽取问题进行建模。在 TMP 的分类体系下，共有 B-T，B-M，B-P，I-T，I-M，I-P，O 这 7 个标签。损失函数为基于 CRF 层 Viterbi 解码方式的 CRF-Loss，CRF-Loss 损失函数通过真实的序列标签路径除以所有序列标签路径的概率和来计算，其中序列标签路径指的是 CRF 层输出的七种可能标签按输入顺序的排列。

4 实验和结果

4.1 数据集

本研究以“(big data) or (machine learning)”为检索式，以主题为检索字段，时间限制为 1990 年-2019 年，在 Web of Science 核心合集中进行文献检索，返回文章 230709 篇。下载每篇文章的标题、作者、摘要、关键短语和年份信息，构成原始数据集。为了获得保证后续研究中更好的数据质量，对元数据信息的缺失情况进行筛查，特别关注作者标注关键短语字段的质量，过滤掉关键短语过少或者在摘要中出现次数过少的文章。首先选择了 250 篇关键短语数量较多的文献作为训练数据，并随机选择剩下文献中的 70000 篇用于后续文献分析研究。



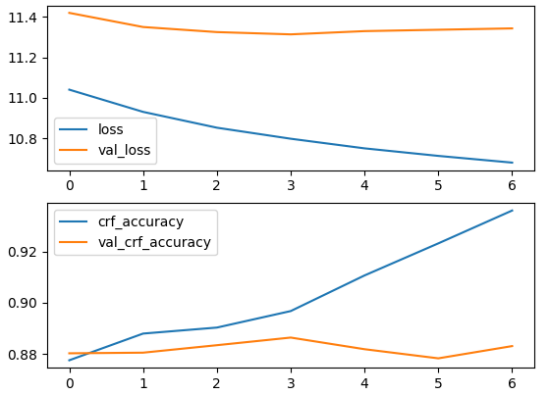
由于对关键短语的功能分类模型进行了更加严谨和完备的定义，本文基于上述 250 篇文献进行了人工数据标注。为了帮助标注专家更好的快速定位摘要整段中的潜在关键短语并进行高效的分类，标注系统在作者关键词的基础上，使用无监督学习、模式匹配等启发式规则对摘要中潜在的候选关键短语进行扩

充。标注系统允许标注专家选出关键短语，并按照 TMP 分类模型进行分类标签的标记。本文作者开发了一个通用的基于浏览器前端的注释工具，如上图所示。相比 Doccano 等流行的序列标注工具，设计了更加符合视觉特点的交互界面，并且能够预先高亮待分类的关键短语。数据标注的结果如下表所示，可以发现 MATERIAL 标签明显比 TASK 和 PROCESS 少。

标签	总数（共 250 篇）	每篇文章平均数量
TASK	1029	4.12
MATERIAL	638	2.55
PROCESS	1543	6.17

4.2 模型训练

将训练数据集以 8-1-1 的比例进一步划分为训练集、验证集和测试集。尽管样本存在不平衡问题，但 BERT 强大的语义表示能力、结合 Bi-LSTM 的上下文记忆能力和 CRF 的规则学习，使模型训练仍然能够从训练集中提取语义功能分类信息。模型在一块 Tesla P100 GPU 进行训练，训练曲线如下图所示。



基于WoS的手动标注数据集训练学习曲线

验证损失在 7 个 EPOCH 后达到收敛并提前停止，根据验证集准确率选择了 EPOCH=6 的模型，准确率 0.883。表 3 显示了对测试集基于 Seqeval 包的多分类评估结果。Seqeval 通过梳理 BIO 信息将序列标注格式转换为短语标签格式，并对每个标签进行评估，提供 F1 分数和多标签分类的微/宏观平均值。

	precision	recall	f1	support
MATERIAL	0.1290	0.1000	0.1127	40
TASK	0.1494	0.1262	0.1368	103
PROCESS	0.4345	0.4650	0.4492	157
micro avg	0.3147	0.3000	0.3072	300
macro avg	0.2959	0.3000	0.2971	300

测试集 seqeval 解析结果

获得分类模型后，为进一步验证抽取、分类的效果，并探索功能分类在实际研究中的用途，对抽取阶段使用的 70000 篇 WoS 文献摘要进行短语抽取和功能分类，用于后续对文献从细粒度的功能角色角度进行分析。模型的预测统计结果如下表所示，平均每篇摘有 1.5 个短语因为 BIO 序列的一致性问题解析失败，抽取并分类成功的三种短语的比例和训练数据比较接近，其中 PROCESS 比例偏大。

	平均抽取数量	平均解析失败
MATERIAL	0.53	
TASK	1.40	
PROCESS	5.45	
总计	7.38	1.50

70000 篇文献预测结果

4.3 文档分析结果

为了更好地多维度分析抽取及分类结果，统计了 TMP 分类结果中的高频率短语（每个类别的前 15 个高频短语），词频数据如下表所示。可以发现高频词比较笼统，但是基本上符合所属类别标签的本质。借此可以分别从 Task、Material、Process 三个维度对文档整体进行简单的主题识别，比如对所测试的 70000 篇 WoS 文章，他们的主要 Task 有聚类、健康、城市、智能、社交等主题；而 Material 则是各种数据集；Process 主要有机器学习、数据挖掘、神经网络等方法，这与获取数据集的检索方法是基本一致的。

Task	数量	Material	数量	Process	数量
clustering	938	data	1353	machine learning	20968
data	626	big data	732	data mining	9943
health	554	gene expression	431	support	4149
urban	439	large	422	clustering	3669
smart	438	datasets	346	neural network	3235
human	422	data sets	325	feature selection	2532
social	370	data set	281	classification	2515
social media	364	training data	255	neural networks	2377
power	351	training	242	machine	1963
network	343	dataset	226	vector machine	1919
cancer	336	big	190	learning	1850
mobile	330	clinical data	158	Support	1841
traffic	327	sensor data	152	big data	1715
online	321	image	147	deep learning	1654

70000 篇文献的预测结果的高频词统计

除了上述分析，基于 TMP 分类的抽取模型还可以结合文献的元数据进行更多丰富的文献研究分析。比如在更大规模的预测数据支撑下，可以利用年份信息探索不同年份数据科学研究的重点任务和方法迁移情况；也可以利用引文信息探索引文网络框架下作者目标任务或者数据集的相似性，并据此实现网络预测分析等。以上多样的文献分析任务都是 TMP 分类模型适合的应用场景。

5 结论

本研究提出了 TMP 关键短语功能分类模型，它对揭示文本主题有重要作用，并可用于广阔的下游任务。虽然由于样本严重不平衡，序列标注结果指标相对较低，但仍然可以看到 TMP 功能分类在细粒度知识抽取领域的巨大前景。在训练方法和标注数据的规模和质量上，还需要未来更多的改进和探索。比如进一步解决标签不平衡问题，特别是削弱 Outside 标签对模型学习的影响，可以在模型训练中设计特殊的损失函数用于功能分类。在 TMP 分类模型的基础上，后续还计划结合元数据来研究更细粒度的文本主题关系提取，为现实世界的场景提供更有意义的应用支持。