计算机科学
COMPUTER SCIENCE

联合知识图谱和预训练模型的中文关键词抽取方法

姚奕，杨帆

引用本文

姚奕，杨帆. 联合知识图谱和预训练模型的中文关键词抽取方法[J]. 计算机科学, 2022, 49(10): 243-251.

YAO Yi, YANG Fan. Chinese Keyword Extraction Method Combining Knowledge Graph and Pre-training Model[J]. Computer Science, 2022, 49(10): 243-251.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800176

联合知识图谱和预训练模型的中文关键词抽取方法

姚 奕 杨 帆
陆军工程大学指挥控制工程学院 南京 210007
(yaoyi226@aliyun.com)

摘 要 关键词表征了文本的主题,是文本概念和主题的凝练。通过关键词,读者可以快速了解文档表达的主旨和思想,从而提升信息检索效率;此外,关键词抽取也可以为自动摘要、文本分类提供支撑。近年来,自动抽取关键词的研究引起了广泛关注,但如何精准地抽取文档的关键词仍是一个挑战。一方面,关键词是人们主观的认识,判断一个词是否是关键词本身具有主观性;另一方面,中文词汇往往具有丰富的语义信息,单纯依赖传统统计特征和主题特征难以准确提炼文本所表达的主旨思想。针对中文关键词抽取中存在的准确率低、信息冗余和信息缺失等问题,提出了一种联合知识图谱和预训练模型的无监督关键词抽取方法。该方法首先利用预训练模型进行主题聚类,并通过一种以句子为单位的聚类方法保证最终选取的关键词对全文内容的覆盖度;同时,通过知识图谱进行实体链接,以此实现精准分词及歧义消除;然后,根据主题信息构建语义词图,并以此为基础计算词语间的语义权重;最后,通过加权的 PageRank 算法进行关键词排序。在 DUC 2001 和 CSL 两个公开数据集和一个单独标注的 CLTS 数据集上,以预测结果的准确率、召回率及 $F1$ 值为指标进行对比实验。实验结果表明,该模型相比多种基线方法,准确率均有所提升,在 CLTS 数据集上与传统统计方法 TF-IDF 相比 $F1$ 值提高了 9.14%,与传统图方法 TextRank 相比 F1 值提高了 4.82%。

关键词:关键词抽取;知识图谱;句嵌入;聚类;图算法;预训练模型
中图法分类号 TP391

Chinese Keyword Extraction Method Combining Knowledge Graph and Pre-training Model

YAO Yi and YANG Fan
College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China

Abstract Keywords represent the theme of the text, which is the condensed concept and content of the text. Through keywords, readers can quickly understand the gist and idea of the text and improve the efficiency of information retrieval. In addition, keyword extraction can also provide support for automatic text summarization and text classification. In recent years, research on automatic keyword extraction has attracted wide attention, but how to extract keywords from documents accurately remains a challenge. On the one hand, the keyword is people's subjective understanding, judging whether a word is a keyword itself is subjective. On the other hand, Chinese words are often rich in semantic information and it is difficult to accurately extract the main idea expressed in the text by solely relying on traditional statistical features and thematic features. Aiming at the problems of low accuracy, information redundancy and information missing in Chinese keyword extraction, this paper proposes an unsupervised keyword extraction method combining knowledge graph and pre-training model. Firstly, topic clustering is carried out by using the pre-training model, and a sentence-based clustering method is proposed to ensure the coverage of the final selected keyword. Then, the knowledge graph is used for entity linking to achieve accurate word segmentation and semantic disambiguation. After that, the semantic word graph is constructed based on the topic information to calculate the semantic weight between words. Finally, keywords are sorted by the weighted PageRank algorithm. Experiments are conducted on two public datasets, DUC 2001 and CSL, and a separate annotated CLTS dataset, the prediction accuracy, recall rate and $F1$ score are taken as indicators in comparative experiments. Experimental results show that the accuracy of the proposed method has improved compared with other baseline methods, $F1$ value is increased by 9.14% compared with the traditional statistical method TF-IDF, and increased by 4.82% compared with the traditional graph method TextRank on CLTS dataset.

Keywords Keyword extraction, Knowledge graph, Sentence embedding, Clustering, Graph-based algorithms, Pre-trained model

到稿日期:2021-08-19 返修日期:2021-12-06
基金项目:军事类研究生资助课题(JY2019C078)
This work was supported by the Military Postgraduate Research Project(JY2019C078).
通信作者:杨帆(ivan10240@163.com)

1 引言

在信息时代，随着各类文本数据的爆炸性增长，许多新闻报道、博客之类的文本数据没有标注关键词，这给读者的阅读和信息搜索带来了不便。因此，研究自动关键词抽取技术，从文本中识别或提炼出与原文所表达意义最相关的一些词或短语，不仅可以将庞杂的文本数据分门别类，还能方便读者快速了解文档所表达的主要概念与核心思想，有效提升在海量数据中搜索有用信息的效率。此外，关键词抽取技术也被广泛应用于文本分类、文本聚类、自动摘要等方面。

相比自然语言处理领域的其他任务，关键词抽取任务始终没有取得突破性的进展，目前仍没有一种很好的、通用的关键词抽取方法被提出。究其原因，关键词自身带有一定的主观色彩，而语义表征更是一个仁者见仁、智者见智的问题，就像一千个读者有一千个哈姆雷特。什么是关键词？为什么有些词语可以作为关键词被抽取出来表征文本主题，而另一些词语应当被忽略？文献[1]对关键词的几种典型定义做了归纳，文献[2-4]对关键词的特性做了归纳。总体而言，关键词是简明而准确地描述文本讲述主题以及中心思想的词或短语。关键词应满足以下 5 个特性。

(1)可读性(Readability)：关键词本身应是有意义的短语[2-3]。例如，“知识图谱”是一个有意义的短语，但是“联合”不是。

(2)相关性(Relevance)：关键词应准确表征原文的主旨及基本思想，不能是与原文无关或关联性弱的词[2-4]。

(3)覆盖性(Coverage)：关键词应对文档主题有较好的覆盖，不能只集中在文档某个主题而忽略了文档的其他主题[2-4]。

(4)一致性(Coherence)：关键词之间应存在语义上的联系，逻辑上保持一致，所表述的内容能够形成一个逻辑统一的整体[3-4]。

(5)简洁性(Conciseness)：关键短语的数量是有限的，并且关键短语集不应包含任何冗余信息，能简洁明了地体现文本主旨[3-4]。

大部分研究通过特征定义来抽取关键词以满足上述特性。文献[5]将现有研究所提出的有效特征归结为候选关键词统计特征、词图结构特征、主题特征、词嵌入向量四大类。这些特征大多基于对目标文档中局部文本信息的分析，多数情况下仅支持作为关键词所需满足的 2～3 种特性。然而，在近几年的研究中，基于特征的关键词抽取方法在性能上基本达到了瓶颈。Yu 等[6]认为文档本身可能无法为关键词抽取任务提供足够的信息，并提出了一种基于 Wikipedia 的关键词抽取方法 WikiRank，这与早期的 CommunityRank[7] 和 SemanticRank[8] 类似，利用 Wikipedia 页面链接作为外部知识。不同的是，早期的研究只是以粗粒度的统计方式使用链接信息，不可避免地引入了许多不相关的关系；而 WikiRank 将关键词的抽取建模为一个优化问题，并给出了相应的解决方案和剪枝方法以降低复杂性。

相比过往使用外部知识库进行背景扩充，知识图谱的结构化知识、实体的关联关系具有更得天独厚的优势——引入的知识与全文语义更相关，引入噪声更少。在一些专业性较强的语料(如军事和医学等领域的专业文章)中，关键词往往只提及一次，仅依赖文本内的信息去提取关键词显然是十分困难的。在这种情景下，外部的知识结构体系就显得尤为重要。然而，利用外部知识的基于图的研究在中文领域稍显空白。鉴于此，本文提出了一种基于中文百科知识图谱 CN-DBpedia[9] 的中文文档关键词抽取方法 CnKGRank。针对中文分词过程中存在的歧义问题，设计了一种基于实体链接和词法规则匹配的候选关键词提取方法，通过对比不同聚类方法以及传统词嵌入聚类和 BERT 模型对关键词抽取性能的影响，验证了 BERT 模型在中文文本主题发现上有更好的效果。此外，针对中文关键词抽取数据集匮乏且标准不统一的问题，翻译了 DUC 2001[10] 数据集，并在中文长文本新闻摘要数据集 CLTS[11] 上进行标注，构建了基于 CLTS 的中文关键词抽取测试集，为关键词抽取在自动摘要中的研究提供了数据集支撑。

2 相关工作

近年来，多种深度学习技术被融入到自然语言处理领域的各项任务中，并表现出了优异的性能。对于关键词抽取任务而言，目前有监督的方法一般通过机器学习模型，使用一个已经标注好关键词的数据集进行训练，将关键词抽取任务转化为分类或序列标注任务。如 Duan 等[12]将统计特征融合到 LSTM 模型中，有效完成了关键词抽取任务。这类方法的效果虽然比大多数无监督的方法好，但是过于依赖语料集，泛化能力弱，难以应用到实际的自动化抽取场景中。而无监督的方法不需要人工标注大量语料集，利用文本本身的结构特征或语义信息来进行关键词抽取仍是目前的主流研究方向。按技术手段，无监督的关键词抽取方法一般分为基于统计的方法、基于主题的方法和基于图的方法等。基于统计的方法以词频、词性、位置等统计特征为指标设置单词的权重，这种方法虽然简单易用，但是准确率不高，在不同的数据集上表现不稳定。因此，本节主要介绍另两种方法。

2.1 基于主题模型的方法

最早的主题概率模型 PLSA(Probability Latent Semantic Analysis)[13] 基于这样一种思想：通常一篇文档由多个主题混合而成，而每个主题都是词汇上的概率分布，可以通过一定的概率选取出某个主题，再通过概率从该主题中选取某个词语，从而得到关键词。Blei[14] 在 PLSA 的基础上加入 Dirichlet 先验分布，提出了 LDA(Latent Dirichlet Allocation)模型，实现了 PLSA 模型的突破性发展。此后，一些变种方法[15-16]相继被提出。这些基于主题模型的方法虽然增大了目标文本的中心主题词被识别为关键词的概率，确保了关键词的覆盖度，但是提取的关键词比较宽泛，有时不能很好地反映文档主题，并且计算复杂度较高，抽取结果对训练数据集的依赖较大。

另一种基于主题的关键词抽取方法则是使用词聚类的方法。早期聚类方法使用文本内的共现关系或者外部知识库(如 Wikipedia)来度量词与词之间的相似度。一些工作[17-18]基于文档内部的统计信息来计算相关性，实现对文档词汇的聚类。但是受文档本身信息的限制，内部统计特征往往无法

为发现文档主题提供足够的信息。随着深度学习技术的发展，Word2Vec[19]和 Glove[20]等语言模型为文本相似度计算带来了新的希望，这些模型在大规模语料数据上利用词的上下文信息进行训练得到词嵌入，增强了单词间的共现关系，语义信息更加丰富，能更准确地计算出词语之间的语义相似度。Wang 等[21]首次将词嵌入应用到关键词提取中，他们提出的 WordAttractionRank 方法在维基百科数据集上计算词嵌入，然后通过加权词嵌入和局部统计特征来计算单词的信息量和得分，在 3 种数据集上均具有好的表现。此后，基于词嵌入的聚类被广泛应用于无监督的关键词抽取中，如 Key2Vec[22]和 WeRank[23]等。这些方法都是先训练模型生成词向量，然后通过计算词向量之间的距离来表示两个词之间的语法，并通过语义之间的相似性来对词语进行排序，最终选取关键词。

2.2　基于图的方法

基于图的方法是当前广为流行的一种方法。Quillian[24]最早提出了语义网络的思想，认为文本中词汇最根本的联系在于语义层，可以依据语义联系将词汇进行连接，从而构建文本的语义网络。基于此，Mihalcea 等[25]将用于计算网页重要程度的 PageRank 算法应用到文本领域，开创了基于图的关键词抽取方法的研究。不同于 PageRank 算法，TextRank 算法构建了有权重的图，所有结点权重分值计算完后，获取排序最高的几个词作为关键词。这种方法无须训练数据，具有很强的适应和扩展能力，并且在一些数据集上的性能可以与有监督的方法相媲美；但是，以滑动窗口作为词共现依据，没有考虑词汇之间的语义特征，也忽略了上下文信息。

为此，Wan 等[26]提出了 ExpandRank 方法。该方法利用少量的近邻文档集来提供更多的知识，辅助词图构建，扩充了词图的信息量，使提取效果得到了显著提升。此后，一些结合统计特征与外部信息的方法相继被提出，如基于主题聚类的 TopicRank[27]、将单词在文本中的位置信息和出现频率融入到词图中的 PositionRank[28]，以及将候选关键词及其主题关系存储在多部图中以加强主题多样性的 MultipartiteRank[29]。这些方法都取得了一定的进步，但在实际应用场景中仍然没有达到人们的期望。大量引入外部的信息可能会淹没文本中单词的真正含义，造成信息过载的问题。为了避免引入额外的噪声，人们开始考虑使用结构化的外部数据。Shi 等[30]首次提出基于知识图谱的关键词抽取模型，他们先基于词嵌入对候选短语进行聚类，同时建立候选词与知识图谱中实体的映射图，然后利用改进的 PPR（Personalized PageRank）算法计算候选词的分数，从而选出关键词。

上述方法在英文文本上均取得了较好效果，但无法直接应用到中文领域。一方面，中文的词语表达与英文不同，词语之间没有明显的分隔界限，并且新词、短语、成语的普遍使用给分词以及短语识别带来了困难；另一方面，中文往往一词多义，甚至一字多义，语义关联与歧义消解也是中文文本处理的难题。而知识图谱结构化的知识信息与完善的语义关联关系，在扩充文本信息，帮助机器理解文本的同时，能有效避免噪声的引入。因此，在关键词抽取任务中引入知识图谱，恰好能在一定程度上解决上述两个难题。

3　模型方法

3.1　模型描述与定义

为了从中文文本中获取满足上述 5 个特性的关键词，本文提出了联合知识图谱和预训练模型的词语排序模型 CnKGRank，其全局结构如图 1 所示。CnKGRank 主要分为 4 个主要步骤：句聚类、实体链接、词图构建和关键词排序，其对应关键词抽取的从句子到候选词再到关键词的过程。更正式地说，给定一个中文文本 D，将其处理为集合$\{S,W\}$，S 表示 D 中所有的句子，W 表示组成 D 的所有词和短语。关键词抽取任务是从 W 中选取一组能全面覆盖文本 D 所表达的概念和主题的词或短语 K。

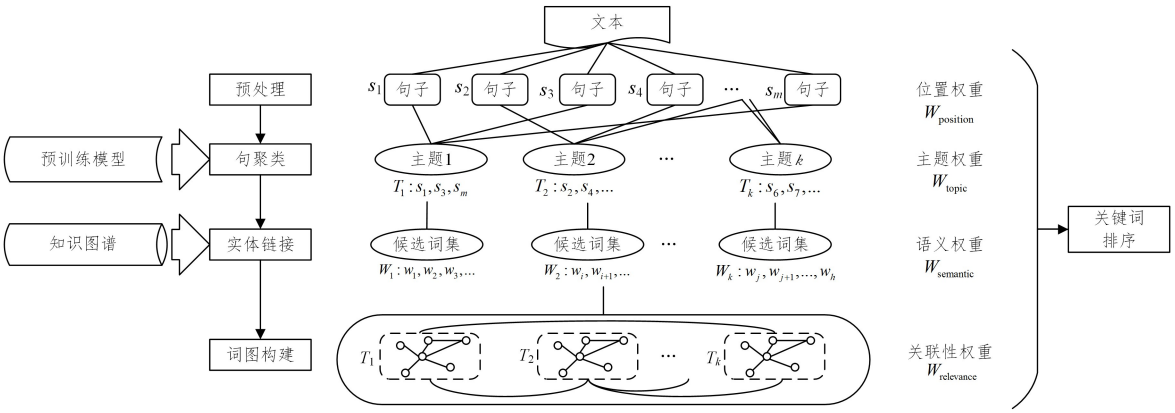

图 1　CnKGRank 模型的全局框架

Fig. 1　Global framework of CnKGrank model

首先，通过文本预处理将文本分割为句子，并记录每个句子的位置信息。然后，利用预训练模型生成每个句子的句向量，通过聚类算法将句子分为多个簇，每个簇代表一个主题。聚类是将句子在向量空间中按语义距离分类，满足覆盖性需求。此后，对主题中的每个句子进行分词，并利用已有知识图谱进行实体链接，按词性匹配规则得到候选关键词。这样处理的好处在于：1）解决了中文分词中存在的歧义问题，满足可读性需求；2）引入了候选词在知识图谱中的背景信息，满足相关性需求。得到候选关键词后，每个主题内部进行词图构建，同时主题之间也依据一定权重进行连接，根据图中心性算法

计算出主题的排名，最后依据主题排名以及主题内部候选关键词的排名计算出全局候选关键词排名，确保一致性与简洁性。

3.2 句聚类

此前的研究大多以词为单位进行聚类，将文本划分为不同词集合的簇，每个簇代表一个主题。然而在实际语言表达中，句子才是语义表达的基本单位，因为不同的词在不同的句子中可能会有不同的语义。例如，“我要吃苹果”和“我用苹果玩游戏”中的“苹果”就有不同的语义，前者表示一种水果，而后者表示一种电子产品，使用相同的词向量来表示这两个句子中的“苹果”会造成聚类的误差。此外，传统 Word2Vec 和 Glove 等词向量模型受限于词汇表，若训练模型中没有涵盖某个词语，则不能准确得到该词的嵌入向量。相比之下，使用 BERT 模型直接生成句向量可以理解上下文的语义，排除了词向量加权引起的误差，在表示语义相似性上更为准确。

因此，本文借助开源的 SimBERT 模型[1)]来生成句向量，然后根据语义距离进行聚类，将语义相关性高的语句划分为一个簇，从而体现文本的主题信息。SimBERT 是融检索与生成于一体的 BERT 模型，由 2200 万相似句组训练得到，在句子相似度任务 STS-B 上有更好的效果[31]。衡量语义距离的常用指标有欧氏距离、余弦相似度、Jaccard 相似系数等。文献[18]证明了余弦相似度可以有效地量化高维数据（如词嵌入、句嵌入）之间的语义相似度，因为向量的方向比大小更重要，因此本文使用余弦相似度来衡量句子之间的语义距离。

对于文本 D，首先利用正则表达式匹配中文的分句符号，如“，”“。”“……”“！”“？”等，将其处理为句子的集合 $S=\{s_1, s_2, \cdots, s_m\}$，$m$ 是文本中的句子总数。然后依次将句子作为输入，利用 SimBERT 模型生成句向量 $\vec{\boldsymbol{S}}=\{\vec{\boldsymbol{s}}_1, \vec{\boldsymbol{s}}_2, \cdots, \vec{\boldsymbol{s}}_m\}^{(n)}$，其中 n 是向量的维度，并依据式(1)计算句向量之间的余弦相似度。

$$Score(\boldsymbol{A},\boldsymbol{B})=\frac{\sum_{i=1}^{n}\boldsymbol{A}_i\times\boldsymbol{B}_i}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{A}_i)^2}\times\sqrt{\sum_{i=1}^{n}(\boldsymbol{B}_i)^2}} \tag{1}$$

其中，$\boldsymbol{A},\boldsymbol{B}$ 代表维度为 n 的向量，$Score(\boldsymbol{A},\boldsymbol{B})$ 计算了向量 $\boldsymbol{A},\boldsymbol{B}$ 之间的余弦相似度。此后再经过聚类算法，将集合 S 分为多个主题的集合 $T=\{T_1, T_2, \cdots, T_k\}$，$T_i(i=1,2,..,k)$ 表示包含一个或多个句子的主题。由于 BERT 模型生成的向量往往是高维的，其在向量空间的分布密度很高，传统的聚类方法不能很好地满足聚类需求。基于图论的谱聚类算法将数据看作空间中的点，通过切图的方式进行聚类，恰好符合句向量在向量空间的分布特征。

对于集合 S 的切图，其目标是将图切成相互没有连接的 k 个子图，子图的集合为 $T=\{T_1, T_2, \cdots, T_k\}$，使其满足 $T_i \cap T_j=\varnothing$ 且 $T_1 \cup T_2 \cup \cdots \cup T_k=S$。对 k 个子图的集合，定义切图的损失函数为 $Ncut$：

$$Ncut(T_1, T_2, \cdots, T_k)=\frac{1}{2}\sum_{i=1}^{k}\frac{CutW(T_i, \overline{T}_i)}{vol(T_i)} \tag{2}$$

其中，$CutW(T_i, T_j)$ 为 T_i 和 T_j 之间的切图权重；$\overline{T}_i$ 为 T_i 的补集；$vol(T_i)$ 为子图权重，用来标示指示向量，通过指示向量可以计算出优化的目标函数。

此外，在多数任务中，降维不但不会降低效果，反而会带来效果上的提升[32]，并且降维可以减少聚类所需的计算量，加快处理速度。因此，在聚类过程中，将高维的词向量降至低维。

3.3 实体链接

实体链接（Entity Linking，EL）指对于给定的一个中文文本（如微博、新闻等），将其中的实体指称项与给定知识图谱中对应的实体进行关联[33]。通过实体链接可以找出句子中的命名实体，并且可以通过知识图谱的语义信息解决歧义问题，提高分词的准确率。实体链接一般包括 3 个主要步骤：指称项提取、候选实体获取和实体消歧。

3.3.1 指称项提取

指称项提取，即在给定句子中抽取出相应的指称项，主要包含命名实体（如机构名词、人名、地名、药物等）。目前命名实体识别的研究主要是基于深度学习的方法，神经网络模型能够自动学习句子特征，相较于其他方法表现更加优异。然而在深度学习的方法中，训练数据很大程度上决定了模型的精度，而缺少训练数据对中文实体识别带来了很大的困难。

此外，不同于英文文本，中文文本的词语之间没有明显的界限，同一句话使用不同的分词方法也会得到不同的分词结果。如例句中的“乔家大院”就可能被分为“乔家大院”和“乔家/大院”，而显然第二种会造成歧义。不同的分词结果影响着句子的语义表达效果，这对后续关键词抽取有着较大影响，这也是目前中文关键词抽取研究中的一个难点。尽管目前有许多成熟的分词工具，如 jieba 和 THULAC[34] 等，但是这些方法都存在一定的不足。jieba 分词结果较快、较全，但是在未登录词识别上存在缺陷；THULAC 虽然准确率高，但是仅支持分词功能，不能进行词性标注。一些方法通过添加用户自定义词典提升分词效果，但是这并不能满足实际应用的需求，因为各个领域的专有名词数量巨大，构建并维护这样一个字典十分困难。而目前流行的知识图谱正好满足这一需求，基于庞大百科数据构建的知识图谱覆盖了大部分的实体概念，并提供了其标签属性信息，为中文实体消歧提供了数据支撑。为了解决数据集依赖和中文分词问题，并降低模型复杂度，本文提出了一种基于知识图谱的实体识别与消歧方法。

对于各个主题中的句子，我们借助 jieba 的分词及词性标注功能，将 T_i 处理为带词性的词集 $t_i=\{[w_1:N], [w_2:A], \cdots, [w_h:VN]\}$，其中，$w_1, w_2, \cdots, w_h$ 是分词后得到的词，N，A，VN 为词性，h 是文本分词后得到的总词数。在实际应用中，绝大多数关键词都是名词或带有形容词的名词短语，基于 POS 标签和 N-grams 是最好的候选关键词选取方法，可以通过词性标签的特定模型直接从现成的名词短语中获得候选词。根据中文用词习惯和词性匹配原则，我们设计了 4 种词法规则提取出指称项集合 $CE=\{ce_1, ce_2, \cdots, ce_n\}$。词法规则如下所示：

$$\begin{cases} N|NT|NZ \\ A|AD+N|NS|NZ \\ N+NV+N \end{cases}$$

1) https://github.com/ZhuiyiTechnology/simbert

其中，$N|NT|NZ$ 为名词，$A|AD$ 为形容词，V 为动词（符号表示以 jieba 词性标注符号为基准）。另外，针对多个名词组合，我们以知识图谱 CN-DBpedia 为知识库 KG，基于最长匹配原则进行选取，即若多个连续的词按该词法规则组合后均能在 CN-DBpedia 中查询到，则选择最多的词组合的短语作为候选关键词。更形象地说，如果 w_iw_{i+1}，$w_iw_{i+1}w_{i+2}$ 均能在 KG 中查询到，则选择 $w_iw_{i+1}w_{i+2}$ 作为候选关键词。

3.3.2　候选实体获取

候选实体获取指从指称项出发找到知识图谱中所有可能的实体组成候选实体集，即对给定 $ce\in CE$，找出 KG 中与之关联的候选映射实体集合 $MCE=\{mce_1,mce_2,\cdots,mce_j\}$。如果找不到，则认为该词不构成命名实体（也可能因为 KG 不完备导致找不到实体映射），将其加入非映射实体集合 NE。然后对候选映射实体集中的实体属性以及实体标签进行处理，通过选取代表性标签以及属性组成实体的背景描述文本。由于描述实体之间的属性与描述信息没有统一的规范，我们仅拼接实体属性中的类别信息以及部分描述信息，得到候选映射实体集的背景描述文本集合 $MCT=\{mct_1,mct_2,\cdots,mct_j\}$，每一个候选映射实体 mce_i 对应一个描述文本 mct_i。

例如，在“晋中市乔家大院景区被文化和旅游部取消旅游景区质量等级”中“乔家大院”实体的候选实体集有：[“乔家大院（山西省著名建筑）”，“乔家大院（2006 年胡玫执导电视剧）”，“乔家大院（朱秀海小说）”，“乔家大院（民族交响组曲）”，“远情（电视剧《乔家大院》主题曲）”]。而候选实体“乔家大院（山西省著名建筑）”的描述信息为“[[‘CATEGORY_ZH’，‘景观景点’]，[‘CATEGORY_ ZH’，‘地理’]，[‘CATEGORY_ZH’，‘地点’]，[‘景区类型’，‘人文景观、历史景观’]，[‘DESC’，‘乔家大院（Qiao Family Courtyard）：是全国重点文物保护单位，国家二级博物馆，国家文物先进单位…]]”。抽取其类别信息“CATEGORY_ZH”字段部分和描述信息“DESC”的部分字段，得到的实体背景信息为“山西省著名建筑景观景点地理地点全国重点文物保护单位”。

3.3.3　实体消歧

实体消歧是根据上下文信息消除一词多义的现象，也就是从 KG 中获取的候选实体 MCE 中选出语义与原文一致的实体。针对实体消歧任务，目前最常用的方法是将其视为二分类问题。对每一个候选实体进行多方位的特征因子抽取，利用一个多层感知机模型对这些特征因子进行融合打分，预测每一个候选实体和指称项的关联分数。由于中文关键词抽取语料集少，我们针对聚类生成的主题句簇，通过计算候选实体的标签文本 mct_i 与该词所在句子 s 之间的余弦相似度来选取与原文语义最近的候选实体。计算见式(3)，选取 ce 所在句子作为候选实体上下文 s，$Score(s,mct_i)$ 表示依次计算 s 与 $mct_i\in MCT$ 所生成的句向量之间的余弦相似度，从中选出相似度值最大标签文本 mct_{gold} 作为 MCT 中与候选实体上下文语义最相似的标签文本。文中算法成功从候选实体集中选取出映射实体“乔家大院（山西省著名建筑）”。

$$mct_{\text{gold}}=\arg\max Score(s,mct_i) \tag{3}$$

然后再根据相似度最高的 mct_{gold} 从候选映射实体集 MCE 中选出映射实体 mce_{gold}，达到实体消歧的效果；并且建立起 $ce\leftrightarrow mce_{\text{gold}}\leftrightarrow mct_{\text{gold}}$ 间的映射关系，同时将正确映射的实体 mct_{gold} 加入映射实体属性集 EA，实现 KG 的语义融入。最后将未能映射的候选实体 ce 放入集合 NE，将消歧得到的 mce_{gold} 放入集合 RE，从而组成候选关键词集合 $W=RE\cup NE$。

通过实体链接生成候选关键词的具体流程见算法 1。*participle*(*s*)表示使用 jieba 分词对句子 *s* 进行分词和词性标注，为方便表示，将词与词性分别记录在数组 *sw*，*sp* 中；*SearchKG*(*query*)是在知识图谱中对候选词 *query* 进行查询，如果查询到对应实体则返回真值，否则返回假值；*disambiguation*(*query*)是对候选词 *query* 在知识图谱中查询到的实体进行消歧，最终返回相应的映射实体 mce_{gold}。

算法 1　Entity Linking Algorithm

```
INPUT: T_i={s_1, s_2, …, }, KG
OUTPUT: CK(candidate keywords)
pattern=[n, nt, nz, a, v, a+n, ad+n, v+n, …]
FOR s IN T_i DO:
    sw, sp←participle(s)
    i=0, j=1
    WHILE i<sw.length:
      query=sw[i]
      query_pos←sp[i]
      IF query_pos IN pattern:
        IF !SearchKG(query_word):
          NE←query
        ELSE:
          WHILE j<sw.length:
            query←query+sw[j]
            query_pos←query_pos+sp[j]
            IF query_pos IN pattern:
            IF SearchKG(query)
              j←j+1
            ELSE:
              query←query[0:-sw[j].length]
              mce←disambiguation(query)
              CK←mce
              j←i; j←i+1
              BREAK
          ELSE:
            i←i+1; j←i+1
            BREAK
      ELSE:
        i←i+1, j←i+1
```

3.4　词图构建

在每个主题内部，引入知识图谱的结构化信息来帮助关键词排序计算。构建一个语义词图 $GW=(V,E,EA)$，其中结点集 V 是候选关键词集 W 的一个子集，即该主题内所有候选词，E 是结点之间组成边集，EA 是结点标签集。GW 是一个带权重无向图，其中的结点依据句共现相连，即出现在同一个句子中的词两两相连，不同句子之间以词共现相连。结点集 V 之间的权重通过式(4)分配，权重代表了词之间的语义相关性，由候选词之间的余弦相似度与位置距离加权得到。结点

集 V 与结点标签集 EA 之间的权重 W_{kg} 由该结点所在句子与其对应的背景标签文本之间的余弦相似度计算得到，如式(5)所示：

$$W_{\text{sematic}_{i,j}}=\alpha\cdot Score(w_i,w_j)+(1-\alpha)\cdot\frac{1}{|wp_i-wp_j+1|} \tag{4}$$

$$W_{kg}=Score(mct_{\text{gold}},s) \tag{5}$$

其中，α 是阻尼系数，$Score(w_i,w_j)$ 用于计算两个词之间的余弦相似度，$1/|wp_i-wp_j+1|$ 表示词之间的位置距离，wp_i 为该候选词在句子中的偏移位置。两个词隔得越近，则其位置距离的权重越大，在语义上更相关，加 1 则是为了适当降低相邻词的位置权重。

在实际应用中，并不总是需要计算所有关键词的分数，因为有些词几乎不可能是关键词，因此我们可以在应用排序算法之前删除一些结点，以达到消除噪声和降低算法复杂度的目的。直观来看，$W_{\text{sematic}_{i,j}}$ 反映了两个词之间的语义相关性，若一个词与其他词之间的相关性均小于阈值，且该词在知识图谱中无映射实体，则该词成为关键词的可能性很小，可以直接删除。另外，结点的度也是我们需要考虑的因素，若一个词在主题中的出现频率较高，与其相连的结点就会偏多，需要对其进行适当剪枝。若一个词连接 u 个以上结点，则删除所有与该结点相连的权重低于第 u 个的边。通常来说，u 设置为 3 或 4。图 2 给出了上文示例文本中一个主题构建的关键词图，图中直线表示结点之间的连接，曲虚线表示结点与结点标签集之间的连接，直虚线表示删除的连接。

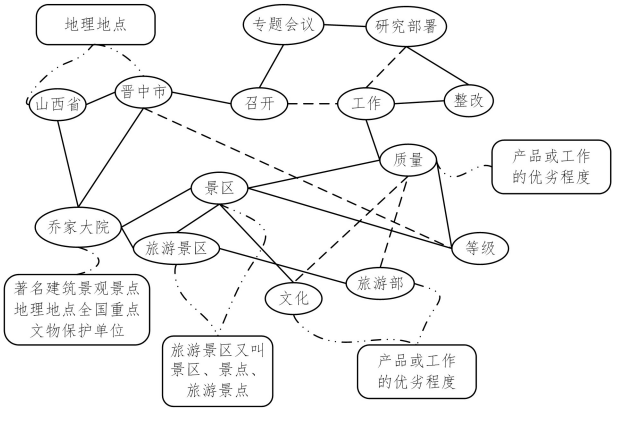

图 2 主题簇中的候选关键词图

Fig. 2 Candidate keyword graph within topic cluster

完成主题内部的词图构建后，需要考虑主题之间的关联性，来表示其在全文中的重要性和代表性。考虑主题之间的关联，将各主题作为结点，构建全文的主题图 $GT=(T,E')$，主题之间的边依据主题中每个句子之间的相关程度进行加权，权重计算如式(6)所示：

$$W_{\text{topic}_{i,j}}=\frac{1}{p'+q'}\sum_{s_i\in t_p}\sum_{s_j\in t_q}rel(s_i,s_j) \tag{6}$$

其中，t_p，t_q 表示两个不同主题；p'，q' 分别为两个主题中候选词的个数；$1/(p'+q')$ 用于归一化主题之间的距离；$rel(s_i,s_j)$ 用于计算句子之间的关联性，如式(7)所示，相关程度由主题中句子之间的余弦相似度与其位置距离的倒数加权得到。

$$rel(s_i,s_j)=\alpha\cdot Score(s_i,s_j)+(1\text{-}\alpha)\cdot\frac{1}{|sp_i-sp_j+1|} \tag{7}$$

其中，$Score(s_i,s_j)$ 表示两个句子之间的余弦相似度；$1/|sp_i-sp_j+1|$ 表示句子之间的位置距离，sp_i 为句子 s_i 在文本中的位置，两个句子相隔越远，则认为其位置距离的权重越小，相关性也就越小。

3.5 关键词选取

前文从语义的角度出发对关键词应具备的特征进行了阐述，并为此设计了相应的方法来提取特征。但如何确定一个词，仍需从数学角度进行分析，通过相应公式进行计算。对于任意 $w_i\in W$，$W=\{w_1,w_2,\cdots,w_h\}$ 表示文本中包含的所有词和短语，当 w_i 为关键词时，应能够使条件概率 $p(W|w_i)$ 最大，即通过关键词 w_i 最能判断出 W 中的其他词。由于每个词之间相互独立，依据朴素贝叶斯假设，条件概率可以转化为词与词之间的转移概率。

$$p(W|w_i)=\prod_{k=1}^{h}p(w_k|w_i) \tag{8}$$

从式(8)中可知，通过估算转移概率即可得到条件概率，从而完成关键词抽取。计算词之间的转移概率与 PageRank 计算网页之间的转移概率思路一致，通过其他与该结点相连的结点来计算该结点的重要程度。因此构建语义图后，可以利用基于改进的 PageRank 算法对每个主题和主题内包含的候选关键词的权重进行计算并排序，主题和关键词的排名代表了其在文本中的重要性和代表性。根据 3.4 节计算的权重为语义图中的每条边分配单独的权重，然后根据主题权重计算出每个主题的得分 $St(T_i)$。在主题内部，同样依据词之间的权重计算出每个候选关键词的得分 $Sw(w_i)$。最后，依据主题与候选词的得分对所有词进行排序，进而选取出最终的关键词，见式(9)－式(11)。

$$St(T_i)=\lambda\cdot\sum_{T_j\in T,j\neq i}\frac{W_{\text{topic}_{j,i}}\cdot St(T_j)}{\sum_{T_k\in T}W_{\text{topic}_{k,j}}}+(1-\lambda) \tag{9}$$

$$Sw(w_i)=\lambda\cdot\sum_{w_j\to w_i}\frac{W_{\text{sematic}_{j,i}}\cdot Sw(w_j)}{\sum_{w_k\to w_j}W_{sematic_{k,i}}}+(1-\lambda) \tag{10}$$

$$R(w_i)=\alpha\cdot Sw(w_i)+(1-\alpha)\cdot St(T_i) \tag{11}$$

其中，λ，α 均为阻尼系数，T_j 是主题集合 T 中除 T_i 的集合，T_k 是主题集合 T 中所有的集合，w_j 是指向 w_i 的前置结点集合，w_k 是 w_i 指向的点的集合。

4 实验和分析

4.1 数据集

限制中文关键词发展的一个重要因素是缺乏标准、统一的关键词数据集。由于很难找到一个适用于所有领域的文本关键词抽取方法，并且大部分研究自己构建的关键词本身就带有一定的主观性，导致了实验结果难以评估。一些研究[35]采用科技论文的摘要和数据集进行测试，还有一些研究[36]爬取网络新闻文本并以网页标签为依据构建关键词作为测试集。而这两者也有所区别，科技论文倾向于使用关键短语，而新闻往往使用短的关键词，并且科技论文摘要本身较短，有时并不能代表论文全文，一些关键词可能并不会出现在摘要中，这影响了关键词抽取评估的性能。为了全面评估 CnKGRank 的性能，本文在 3 个数据集即 DUC 2001，CSL[1)] 和 CLTS 上进行实验。表 1 列出了有关数据集的详细信息。

表 1　数据集的统计信息

Table 1　Statistics information of datasets

Datasets	Type	Text number	Average text length/(word)	Average number of keywords
DUC 2001	News	308	1 300	8
CSL	Abstract	3 000	275	4
CLTS	News	200	1 364	5

(1)DUC 2001。该数据集包含 308 篇新闻文章,文件的平均长度约为 700 字,每个文档被手动分配大约 10 个关键词。选择该数据集是因为许多优秀的基准方法都在该数据集上进行测试。但是该数据集为英文,本文借助百度翻译[2)]将该数据集翻译为中文,并手动对一些关键词与文本进行了修正。如"firefighters"可能被翻译为"消防人员""消防员"或者"消防队员",本文则手动将其统一修改为"消防队员"。

(2)CSL。该数据集取自部分中文社会科学和自然科学核心期刊的论文摘要及关键词,其中适用于关键词抽取任务的 csl_public 包含 2.6 万短文本。此外,CSL 使用 tf-idf 生成伪造关键词,并将其与论文真实关键词混合来构造摘要-关键词对。该数据集主要用于有监督的关键词抽取,本文在此只进行对比实验,因此选取其测试集进行测试。

(3)CLTS。CLTS 是基于中文新闻网站 ThePaper.cn 的中文长文本摘要数据集。数据集的最终版本包含超过 180 000 个长序列对,其中每一篇文章由多个段落组成,每个摘要由多个句子组成,文本平均长度为 1 364 个字。选取该数据集的目的是测试本文模型在长文本新闻数据上的性能。由于该数据集不包含关键词,我们选取了其训练集上的前 200 篇新闻,通过 5 名专家结合文本和摘要内容对每篇文章标注 4～7 个关键词,然后通过统计选取出选取率最高的 5 个词,以保证关键词尽可能客观。

4.2　评价指标

为客观评估关键词抽取效果,对不同数据集,分别记录不同的方法在该数据集上的准确率、召回率和 F1 值(Precision/Recall/F1-Score)。

$$Prescision=\frac{C_{\text{right}}}{C_{\text{extract}}} \tag{12}$$

$$Recall=\frac{C_{\text{right}}}{C_{\text{standard}}} \tag{13}$$

$$F_1\text{-}Score=\frac{2Prescision*Recall}{Prescision+Recall} \tag{14}$$

其中,C_{right} 是一个方法所有准确抽取的关键词数目,C_{extract} 是所有抽取的关键词数目,而 C_{standard} 是所有人工标注的标准答案数目。

4.3 实验结果与分析

4.3.1　CnKGRank 与对比算法的 F1 值比较

目前中文关键词抽取研究较少,这些中文数据集缺少普适的对比数据。为了与本文的方法进行比较,我们重新实现了基线方法 TF-IDF,TexkRank,TopicRank 以及 MultipartiteRank。这 4 种方法 F_1 值是传统统计方法、图模型方法以及基于词嵌入的主题聚类方法的代表性方法,MultipartiteRank 是 TopicRank 的扩展,将候选关键词及主题关系存储在多部图中,进一步增强了主题特征。与这 4 种方法进行比较,可以充分体现本文所提方法的性能。

表 2 列出了本文方法与另外 4 种方法在 3 组数据集上准确率、召回率与 $F1$ 值的结果,其中 W 表示滑动窗口大小,K 表示抽取关键词的数目。在测试所有 4 种方法时,针对 3 种数据集关键词的平均标注数量不同,为每个数据集设置了不同的关键词预测数量。通过对比实验数据可以看到,CnKGRank 在 3 个数据集上相比其他 4 种方法 $F1$ 值都有一定的提高,特别是在 CLTS 数据集上,相比其他 3 种方法,$F1$ 值分别提高了 9.14%,4.82%,3.05%和 1.77%。这表明了知识图谱的语义信息和预训练模型的词嵌入特征在关键词提取任务中的有效性。同时,传统的 TF-IDF、基于图的 TexkRank 方法与主题聚类的 TopicRank 和 MultipartiteRank 方法均在短文本数据集 CSL 上表现一般。究其原因,短文本内容较少,用 TexkRank 方法构建的词图缺乏足够的信息来选取出关键词,而以词为单位进行聚类同样不能很好地得到主题特征;另一方面,CSL 数据集以论文摘要为主,含有的命名实体较多,这为传统方法带了一定困难,而且论文中标注的关键词有部分没有在摘要中出现,这直接影响了关键词抽取算法的准确率。而 CnKGRank 借助于知识图谱蕴含的专家知识以及语义信息,有效扩充了文本的内容,实验也进一步证明了该方法的有效性。

表 2　CnKGRank 与其他方法在 3 种数据集上的对比

Table 2　Comparison of CnKGRank and other methods on three datasets

(单位:%)

Datasets	Methods	Parameters	*Precision*	*Recall*	*F1-Score*
DUC 2001	TF-IDF	$K=8$	9.98	9.92	9.95
	TextRank	$W=5,K=8$	10.67	10.61	10.64
	TopicRank	$W=5,K=8$	11.89	11.82	11.86
	MultipartiteRank	$K=8$	12.78	12.71	12.75
	CnKGRank	$K=8$	13.47	13.39	13.43
CSL	TF-IDF	$K=4$	10.18	11.43	10.77
	TextRank	$W=5,K=4$	8.93	10.03	9.44
	TopicRank	$W=5,K=4$	10.31	11.58	10.91
	MultipartiteRank	$K=4$	11.06	12.42	11.70
	CnKGRank	$K=4$	12.16	13.66	12.86
CLTS	TF-IDF	$K=5$	20.30	19.63	19.96
	TextRank	$W=5,K=5$	24.70	23.89	24.29
	TopicRank	$W=5,K=5$	26.50	25.63	26.06
	MultipartiteRank	$K=5$	27.80	26.89	27.34
	CnKGRank	$K=5$	29.60	28.63	29.11

另外,为了测试了不同关键词抽取数量对 $F1$ 值的影响,本文在 CLTS 数据集上选取了不同的 K 值进行实验,其结果如图 3 所示。随着关键词抽取数量的增加,模型准确抽取的关键词数量增加,召回率有所提升,但是由于抽取的词语数量基数增大,准确率会下降。在实际应用中不可能抽取过多的关键词,需要用尽可能少的关键词来完整表达原文传达的主题与思想,因此将准确率与召回率放在同等重要的位置,而这也是选择 F1 值作为评价指标的重要原因。

1) https://github.com/P01son6415/CSL

2) https://fanyi-api.baidu.com/

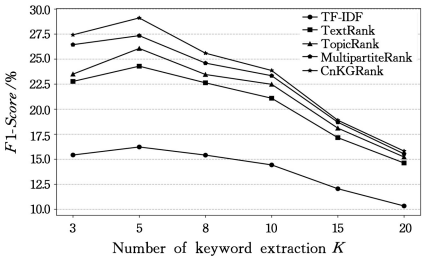

图 3 关键词抽取数量对 F1 值的影响

Fig. 3 Influence of number of extracted keywords on F1 value

4.3.2 消融实验

为了探究知识图谱中语义信息与预训练模型中的句向量特征对关键词抽取的影响，我们进行了消融实验。将 CnKGRank 分为 3 个版本。1)与 TopicalRank 方法类似，仅根据分词工具和词性匹配原则生成候选关键词，然后通过句聚类方法生成候选关键词，用 CnKGRank(+cluster)表示；2)仅通过知识图谱进行实体链接，没有通过句聚类来生成主题，用 CnKGRank(+el)表示；3)联合句聚类与实体链接的方法，用 CnKGRank(+union)表示。

图 4 给出了消融实验的结果。从图中可以看到，仅使用 BERT 模型进行句聚类后的方法 CnKGRank(+cluster)与其他方法相比 $F1$ 值提升不大，在长文本数据集 DUC 2001 和 CLTS 上与 TopicRank 相比，$F1$ 值分别提升了 0.72% 和 0.78%，在短文本数据集 CSL 上 $F1$ 值比 TextRank 提升了 0.46%，比 TopicRank 方法下降了 1.01%。这表明了在未引入知识图谱的情况下，短文本数据集上的词聚类比句聚类有更好的效果，因为短文本句子较少，可能只有 2～3 句话，这种情况下使用句子进行聚类难以得到有效的主题特征。但 CnKGRank(+el)方法在引入知识图谱的语义信息后，$F1$ 值得到了较大提升，在 3 个数据集上相比 TopicRank 方法，$F1$ 值分别提升了 1.25%，1.03%和 1.67%。

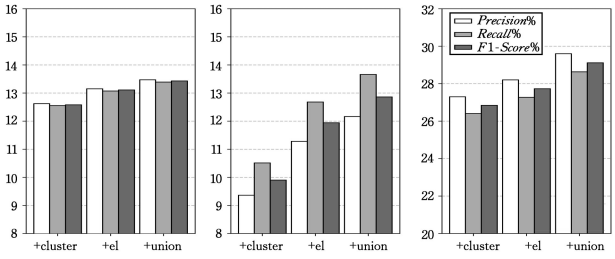

图 4 消融实验结果对比

Fig. 4 Comparison of the results of ablation experiments

消融实验证明了知识图谱中语义信息在关键词抽取任务中具有更好的效果，因为通过实体链接引入知识图谱中的语义信息和关联关系之后，消除了部分词的歧义，使得一些词被正确选为候选关键词。例如，在 DUC 2001 数据集中，一篇文章的关键词为“光辉道路游击队”，传统分词算法会将其分为“光辉、道路、游击队”，而经过实体链接之后，“光辉道路游击队”则被准确地从句子中分割出来，并被选为关键词。相比之下，预训练模型的句向量特征对关键词抽取效果的提升影响较小，并且在短文本处理中的表现不如以词语为单位进行聚类的方法。但是就长文本而言，以句为单位进行主题聚类能够加快模型的处理速度，并且更好地保留语义信息。因而在实际应用中，可以针对不同的文本数据，灵活使用不同的方案。

结束语 本文提出了一种联合知识图谱与预训练模型的无监督中文关键词抽取方法 CnKGRank。该方法首先利用聚类算法和预训练模型将词嵌入向量特征融入词图构建中，将文本句子按主题分类；然后借助知识图谱的实体属性弥补当前中文文本分词工作的不足，通过实体链接将文本与知识图谱中的实体联系起来，将语义信息融入文本；最后依据改进的 PageRank 算法获得关键词的排名，选取出一组与原文本语义最相近且主题覆盖率最高的关键词。实验表明，所提方法优于目前主流的基于图的方法。可以预见的是，未来的自然语言处理领域与知识图谱之间必然是相辅相成的，NLP 的发展促进知识图谱的构建，而知识图谱通过引入知识促进 NLP 的发展。随着深度学习技术的飞速发展，语义特征将会更深入地融入各个任务，而如何利用预训练模型和已有知识图谱使各项任务不再依赖于标注好的语料集仍是我们努力的方向。此外，未来我们将继续标注 CLTS 语料集，致力于构建出一个优质的中文长文本新闻关键词抽取语料集。

致谢 特别感谢复旦大学知识工厂实验室提供知识图谱 CN-DBpedia 的调用 APIKEY，使本文的实验得以顺利进行；另外，感谢中国科学院信息工程研究所刘晓君等人提供的 CLTS 数据集以及苏剑林提供的开源 SimBERT 模型，给予本文实验较大的帮助。

参 考 文 献

[1] ZHAO J S, ZHU Q M, ZHOU G D, et al. Review of research in automatic keyword extraction[J]. Journal of Software, 2017, 28(9): 2431-2449.

[2] LIU Z Y. Research on keyword extraction using document topical structure[D]. Beijing: Tsinghua University, 2011.

[3] CHEN T, MIAO D, ZHANG Y. A Graph-Based keyphrase extraction model with three-way decision[C]// Proceedings of the Rough Sets-International Joint Conference. Havana, Cuba, 2020: 111-121.

[4] DING Z, ZHANG Q, HUANG X. Keyphrase extraction from online news using binary integer programming[C]// Proceedings of the 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand, 2011: 165-173.

[5] CHANG Y C, ZHANG Y X, WANG H, et al. Features oriented survey of state-of-the-art keyphrase extraction algorithms[J]. Journal of Software, 2018, 29(7): 2046-2070.

[6] YU Y, NG V. WikiRank: Improving keyphrase extraction based on background knowledge[J]. arXiv: 1803.09000, 2018.

[7] GRINEVA M, GRINEV M, LIZORKIN D. Extracting key terms from noisy and multitheme documents[C]// Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, 2009: 661-670.

[8] TSATSARONIS G, VARLAMIS I, NORVAG K. SemanticRank: Ranking keywords and sentences using semantic graphs[C]// Proceedings ofthe 23rd International Conference on Computational Linguistics. Beijing, 2010: 1074-1082.

[9] BO X, YONG X, LIANG J, et al. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System[C]// Proceedings of the 30th International Conference on Industrial Engineering and

Other Applications of Applied Intelligent Systems(IEA/AIE 2017). Arras, France, 2017: 428-438.

[10] OVER P. Introduction to DUC 2001: An intrinsic evaluation of generic news text summarization systems[C]// Proceedings of the Document Understanding Conference. 2001.

[11] LIU X, ZHANG C, CHEN X, et al. CLTS: A new chinese long text summarization dataset[C]// Proceedings of the Natural Language Processing and Chinese Computing(NLPCC 2020). Cham: Springer, 2020: 531-542.

[12] DUAN J Y, YOU S X, ZHANG M, et al. Keyword Extraction Based on Multi-feature Fusion[J]. Computer Science, 2020, 47(S2): 73-77.

[13] HOFMANN T. Probabilistic latent semantic indexing [J]. Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, 1999, 51(2): 50-57.

[14] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[15] PU X, JIN R, WU G, et al. Topic Modeling in Semantic Space with Keywords[C]// Proceedings of the 24th ACM International Conference on Information and Knowledge Management. New York, 2015: 1141-1150.

[16] LIU X J, XIE F. KeywordExtraction Method Combining Topic Distribution with Statistical Features[J]. Computer Engineering, 2017, 43(7): 217-222.

[17] ALREHAMY H H, WALKER C. SemCluster: Unsupervised Automatic Keyphrase Extraction Using Affinity Propagation [C]// Advances in Computational Intelligence Systems(UKCI 2017). 2017: 222-235.

[18] AWAN M N, BEG M O. TOP-Rank: A TopicalPostionRank for Extraction and Classification of Keyphrases in Text [J]. Journal of Computer Speech Language, 2021(65): 101-116.

[19] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space [J]. arXiv: 1301.3781v3, 2013.

[20] PENNINGTON J, SOCHER R, MANNING C. Glove: Global Vectors for Word Representation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP 2014). Doha, Qatar, 2014: 1532-1543.

[21] WANG R, LIU W, MCDONALD C. Corpus-independent Generic Keyphrase Extraction Using Word Embedding Vectors[C]// Proceedings of the Software Engineering Research Conference. 2015.

[22] MAHATA D, KURIAKOSE J, SHAH R R, et al. Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings [C]// Proceedings of NAACL-HLT. New Orleans, Louisiana, 2018: 634-639.

[23] ZHANG Y, LIU H, WANG S, et al. Automatic keyphrase extraction using word embeddings [J]. Journal of Soft Computing, 2020(24): 1-16.

[24] QUILLIAN M R. Semantic networks [J]. Approaches to Knowledge Representation Research Studies, 1968, 23(92): 1-50.

[25] MIHALCEA R, TARAU P. TextRank: Bringing Order into Text[C]// Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain, 2004: 404-411.

[26] WAN X, XIAO J. Single Document Keyphrase Extraction Using Neighborhood Knowledge[C]// Proceedings of the 23rd AAAI Conference on Artificial Intelligence. Palo Alto, 2008: 855-860.

[27] BOUGOUIN A, BOUDIN F, DAILLE B. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction [C]// Proceedings of International Joint Conference on Natural Language Processin. Nagoya, Japan, 2013: 543-551.

[28] FLORESCU C, CARAGEA C. A Position-Biased PageRank Algorithm for Keyphrase Extraction[C]// Proceedings of the Association for the Advancement of Artificial Intelligence. San Francisco, California, 2017: 582-592.

[29] BOUDIN F. Unsupervised Keyphrase Extraction with Multipartite Graphs[C]// Proceedings of NAACL-HLT. New Orleans, Louisiana, 2018: 667-672.

[30] SHI W, ZHENG W, YU J X, et al. Keyphrase Extraction Using Knowledge Graphs [C]// Asia-Pacific Web(APWeb) and Web-Age Information Management(WAIM) Joint Conference on Web and Big Data. Cham: Springer, 2017.

[31] GAO T, YAO X, CHEN D. SimCSE: Simple Contrastive Learning of Sentence Embeddings [J]. arXiv: 2104.08821, 2021.

[32] SU J, CAO J, LIU W, et al. Whitening Sentence Representations for Better Semantics and Faster Retrieval [J]. arXiv: 2103.15316, 2021.

[33] JI H, GRISHMAN R, DANG H T, et al. Overview of the TAC 2010 knowledge base population track[C]// Proceedings of the Third Text Analysis Conference(TAC). Gaithersburg, Maryland, 2010.

[34] SUN M S, CHEN X X, ZHANG K X, et al. THULAC: An Efficient Lexical Analyzer for Chinese[EB/OL]. https://nlp.csai.tsinghua.edu.cn/project/thulac/.

[35] XIA T. Extracting Key-phrases from Chinese Scholarly Papers [J]. Data Analysis and Knowledge Discovery, 2020, 4(7): 76-86.

[36] LIANG Y. Chinese keyword extraction based on weighted complex network[C]// Proceedings of International Conference on Intelligent Systems and Knowledge Engineering(ISKE). Nanjing, China, 2017: 1-5.

YAO Yi, born in 1981, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include software engineering, natural language processing and knowledge graph.

YANG Fan, born in 1997, postgraduate. His main research interests include knowledge graph and natural language processing.

（责任编辑：柯颖）