计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200300121

基于多特征融合的关键词抽取

段建勇　游世薪　张　梅　王　昊
北方工业大学信息学院　北京 100144

摘　要　随着互联网的发展,网页数据以及新媒体文本等数据日益增多,全文信息检索的效率已经不足以支撑海量数据的检索,因而关键词抽取技术广泛应用于搜索引擎(如百度搜索)和新媒体服务等领域(如新闻检索)。融合模型是一种使用 BiLSTM-CRF 结构并融合多重手工特征的模型,可以更有效地完成关键词抽取任务。融合模型在词嵌入特征的基础上,融入了词性、词频、词长和词位置特征,多维度的特征信息可以更加全面地辅助模型提取到关键词的深层特征信息。融合模型将深度学习的广覆盖度、高学习能力等特点与手工特征的精确表达能力相结合,以进一步提高特征挖掘能力并缩短训练所需时间。此外,该模型使用了一种新的"LMRSN"标记方法,可以更有效地完成关键短语的抽取。实验结果表明,融合模型在与传统模型的对比中取得了 62.08 的 F1 分值,性能远高于传统模型。

关键词:抽取;深度学习;特征融合;信息检索;长短期记忆网络
中图法分类号　TP391.1

Keyword Extraction Based on Multi-feature Fusion

DUAN Jian-yong, YOU Shi-xin, ZHANG Mei and WANG Hao
School of Information, North China University of Technology, Beijing 100144, China

Abstract　With the development of the Internet, webpage data, new media text and other data are increasing, the efficiency of information retrieval based on full text is not enough to support the retrieval of massive data, so the keyword extraction technology is widely used in search engines (such as Baidu search) and new media services (such as news retrieval). The fusion model is a model that uses the BiLSTM-CRF structure and fuses multiple manual features, which can more effectively complete the task of keyword extraction. Based on the features of words embedding, the fusion model incorporates the features of part of speech, word frequency, word length and word position. Themultidimensional feature information can help the model to extract deep keyword feature information more comprehensively. The fusion model combines the features of deep learning, such as wide coverage and high learning ability, with the ability of accurate expression of manual features to further improve the feature mining ability and shorten the training time. In addition, a labeling method called LMRSN is adopted in this model to extract key phrases more effectively. Experimental results show that the fusion model achieves F1 score of 62.08 in comparison with the traditional model, and its performance is much better than that of the traditional model.

Keywords　Keyword extraction, Deep learning, Feature fusion, Information retrieval, Long and short term memory network

1　引言

关键词抽取任务是从文本或文本集合中自动抽取主题性或重要性的词或短语的任务。关键词表征了文本中主题性和关键性内容,是文本内容理解的最小单位。在获取到文本的关键词信息后,可以通过关键词对文本进行分类、聚类等操作,从而提高文本的检索效率。具体示例如图 1 所示。

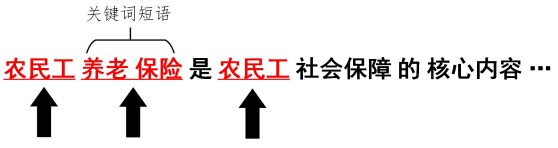

图 1　关键词、关键短语示例

Fig. 1　Examples of keywords and key phrases

其中,箭头所指为应当抽取出的关键词,如"农民工"。"养老"与"保险"是一种短语组合,所以应当抽取"养老保险"整体关键词短语。

融合模型使用序列标注的方式实现关键词抽取任务。业界许多领域已经证明,LSTM 模型在解决序列标注任务时拥有优异的表现。本文中,我们通过使用双向的 LSTM 模型,使融合模型可以从文本正反两个方向学习关键词特征。

在传统关键词抽取模型中,词频、词语上下文等特征信息是构建词语关键性的重要指标。融合模型在使用深度学习神经网络的前提下,将词性、词频、词长、词位置等多维度人工特征融合入词嵌入信息。通过实验发现,融入特征可以使模型快速学习到关键词的特征信息,并且融合模型特征提取与手工特征可以避免遗漏关键信息,提供多维度的信息可以由模型在迭代中学习每种手工特征的重要性权值。

序列标注任务中,标记方法同样会影响整体实验的效果。

基金项目:国家自然科学基金(61972003,61672040)
This work was supported by the National Natural Science Foundation of China (61972003, 61672040).
通信作者:段建勇(duanjy@ncut.edu.cn)

文中提出使用 LMRSN 标记法来提高标记的效率，并与 YN 标记法进行对比。LMRSN 标记方法可以准确地标记关键词短语的边界，提供了关键词短语抽取的解决办法。本文 3.3 节将详细介绍这种方法。

2 相关工作

关键词抽取作为自然语言处理领域与信息检索领域的基础任务，已经被研究了很多年。在这些研究中，关键词抽取任务被分为几个主要流派，其中最著名的是基于统计法的流派以及基于网络图方法的流派。

2.1 基于统计方法的流派

Salton 等[1]在 1988 年提出了 TF-IDF，其中 TF 被称为词频，表示一个词可以描述整个文本的能力；IDF 被称为逆文档频率，表示该词的区分度。如式(1)－式(4)所示，TF-IDF 方法可以通俗的表示为：一个词语在该文本中出现的频率高，而在其他文本中出现的次数少，则证明这个词语有高概率成为该文本的关键词。

$$tf_i=\frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

$$idf_i=\log\frac{|D|}{|D_i|+1} \tag{2}$$

$$w_{ij}=tfidf_{ij}=tf_{ij}\times idf_i \tag{3}$$

$$w_{ij}=\frac{w_{ij}}{\sqrt{w_{kj}}} \tag{4}$$

其中，n_{ij} 是词 t_i 在文档 d_j 中的出现次数，$|D|$ 表示语料库中的文档总数，$|D_i|$ 表示语料库中包含词 t_i 的文档总数，w_{ij} 表示词 t_i 在文档 d_j 中的归一化权重。

TF-IDF 方法可以快速有效地计算出一个词作为关键词的概率大小。但是由于该方法过多地依赖词频特征，导致特征提取不够全面。并且，在计算 IDF 值时，由于没有考虑特征项在类内与类间的分布信息，会导致一些在少量文档中存在分布的特殊生僻词条计算的 IDF 值异常高，从而影响模型的整体效果[2]。Besils 等[3]提出了 TF * IWF * IWF 算法，IWF 表示语料库词语总数与待查文本中该词出现次数之比的对数，在一定程度上解决了 IDF 带来的问题。

TFIDF 方法在关键词抽取任务上简单而且有效，利用词频与逆文档频率两个因素进行有效计算。显而易见，文章中词语的关键性高度于依赖词语出现的频率。但不仅这一特征，许多其他特征甚至深层特征同样会影响词语的关键性。我们的模型在使用词频作为特征的同时加入了大量其他特征，并且利用神经网络进行深层次的特征挖掘，更加充分地利用文章中的信息。

2.2 基于网络图方法的流派

基于随机游走法可以将节点序列构成一张图，从而计算各个节点之间的重要性。著名的 PageRank 算法就是利用这种思想提取了互联网中各个网页的重要性。

Mihalcea 等在 2004 年提出了 TextRank 算法[4]，将 PageRank 算法应用于关键词抽取领域。TextRank 算法无需训练数据，是一种无监督的学习方法，具有很强的适应性与扩展性。并且，TextRank 的收敛速度很快，所需要的训练时间很短。TextRank 算法中节点重要性由式(5)、式(6)计算得到：

$$t=\sum_{v_j\in Adj(v_j)}\frac{w_{ji}}{\sum_{k\in Adj(v_j)}w_{jk}}s(v_j) \tag{5}$$

$$S(v_i)=(1-d)+d\times t \tag{6}$$

其中，$S(v_i)$ 表示节点 v_i 的 PageRank 值或 TextRank 值，d 为阻尼系数。

基于网络图的方法可以在词语间构造出一张权重网络，从而得到每个词语的重要性，在特征提取层面更多地依赖词语在文章中出现的频率以及上一级词语的重要性。这些特征可以被人为总结，并在大量数据中自动提取，在关键词抽取任务中有着不错的表现，但特征提取深度依然较浅。我们使用的神经网络模型可以有效解决这个问题。

3 方法

本节首先介绍了融合模型结构，其次在 3.2 节介绍手工特征提取方式，最后在 3.3 节介绍本文的 LMRSN 标记方法。

3.1 模型结构

我们选择使用 BiLSTM-CRF[5]模型作为融合模型的基础结构来完成关键词抽取任务。融合模型的神经网络结构如图 2 所示，该模型融合多维度的人工特征(F 表示)与词嵌入特征，将融合向量输入 BiLSTM 模型，最后，模型中学习到的特征通过 CRF 层并获取标签。

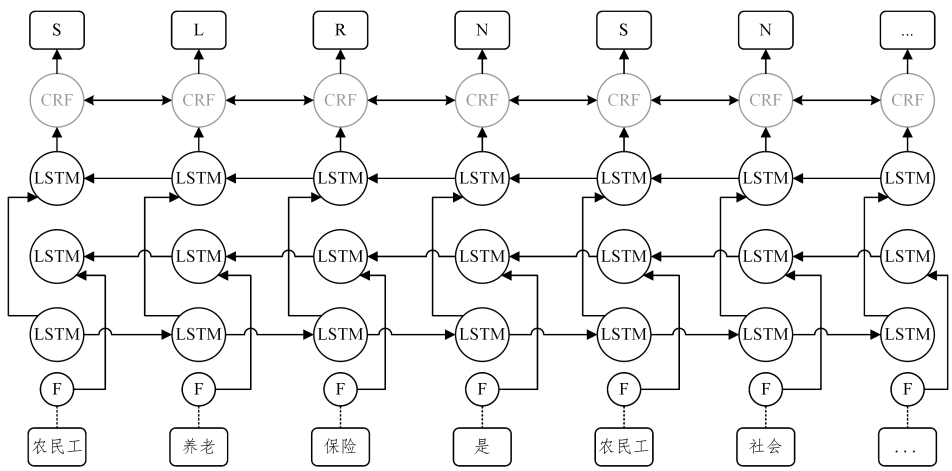

图 2 融合模型结构

Fig. 2 Structure of fusion model

3.1.1 词嵌入层

对于每一个词语，融合模型使用 GloVe 初始化词嵌入向量。如式(7)所示，利用词嵌入查找函数将每个词语转化为词向量。为了获取每个词语多维度的特征，融合模型提取了每个词语的词性、词频、词长和词位置特征，并将这些特征与词嵌入向量连接起来。

$$x_{b,e}^{w}=e^{w}(c_b,c_{b+1},\cdots,c_e) \tag{7}$$

3.1.2 双向长短期记忆网络层

Long Short Term Memory(LSTM)是 Recurrent Neural Network(RNN)的一个变种，LSTM 在 RNN 的基础上增加了遗忘门、输入门和输出门，所以可以保存部分前后文记忆。其中，遗忘门能决定应该丢弃或保留哪些信息；输入门用来更新单元状态；输出门则可以决定下一个隐藏状态的值。LSTM 的构造如图 3 所示。

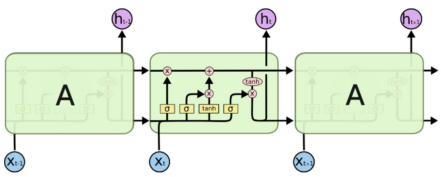

图 3 LSTM 结构

Fig. 3 Structure of LSTM

在 t 时刻，中间结果与最终结果的方程如式(8)－式(13)所示：

$$i_t=\sigma(W_i h_{t-1}+U_i x_t+b_i) \tag{8}$$

$$f_t=\sigma(W_f h_{t-1}+U_f x_t+b_f) \tag{9}$$

$$g_t=\tanh(W_{c-1} h_{t-1}+U_c x_t+b_c) \tag{10}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot g_t \tag{11}$$

$$o_t=\sigma(W_o h_{t-1}+U_o x_t+b_o) \tag{12}$$

$$h_t=o_t\odot\tanh(c_t) \tag{13}$$

双向长短期记忆(Bidirectional Long Short Term Memory，BiLSTM)网络层是由前向 LSTM 与后向 LSTM 组合而成。在自然语言处理任务中，BiLSTM 可以建模上下文信息，解决了 LSTM 仅可以从前文获得信息而不能获取后文信息的问题，从而使模型获得更好的效果。在关键词抽取任务中，由于词语的关键性与文本上下文相关，BiLSTM 模型结构可以更有效地提取到词语的关键性特征。

3.1.3 条件随机场层

在自然语言处理任务中，相邻的词语会互相影响，每个词语的特征概率会受到周围词语的特征影响。例如动词的下一个词语更大的概率为名词，而连续动词的可能性会降低。条件随机场(CRF)可以有效地预测此类情况。本文对关键词和关键短语的标记方法存在顺序性，例如 L 标签的下一个词语肯定是 M 或者 R 其中之一，并且连续词语组成的关键词短语长度也是一项重要特征。因此，我们引入 CRF 层来预测这些特征。

使用 $x=\{x_1,x_2,\cdots,x_n\}$ 表示输入的序列词语，例如“农民工”，使用 $y=\{y_1,y_2,\cdots,y_n\}$ 表示输出的标签序列，例如 S。给定 x 的每个标签序列的概率值如式(14)－式(16)所示：

$$P_1=\prod_{i=1}^{n}\varphi_i(y_{i-1},y_i,x) \tag{14}$$

$$P_2=\sum_{y'\in y(x)}\prod_{i=1}^{n}\varphi_i(y'_{i-1},y'_{i-1},x) \tag{15}$$

$$p(y|x;\boldsymbol{W},\boldsymbol{b})=\frac{P1}{P2} \tag{16}$$

其中，$Y_{(X)}$ 表示 x 集合中所有可能的标签序列，$p(y|x;\boldsymbol{W},\boldsymbol{b})$ 则表示给定 x 的每个标签序列的概率。

式(16)中，$\boldsymbol{W}$ 表示权值转移矩阵，$\boldsymbol{b}$ 代表偏置转移矩阵。两个矩阵中的每一个参数表示标签的转移得分。融合模型使用式(17)中的对数极大条件似然估计来预测概率。我们需要找出一个合适的条件概率，使损失函数达到最大值，从而确定最终的标签序列。

$$L(\boldsymbol{W},\boldsymbol{b})=\sum_i\log p(y|x;\boldsymbol{W},\boldsymbol{b}) \tag{17}$$

通常，使用 Viterbi 算法[6]来训练 CRF 模型可以取得不错的效果，所以我们使用同样的方法来预测融合模型。

3.2 特征

词性、词频、词长、词位置这些词语的特性可以辅助模型分析关键词的特征。例如：名词有更高概率成为关键词，而动词成为关键词的概率较低。同时，这些特征在传统关键词抽取方法中被证明可以有效提高关键词抽取的效率。如图 4 所示，我们将提取的各种特征与词嵌入特征拼接从而获得融合特征向量。

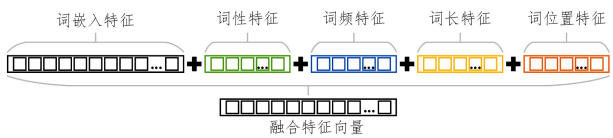

图 4 融合特征向量

Fig. 4 Fusion of feature vectors

3.2.1 词性

词性是以语法特征为主要依据并且兼顾词语含义的划分结果。在关键词抽取任务中，词性可以作为重要的衡量指标。我们可以很容易地发现名词性的关键词数量远大于其他词性的关键词数量，但是很难精确定义词性在关键词特征中的重要性权值，所以本文将词性融入词嵌入特征，经过神经网络进行特征分析与提取。

3.2.2 词频

词频代表着一个词在文本中的出现频率。传统的关键词抽取方法，如 TF-IDF，将词频作为一个极其重要的指标进行计算。词语的出现频率可以在很大程度上影响其作为关键词的概率，所以本文将词频作为一种重要的手工特征。

3.2.3 词长

关键词作为文章的主题性词语，包含大量关键信息。而词的长度会一定程度影响词语的表达能力。例如，长度为 1 的词语很难表达出重要的主题含义，而过长的短语又会导致语义发散。

3.2.4 词位置

在文本表达中，通常关键性词语的位置会更加考究。因此，关键词在文本中的分布具规律性，且每个位置的词语成为关键词的概率会有所不同，因此，融合模型提取了词位置特征，通过深层神经网络进行拟合以获取其特征信息。

3.3 标记方法

融合模型使用 LMRSN 标记方式来标记训练数据。如图 5 所示，其中 L，M，R 3 种标签用来标记关键词短语，L 标记关键词短语的开头词语；M 标记关键词短语的内部词语；R 标记关键词短语的结束词语。S 标签用来标记单独词语的关

键词。对于非关键词的词语使用 N 作为标签标记。

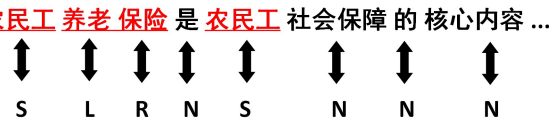

图 5 LMRSN 标记方法实例

Fig. 5 Example of LMRSN

常用的 YN 标记方式使用 Y 标记关键词信息；N 标记非关键词信息。LMRSN 标记方式与传统 YN 标记方式相比可以标记出关键词的边界信息，从而在关键词短语抽取中取得了更优异的表现。

4 实验

本节首先介绍了实验的数据预处理以及评价指标，其次介绍了在评估融合模型效率时所进行的各种对比实验。本文从多角度出发进行实验：1)将融合模型与传统基准模型在同样数据集下进行实验对比；2)针对不同人工特征的组合进行实验对比；3)对词嵌入的方式进行对比与探索；4)分别对 LMRSN 标记法与 YN 标记法进行实验，证明标记方式对模型整体效率的影响。最后，我们对融合模型中的超参数进行实验对比。

4.1 数据预处理

传统的关键词抽取任务多由无监督学习完成，而目前还没有公开的标记好的大型数据集。因此，研究人员经常通过抓取一些网站中的富文本信息或使用通用的数据源来创建数据集。在本文中，我们获取了 164236 条新闻文本数据作为训练数据，每一条数据由新闻标题、新闻内容和新闻关键词 3 项组成。

本文对实验数据做了简单的预处理：首先对数据进行分词，并根据新闻关键词信息在内容中进行 LRMSN 标记，其次提取了每个词语的词性、词频、词长和词位置特征。如表 1 所列，最终获得了 116 236 条训练数据、32 000 条测试数据、16 000 条验证数据。

表 1 数据预处理详情

Table 1 Data preprocessing details

	Training set	Test set	Validation set
Amount	116236	32000	16000
Percentage/%	70.78	19.48	9.74

4.2 评估标准

本文使用精确度、召回率和 F1 值作为评估标准，如式(18)－式(20)所示，这是评估模型效果的标准方法：

$$p=\frac{TP}{TP+FP} \tag{18}$$

$$r=\frac{TP}{TP+FN} \tag{19}$$

$$F1=\frac{2\times p\times r}{p+r} \tag{20}$$

其中 TP 为预测正确的关键词数量；FP 是被错误预测为关键词的单词数量；FN 表示未检测到的关键词数量。

4.3 与传统模型比较

本文选择两种传统模型作为基准模型：基于统计学的 TF-IDF 模型与基于网络图的 TextRank 模型。同时，我们对使用单向 LSTM-CRF 结构构建的融合模型与使用 BiLSTM-CRF 结构的融合模型进行实验对比。如表 2 所列，可以看到，与传统模型相比，融合模型具有更好的性能，并且 BiLSTM-CRF 的结构可以更加充分地提取词语的关键特征。

表 2 与传统模型对比实验结果

Table 2 Experimental results compared with traditional models

Model	*Precision*		*Recall*		*F1*	
	Test	Dev	Test	Dev	Test	Dev
TF-IDF	39.22	39.09	43.57	43.25	41.28	41.06
TextRank	38.15	38.12	41.80	41.72	39.89	39.84
Fusion model based LSTM	59.98	59.72	50.17	50.01	54.64	54.44
Fusion model based BiLSTM	66.94	66.84	58.01	57.95	62.16	62.08

4.4 特征组合对比实验

本文选择 4 种手工特征：词性(pos)、词频(frq)、词长(len)和词位置(loc)，每种手工特征的组合均会导致不同的模型拟合效果。如表 3 所列，本文尝试了手工特征间的各种组合。其中，emb 代表词嵌入特征。

表 3 特征组合对比实验结果

Table 3 Experimental results of feature combination comparison

Combination of features	*Precision*	*Recall*	*F1*
emb	60.78	52.51	56.35
emb+pos	62.72	53.15	57.54
emb+frq	63.15	54.09	58.27
emb+len	60.92	52.55	56.43
emb+loc	60.77	54.18	57.29
emb+pos+frq	66.98	55.24	60.55
emb+pos+len	62.78	53.09	57.53
emb+pos+loc	61.79	53.05	57.09
emb+frq+len	63.03	53.56	57.91
emb+frq+loc	62.47	55.98	59.05
emb+len+loc	59.79	56.22	57.95
emb+pos+frq+len	63.97	54.27	58.72
emb+pos+frq+loc	66.84	57.95	62.08
emb+pos+len+loc	64.36	56.28	60.05
emb+frq+len+loc	63.69	55.12	59.10
emb+pos+frq+len+loc	68.01	54.19	60.32

通过表 3 分析可得，不同的特征组合可以不同程度地提高模型的效率。

在融入单特征时，词频以及词性对模型的影响最高，而词位置可以有效地提高模型的召回率。当同时融入两种特征时，词性和词频的组合获得了最好的效果；词频和词长的组合反而使模型的效果低于单独融入词频的组合。词频与词位置的组合同样可以在一定程度上提高模型的效果。当加入 3 个特征进行融合时，词性、词频、词位置的组合达到了最优的结果；当为词性、词频的组合加入词长特征后反而降低了模型的效果。4 种特征全部组合后，精确率有了显著提升，但是召回率却无法得到一个较高的值，导致整体评分未能达到最优。

经过分析得到最优的特征组合方式为词嵌入特征、词性特征、词频特征和词位置特征的组合，这些特征使模型的 F1 得分达到了 62.08。并且，我们发现融入词位置可以有效提升模型的召回率，但是融入词长却可能导致模型的效果下降。

4.5 词嵌入对比实验

本文对比了两种流行的词嵌入方式：Word2vec 与 GloVe，二者在融合模型中均有良好的表现。如表 4 所列，当使用 GloVe 模型时，模型的整体效率更高。

表 4 词嵌入对比实验结果

Table 4 Comparative experimental results of word embedding

Word embedding	*Precision*	*Recall*	*F*1
Word2vec	65.59	57.27	61.29
GloVe	66.84	57.95	62.08

4.6 标记方法对比试验

本文提出使用 LMRSN 标记法，这种标记法可以更有效地抽取关键词短语，对整体的使用效果有显著提升。同时，本文也使用 YN 标记进行了测试实验。如表 5 所列，本文提出的 LMRSN 标记法确实能够有效提高融合模型的最终效果。

表 5 标记方法对比实验结果

Table 5 Experimental results comparison of labeling methods

Labeling methods	*Precision*	*Recall*	*F*1
YN	59.76	55.12	57.35
LMRSN	66.84	57.95	62.08

4.7 超参数实验

本文尝试了多种超参数的组合，以获取实验的最佳结果。本文尝试考虑了神经网络层数、每层节点数以及 batch-size 的大小。

将每层神经元个数控制为 100，通过图 6 可以观察到，当神经网络层数为 2，3 层时，神经网络层数对实验结果的影响很大，达到 4 层时模型已经可以充分拟合特征。

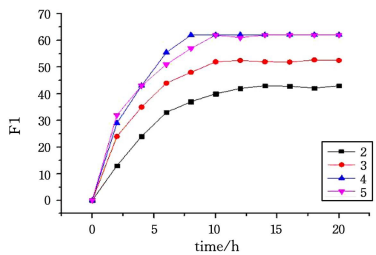

图 6 神经网络层数实验结果

Fig. 6 Experimental results of neural network layer selection

将神经网络层数控制为 4 层，使用 5 组不同的神经元个数进行实验，实验结果如图 7 所示。可以发现神经元个数在 100 时可以达到最佳效果，而神经元数量在 100 以上时，训练时间不断加长，但是训练的效果却没有显著提升。

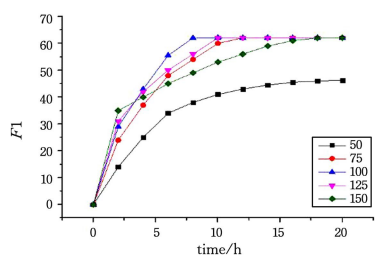

图 7 每层神经元个数实验结果

Fig. 7 Experimental results of number of neuron per layer

我们使用不同的 batch-size 进行实验对比，如表 6 所列，batch-size 不会显著影响实验的结果，但是会影响训练的时间长度。经过实验，我们选择使用 64 作为融合模型最终确定的 batch-size。

表 6 Batch-size 对比实验结果

Table 6 Comparative experimental results of batch-size

Batch-size	Epochs	Times/h
32	45	14
64	29	8
128	18	9
500	9	9.5
1000	7	11

结束语 本文使用 BiLSTM-CRF 结构融合多重手工特征构建的融合模型来完成关键词抽取任务，在与传统方法的对比中，融合模型取得了更好的 F1 分值。我们提取了多种词语手工特征，并将这些特征以一定的组合融入词嵌入特征，并且分析得到的词性、词频、词位置的特征组合可以使我们的模型得到最好的效果。同时，我们分析了词嵌入方式对融合模型的影响。本文提出了一种更加适合关键词抽取的 LMRSN 标记方法，使用这种标记方法可以使融合模型更好地提取词语与短语的关键特征，从而取得更好的模型效果。未来我们将尝试更多手工特征与神经网络结构的融合，以激发深度学习在关键词抽取任务上更大的潜力。

参 考 文 献

[1] SALTON G, BUCKLEY C. Term-Weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.

[2] HUANG L, WU Y P, ZHU Q F. Research and Improvement of TFIDF Text Feature Weighting Method[J]. Computer Science, 2014, 41(6): 204-207.

[3] BESILS R, MOSCHITTI A, PAZIENZA M. A text classifier based on linguistic processing[C]// Proc. of the Int'l Joint Conf. on Artificial Intelligence. UCAI, 1999: 3640.

[4] MIHALCEA R, TARAU P. TextRank: Bringing order into text [C]// Proc. of the EMNLP 2004. Unt Scholarly Works, 2004: 404411.

[5] MA X Z, HOVY E. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv: 1603.01354.

[6] VITERBI A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm[J]. IEEE Transactions on Information Theory, 1967, 13(2): 260-269.

DUAN Jian-yong, born in 1978, Ph.D, professor, is a member of China Computer Federation. His main research interests include natural language processing and so on.