基于 BiLSTM-CRF 的关键词自动抽取

陈 伟[1] 吴友政[2] 陈文亮[1] 张 民[1]
(苏州大学计算机科学与技术学院 江苏 苏州 215006)[1] (爱奇艺人工智能研究组 北京 100080)[2]

摘 要 关键词自动抽取是自然语言处理(Natural Language Processing,NLP)的一项重要任务,给个性化推荐、网购等应用提供了重要的技术支撑。针对关键词自动抽取问题,提出一种新的基于双向长短期记忆网络条件随机场(Bidirectional Long Short-Term Memory Network Conditional Random Field,BiLSTM-CRF)的方法,并将该问题刻画为序列标注问题。首先,该方法通过对输入的文本进行建模,把文本表示为低维高密度的向量;然后,使用分类算法对各个词进行分类;最后,使用 CRF 对整个标注序列进行解码,得到最终结果。在一个大规模的真实数据中进行实验,结果表明该方法较基准系统性能提高约 1 个百分点。

关键词 自然语言处理,关键词抽取,条件随机场,长短期记忆网络
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.06.001

Automatic Keyword Extraction Based on BiLSTM-CRF

CHEN Wei[1] WU You-zheng[2] CHEN Wen-liang[1] ZHANG Min[1]
(School of Computer Sciences and Technology,Soochow University,Suzhou,Jiangsu 215006,China)[1]
(IQIYI Artificial Intelligence Research Group,Beijng 100080,China)[2]

Abstract Automatic keyword extraction is an important task of natural language processing (NLP),which provides technical support for personalized recommendation,online shopping and other applications. For the task,a new keyword extraction method based on bidirectional long short-term memory network and conditional random field (BiLSTM-CRF) was proposed. In the method,the extraction task is regarded as the sequence labeling problem. Firstly,the input text is represented as a low-dimensional,high-density vector. Then,a classification algorithm is used to predict the tags of the words. Finally,a CRF layer is used to decode the whole sequence to get the tagging result. Experiments were conducted on large scale real data,and the results show that this way can improve about 1% compared with the base system.

Keywords Natural language processing,Keyword extraction,Conditional random field,Long short-term memory network

1 引言

关键词一般是单个词或者由多个词组成的短语,是指能反映文本主题或者意思的概括性词或者短语,如论文中的 Keywords 字段、新闻的标签等。把由单个词组成的关键词称作简单关键词(Simple Word,SW);由多个词组成的关键词称作复杂关键词(Complicate Word,CW)。文中将这两种统称为关键词。人们根据文档中提供的关键词,可以快速了解文档内容、把握文档主旨。同时,关键词被广泛应用于新闻报道、科技论文及文献等领域,以便人们高效地管理和检索文档[1]。

进入 Web2.0 时代,关键词的自动抽取已被广泛应用于搜索引擎(如 Google、百度)、新闻服务(如新闻订阅)以及购物网站(如亚马逊、京东、淘宝)。它们根据用户的历史行为,来推荐一些用户感兴趣的广告、新闻和商品等相关服务。同时,关键词在信息检索、文本聚类、分类和文档摘要等 NLP 任务中也发挥着重要作用。例如,在文本聚类时,可以将关键词相似的多篇文档看成一个簇,这样就可以大大地提高 K-Means 聚类的收敛速度;从某天所有新闻中提取出这些新闻的关键词,就可以大致知道那天发生了什么事情。

由此可见,关键词是信息时代人们获取信息、管理和检索资源的重要手段和便捷工具[1]。关键词自动抽取技术为人们在互联网的海量信息中检索知识提供了重要支撑,而个性化推荐技术与关键词自动抽取也有着紧密而重要的联系。

然而,关键词自动抽取面临着两大主要挑战:主观性和复杂性。主观性是指一个标题或者一篇文档,不同人的认知范围不同、看法角度不一,导致其对某一类型或题材的标题的偏好也不同,这样就会影响训练数据的标注质量。给出如下例子:

标题:《熟悉的问道 2》为了收视率岳云鹏当场倒立

关键词:孙坚 李咏 熟悉的味道 2 岳云鹏 收视率 倒立

其中,"收视率"和"倒立"是否作为关键词,不同人有不同的看法。而"孙坚"和"李咏"并没有在标题中出现,而是根据

本文受国家自然科学基金资助项目(61572338),江苏省高校自然科学研究重大项目(16KJA520001),CCF-腾讯科研基金资助。

陈 伟(1989-),男,硕士生,主要研究方向为自然语言处理,E-mail:947869167@qq.com;吴友政(1976-),男,博士,主要研究方向为自然语言处理、信息抽取、语音识别、预测等,E-mail:wuyouzheng@qiyi.com;陈文亮(1977-),男,博士,教授,主要研究方向为自然语言处理、推荐系统、信息抽取、知识图谱,E-mail:wlchen@suda.edu.cn(通信作者);张 民(1970-),男,博士,教授,主要研究方向为自然语言处理、机器翻译、人工智能,E-mail:minzhang@suda.edu.cn。

标题内容打上的。我们把这种基于标题内容打上的标签称为抽象标签。抽象标签并不是本文所研究的目标,本文所研究的关键词均来自于标题。

复杂性是针对一些有歧义的词,即在一个领域下是普通词,在另一个领域下可能是专用名词。如“传奇”这个词,在日常生活中可能就是一个普通词;但在游戏领域,可能就是一款游戏的名字。给出如下例子:

标题:传奇好玩的游戏,召唤小伙伴!

关键词:传奇

针对关键词抽取这个任务,传统方法大致可分为无监督方法和有监督方法。无监督方法主要是利用 TFIDF 等统计信息来寻找重要词。有监督方法主要是在一个有标注的数据集上训练一个分类器,将关键词抽取任务转化为二分类问题,也就是判断每个候选关键词是否为关键词的二分类问题。有监督方法能综合利用更多的信息,比无监督方法有更大的优势,实验效果也较好[2-3]。但是,把关键词自动抽取任务看作分类问题存在一些问题,其中最主要的问题是它对每个候选词进行单独处理,忽略了文本中句子结构的有效信息,造成模型分类的性能较差。

针对分类思想解决此任务的不足,本文将关键词抽取任务转化为序列标注问题。本文基于双向 LSTM 的深度学习框架,结合 CRF 模型,构建新的关键词自动抽取系统。在本文的方法中,不需要构建人工特征模板和规则,因而可以方便、快捷地构建关键词自动抽取系统。在大规模的真实数据上的实验结果表明,双向 LSTM-CRF 模型能够获得比传统 CRF 模型更好的效果。

2 相关工作

目前,关键词抽取主要有两种方式:1)关键词分配,即预先定义一个关键词词库,对于一篇文章,从词库中选取若干词语作为文章的关键词;2)关键词抽取,即从文章的内容中寻找一些词语作为推荐关键词[1,4]。

对于关键词分配,一般要求词库是某个或某些领域的专业词汇,或者看作是与一个或多个领域相关的专业词典。这些词典一般都是由专家手工编纂的,有质量保证,但费时费力,而词典的大小和覆盖度决定了关键词分配的范围和效果。当切换到一个新的领域时,又需要重新构建词典,无法满足如今网络时代的大规模应用和推广需求。

对于关键词抽取,大致可分为无监督方法和有监督方法。无监督方法会利用 TFIDF 等统计信息,选取 top K 作为关键词[5-7]。这些方法无需人工标注训练集合的过程,因此更加快捷,但无法有效地综合利用词法和语义信息对候选关键词进行排序。而在有监督方法中,将关键词抽取问题转换为判断每个候选关键词是否为关键词的二分类问题[8-11],它需要一个已经标注关键词的文档集合来训练分类模型,目标是在一个有标注的数据集上训练一个分类器,以便决定候选词中哪些是关键词。不同的机器学习算法可以训练出这样一个分类器,如贝叶斯算法[12]、决策树算法[4]、bagging[13]、boosting[14]、最大熵算法[15]、多层感知机[16]和 SVM 算法[17]。但是,把关键词抽取问题看作分类问题存在一些问题,最主要的问题是它对每个候选词进行单独处理,忽略了文本中句子结构的有效信息,造成模型分类的性能较差。

基于分类思想解决此任务的不足,另外一种思路是将关键词自动抽取任务转化为序列标注问题来解决。传统的最常用的解决序列标注问题的方案是隐马尔可夫(Hidden Markov Model,HMM)、最大熵(Maximum Entropy,ME)和条件随机场(Conditional Random Fields,CRF)等模型。其中 CRF 是目前解决序列标注问题最主流的做法,性能也最好,目前已被广泛应用于 NLP 的各种任务中,如分词、词性标注、命名实体识别等,并且取得了非常好的效果。基于此,本文也将 CRF 应用于关键词自动抽取任务中,并将其作为基准系统。

但是,诸如 CRF 等传统的机器学习算法往往依赖人工设计的特征,而一个特征是否有效往往需要多次尝试与选择,因此人工设计一系列好的特征既费时又费力,而模型的好坏与特征工程的构建有很大关系。

近些年,随着深度学习的兴起,其已被广泛应用于 NLP 的各种任务中,如分词、词性标注、命名实体识别、情感分析等,且取得了一定的成果。长短期记忆网络(Long Short-Term Memory Networks,LSTM)作为其中的代表,对处理诸如分词、词性标注、命名实体识别等长序列依赖问题非常有效[18],具有天然的优势。LSTM 会对前面的信息进行记忆并将其应用于当前输出的计算中,而且隐藏层之间的节点是有连接的,这与传统的神经网络模型不同。同时,隐藏层的输入不仅包括输入层的输出,还包括上一时刻隐藏层的输出。而结合 LSTM 网络和 CRF 网络,通过 LSTM 层可以高效地使用前后上下文的特征,通过 CRF 层使用标签信息,综合利用多种信息,使性能更好[18]。

3 数据预处理

3.1 构建黑名单词典

在构建黑名单词典前,首先将线上人工标注的所有标签导出,对其进行分析,如词性、长度等,同时考虑主观及客观等原因。发现以下几个问题:

1)一些普通词可能是歌名、人名或者电影名等,如“我的”“高兴”;

2)标注人员在标注过程中可能误标,如把一些普通词当作关键词标出;

3)一些游戏 APP 及一些解说的名字有歧义,如“传奇”“一一”,去除这些有歧义或者普通的标签,避免给模型引入一些噪音。

通过上述对数据的分析,针对发现的问题,安排两人分别对同一份导出的标签文件构建黑名单,可以利用各种信息(如标签的长度、标签的词性及类型、标签出现的次数及标注的次数等)加以识别;最后,取两人识别的黑名单词典的交集作为最终结果。

3.2 自动补充标注关键词

对数据进行补充标注,主要基于如下考虑:1)标注人员对视频标题进行标注时,带有一定的个人主观性;2)无法保证标注人员在标注数据过程中时时刻刻都全神贯注、集中注意力,易出现误标、漏标等现象;3)标注人员的认知范围不同,即标注人员平时所关注的领域不同,导致对某一类型或题材的视频标题的偏好也不同。补充标注主要针对上述这些情况,目的是丰富视频标题的标注信息,以便让模型学习到更多的知识。补充标注所用的词典是根据训练数据在线计算得到的,公式如下:

$$ReceiveRate(W)=LabelNum(W)/TotalNum(W) \quad (1)$$

其中，W 表示由标注人员标注的 SW/CW；$ReceiveRate(W)$ 表示 W 的接受率；$LabelNum(W)$ 表示 W 在训练数据中被标注的个数；$TotalNum(W)$ 表示 W 在训练数据中出现的总次数。

根据上述公式，统计在训练数据中出现的所有的 SW/CW，然后设定阈值，选取高接受率的 W 作为补充标注的词典。在其他条件相同的情况下，通过对比实验可以发现，对训练数据进行自动补充标注对结果的提升很大，不论是对 SW 层面的抽取还是 CW 层面的抽取，F 值均提升了 10 余个百分点，效果尤其显著。

4　基于 LSTM-CRF 的抽取模型

4.1　抽取框架

在本文中，关键词的自动抽取框架如图 1 所示。

图 1　关键词抽取框架

4.2　基于序列标注模型的关键词抽取方法

针对关键词自动抽取这个任务，已有的研究要么是基于 TFIDF 等统计信息的无监督方法，要么是基于分类思想的有监督方法。如引言所述，分类模型对每个候选词进行单独处理，忽略了文本中句子结构的有效信息。为了更好地综合利用更多信息，如词法、句法等，本文对关键词抽取重新建模，将之作为一个序列标注问题。针对序列标注模型的求解，在传统的机器学习方法中，表现较好的是条件随机场方法，即 CRF 方法；在神经网络方法中，表现较优的是循环神经网络模型，即 RNN(Recurrent Neural Networks)模型。

我们使用了 5-Tag 标签集合，即 B-WORDTAG，B-PHRASETAG，I-PHRASETAG，E-PHRASETAG，O，其中，B-WORDTAG 表示 SW，即单个词的标签，假设句子中 SW 是“刘德华”，则对应的标注是“刘德华/B-WORDTAG”。而 B-PHRASETAG，I-PHRASETAG 和 E-PHRASETAG 分别表示 CW 的开头、中间和结尾，假设句子中 CW 是“北京 天安门城楼”，则对应的标注是“北京/B-PHRASETAG 天安门/I-PHRASETAG 城楼/E-PHRASETAG”。O 表示 other，也就是除 SW 和 CW 之外的其他词所要标注的标签。

4.3　BiLSTM-CRF 模型

用传统方法进行关键词自动抽取时，需人工构建大量的特征工程，费时费力，相当复杂。而相较于传统的基于统计的机器学习方法，神经网络主要是用词向量作为系统输入，可以避免人工构建特征工程。词向量是在一个连续低维空间中表示词，该向量可以捕获词的语义和句法信息，即相似的词有相似的向量表示，如 v(“国王”)-v(“王后”) ≈v(“男”)-v(“女”)。此外，词向量已经广泛应用于 NLP 的各种任务中，且均得到较传统方法更好的结果[19-23]。

因此，在最近几年里，词向量作为词法或者语义特征，被广泛应用在 NLP 任务中。词向量是在一个较大的无标注海量数据上训练的，这一特性使得它在一个新领域或者新语言上仍然可以使用。

4.3.1　LSTM 网络

RNN(Recurrent Neural Networks)已被广泛应用于 NLP 的各种任务中，如词向量表示、词性标注、语音识别等。RNN 能保存历史信息并将其应用到当前输出中，这些特性对序列标注问题的解决很有帮助。图 2 是一个 RNN 结构，包括一个输入层 Input、一个隐藏层 Hidden 和一个输出层 Output，其中 Input 代表输入特征，Output 代表标签结果。

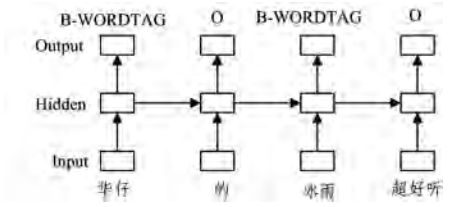

图 2　RNN 结构图

输入层的词特征可以用 one-hot 向量或者低维稠密的向量表示，无论用哪一种向量表示，每个词的维度都是相同的。输出层是关于标签的概率分布，它用相同的维度作为标签的大小。与传统神经网络相比，RNN 隐藏层的输入不仅包括输入层的输出，还包括上一时刻隐藏层的输出，这一层被用来存储历史信息。隐藏层和输出层可用如下公式计算：

$$\begin{aligned} h(t)&=f(Ux(t)+Wh(t-1))\\ y(t)&=g(Vh(t)) \end{aligned} \tag{2}$$

其中，U，W 和 V 为权重，是模型要学习出的参数；f 和 g 分别为 sigmoid 和 softmax 激活函数，如下：

$$\begin{aligned} f(z)&=\frac{1}{1+e^{-z}}\\ g(z_m)&=\frac{e^{z_m}}{\sum_k e^{z_k}} \end{aligned} \tag{3}$$

本文中所说的 Long Short-Term Memory Networks (LSTM)和 RNN 在结构上并没有什么不同，只是使用了不同的函数去计算隐藏层的状态。图 3 给出 LSTM 的一个 cell。

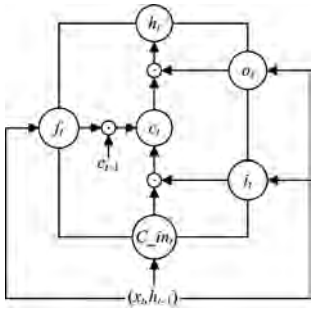

图 3　LSTM 的 cell 图

一个完整的 LSTM 的 cell 可由下式表示：

$$\begin{aligned} i_t&=g(W_{xix_t}+W_{h_i}h_{t-1}+b_i)\\ f_t&=g(W_{xfx_t}+W_{hf}h_{t-1}+b_f)\\ o_t&=g(W_{xox_t}+W_{ho}h_{t-1}+b_o)\\ c_t&=f_tc_{t-1}+i_t\tanh(W_{xcx_t}+W_{h_c}h_{t-1}+b_c)\\ h_t&=o_t\tanh(c_t) \end{aligned} \tag{4}$$

其中，g 代表激活函数；i，f，o 和 c 分别代表输入门、遗忘门、输出门和最后 cell，它们与 h 具有相同的维度大小。图 4 给出了 LSTM 的序列标注模型，使用了上述提及的 cell。

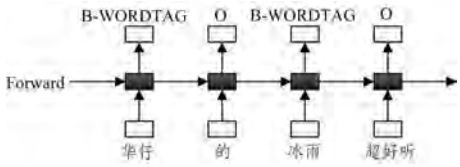

图 4　LSTM 序列标注框架图

4.3.2　CRF 网络

我们不仅使用 BiLSTM 对标签建模，而且结合使用 CRF

网络对模型进行共同建模。对于一个给定的输入序列 X，我们想得到一个预测序列 y，定义如下得分函数[23]：

$$s(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{5}$$

其中，A 和 P 分别对应转移分数矩阵和经过 BiLSTM 网络输出的分数矩阵。

在训练的过程中，我们要最大化正确标签序列的似然概率，如下：

$$\log(p(y|X))=s(X,y)-\log(\sum_{\bar{y}\in Y_X}e^{s(X,\bar{y})}) \tag{6}$$

其中，Y_X 是指一个输入序列 X 对应的所有可能的标签序列。在解码时，利用动态规划算法，如 Viterbi 算法，通过如下公式预测其最大得分的输出序列：

$$y^*=\arg\max_{\bar{y}\in Y_X}s(X,\bar{y}) \tag{7}$$

4.3.3 LSTM-CRF 网络

在序列标注任务中，对于某一给定的状态，我们会用到它过去的信息和将来的信息。因此，对于特定状态，我们可以使用双向的 LSTM，用前向状态来使用它过去信息，用后向状态来使用它将来的信息。双向 LSTM 和 CRF 结合而成的 BiLSTM-CRF 模型如图 5 所示。

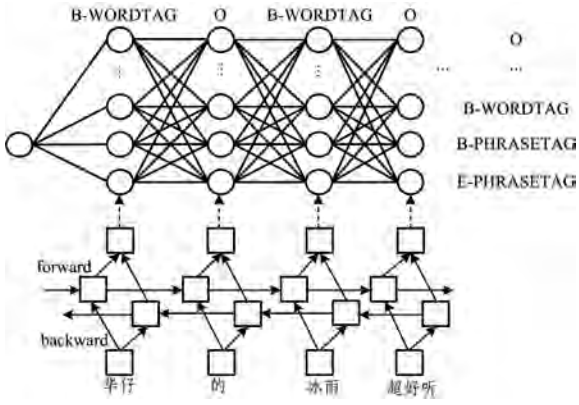

图 5 BiILSTM-CRF 系统框架图

此模型通过 LSTM 层可以充分利用过去的信息和将来的信息，通过 CRF 层能利用标签信息，即可以充分利用过去和将来的标签信息来预测当前的标签。

5 实验设计与结果

5.1 训练数据

本文使用的实验数据均是爱奇艺线上的图文视频标题，标题的长度多集中在 20～50 字。标注人员标注的原始训练数据规模约有 660k，但是由于标注人员在标注数据时不一定每时每刻都全神贯注、集中注意力，会出现误标、漏标等情况；同时，数据也有噪声，会出现错别字、句子不通顺等现象；另外，每个人的认知范围不同，如有些标注人员对娱乐、综艺感兴趣，有些标注人员对游戏、电竞感兴趣。以上种种原因，会导致数据质量也并不十分理想。

鉴于上述原因，首先，过滤掉特别长的视频标题，如长度大于 100 字的标题。主要考虑如下：1)一般正常的标题长度是 20～50 字，大于 100 字后，不太正常，可作为垃圾数据扔掉；2)若标题太长，模型在不能学习太多信息的情况下，可能还会引入噪音。其次，过滤掉带抽象标签的数据，即根据标题的内容打上的标签，此标签并没有在标题中出现。如下列例子：

标题：快来看，你的偶像是哪位？

关键词：黄家驹 刘德华 张学友 范冰冰 章子怡 舒淇

5.2 测试数据

从线上图文视频标题随机采样 1.6k 数据，并且让 3 人同时标注这份数据，将 3 人标注结果的交集作为本实验的测试数据。除了对测试数据进行中文分词处理外，不进行任何数据预处理操作。

5.3 评测方法

本文采用 Min-Max 集合评测方法，将 3 人标注结果的交集和并集作为 ground-truth，在计算正确率时，使用最大集合；在计算召回率时，使用最小集合。评价标准为正确率(Precision)、召回率(Recall)、F 值(Fscore)。选择 Min-Max 集合评测方法的原因是：1)线上业务的需求及对用户反馈数据的分析；2)对标注人员标注结果的分析。

5.4 基于 CRF 模型的基准系统

条件随机场模型是 Lafferty 等[24]在最大熵模型和隐马尔可夫模型的基础上提出的一种无向图学习模型，是一种用于标注和切分有序数据的条件概率模型。CRF 在一系列序列标注任务上取得了非常好的效果，因此将它作为基准系统。本文使用 crfsgd 工具来实现我们的基准系统，采用默认参数。按照 CRF 工具的语料格式要求，针对 SW 层面和 CW 层面这两种识别对象，同时结合对数据的分析和任务的需求，我们做了大量的特征工程，构建了非常强壮的特征模板。模板的基本格式为%x[行、列]，它用于确定输入数据中的一个词例。行，表示%x 相对于当前词例的行数；列，表示%x 在列上的绝对列数。以“京东/nz 618/m 单品/n 销量/n：/w 坚果/n pro/n 成/v 黑马/n NUL/w iphone/n 落后/a”这个标题来作为特征模板示例，假设当前词为“坚果”，本研究所采用的 CRF 特征训练模板如表 1 和表 2 所列。

表 1 特征列

text	part-of-speech	word weight	word length	the beginning of the word	label
京东	nz	25	2	1	B-PHRASETAG
618	m	12	3	0	E-PHRASETAG
单品	n	11	2	0	B-OTHER
销量	n	14	2	0	B-OTHER
：	w	0	1	0	B-OTHER
坚果	n	15	2	0	B-PHRASETAG
pro	n	11	3	0	E-PHRASETAG
成	v	0	1	0	B-OTHER
黑马	n	12	2	0	B-OTHER
NUL	w	0	1	0	B-OTHER
iphone	n	13	6	0	B-WORDTAG
落后	a	11	2	0	B-OTHER

表 2　特征模板及示例

以“坚果”为例		
Unigram		
word feature	U00:%x[-2,0]	销量
	U01:%x[-1,0]	:
	U02:%x[0,0]	坚果
	U03:%x[1,0]	pro
	U04:%x[2,0]	成
	U05:%x[-1,0]/%x[0,0]	:/坚果
	U06:%x[0,0]/%x[1,0]	坚果/pro
pos feature	U10:%x[-2,1]	n
	U11:%x[-1,1]	w
	U12:%x[0,1]	n
	U13:%x[1,1]	n
	U14:%x[2,1]	v
	U15:%x[-2,1]/%x[-1,1]	n/w
	U16:%x[-1,1]/%x[0,1]	w/n
	U17:%x[0,1]/%x[1,1]	n/n
	U18:%x[1,1]/%x[2,1]	n/v
	U181:%x[-1,1]/%x[1,1]	w/n
	U19:%x[-2,1]/%x[-1,1]/%x[0,1]	n/w/n
	U190:%x[-1,1]/%x[0,1]/%x[1,1]	w/n/n
	U191:%x[0,1]/%x[1,1]/%x[2,1]	n/n/v
word weight feature	U20:%x[-1,2]	“:”的权重
	U21:%x[0,2]	“坚果”的权重
	U22:%x[1,2]	“pro”的权重
	U23:%x[-1,2]/%x[0,2]	“:”的权重/“坚果”的权重
	U24:%x[0,2]/%x[1,2]	“坚果”的权重/“pro”的权重
cross feature	U50:%x[0,1]/%x[0,2]	n/“坚果”的权重
	U51:%x[0,1]/%x[0,3]	n/“坚果”的词长度
	U52:%x[0,2]/%x[0,3]	“坚果”的权重/“坚果”的词长度
	U53:%x[0,1]/%x[0,2]/%x[0,3]	n/“坚果”的权重/“坚果”的词长度
	U54:%x[0,1]/%x[0,2]/%x[0,4]	n/“坚果”的权重/是否是起始位置(是:1;否:0)
	U55:%x[0,2]/%x[0,4]	“坚果”的权重/是否是起始位置(是:1;否:0)
	U58:%x[0,1]/%x[0,4]	n/是否是起始位置(是:1;否:0)
Bigram	B	

5.5　BiLSTM-CRF 模型

5.5.1　词向量

与英文文本不同，中文文本并不是事先分好词的。对于每一个词的词向量，无法在一个未经分词的语料上训练得到。对于中文分词，本文使用的是爱奇艺 NLP 团队自己开发的中文分词工具。本文使用预训练好的词向量，大小为 400 维。

5.5.2　词性特征

由于大部分关键词符合一定的词性模式[13]，如“形容词＋名词”是最常见的模型，因此可将词性作为一个很强的分类特征。对比实验结果如表 3 所列，all-pos 指所有词性。

表 3　BiLSTM-CRF 中加入词性的对比实验(1)

		exp1	exp2	exp3	exp4
Experiment-Method		LSTM	BiLSTM	BiLSTM-CRF	BiLSTM-CRF
Train-data		230 k	230 k	230 k	230 k
Test-data		1.6 k	1.6 k	1.6 k	1.6 k
Add_pos_or_not		no	no	no	yes(all-pos)
SW	Recall/%	85.3	84.3	84.5	86.1
	Precision/%	73.5	75.3	85.5	86.6
	Fscore/%	78.9	79.6	85.0	**86.4**
CW	Recall/%	31.1	47.6	52.4	54.7
	Precision/%	48.0	51.6	64.6	62.7
	Fscore/%	37.8	49.5	57.8	**58.4**

PS：LSTM 和 BiLSTM 实验参数设置为 *RANDOM_SEED*=1337，*MAX_SEQUENCE_LENGTH*=100，*WORD_EMBEDDING_DIM*=400，*BATCH_SIZE*=64，*Loss*=cross entropy，*ACTIVATION*=tanh，*Optimizer*=RMSprop，*LEARNING_RATE*=0.001；

BiLSTM-CRF 实验参数设置为 *RANDOM_SEED*=1337，*MAX_SEQUENCE_LENGTH*=100，*WORD_EMBEDDING_DIM*=400，*BATCH_SIZE*=100，*Loss*=crf.sparse_loss，*ACTIVATION*=tanh，*Optimizer*=RMSprop，*LEARNING_RATE*=0.001，下同。

通过对比 exp1，exp2 和 exp3 的实验数据可以发现，BiLSTM-CRF 在 SW 层面和 CW 层面的性能具有较为显著的提高。通过对比 exp3 和 exp4 可以发现，在 SW 层面上，Fscore 从 85.0%提高到 86.4%，提升了 1.4 个百分点；在 CW 层面上，Fscore 从 57.8%提高到 58.4%，提升了 0.6 个百分点。数据表明，加入词性这一特征，可以提高模型的性能。

此外，通过对比分析实验结果，综合对人工标注的标签及词性等分析发现，SW/CW 的词性多集中在名词等词性上，而

其他一些词性很少涉及。因此，将人工标注的标签按词性的个数由多到少排序，选择 top10 的词性，其他词性为 other，再次进行对比实验，结果如表 4 所列，其中 top10-p05 指 top10 的词性。

表 4 BiLSTM-CRF 中加入词性的对比实验(2)

		exp1	exp2	exp3	exp4	exp5
Experiment-Method		LSTM	BiLSTM	BiLSTM-CRF	BiLSTM-CRF	BiLSTM-CRF
Train-data		230 k	230 k	230 k	230 k	230 k
Test-data		1.6 k	1.6 k	1.6 k	1.6 k	1.6 k
Add_pos_or_not		no	no	no	yes(all-pos)	yes(top10-pos)
SW	Recall/%	85.3	84.3	84.5	86.1	86.8
	Precision/%	73.5	75.3	85.5	86.6	86.7
	Fscore/%	78.9	79.6	85.0	**86.4**	**86.7**
CW	Recall/%	31.1	47.6	52.4	54.7	55.9
	Precision/%	48.0	51.6	64.6	62.7	63.6
	Fscore/%	37.8	49.5	57.8	**58.4**	**59.5**

exp4(all-pos)与 exp5(top10-pos)表明，相较于加入全部词性，加入标签的 top10 词性对 SW/CW 更有效，较之前未加词性，性能提升了 1.7 个百分点。

5.5.3 主要实验结果

从实验数据(见表 5)上看，无论在 SW 层面还是 CW 层面，BiLSTM-CRF 模型均比基准系统 CRF 模型提高了 0.9 个百分点。

表 5 实验结果

		Baseline-CRF Model	BiLSTM-CRF Model
Train-data		230 k	230 k
Test-data		1.6 k	1.6 k
Add_feature-POS		yes	yes(top10-pos)
SW	Recall/%	84.5	86.8
	Precision/%	87.1	86.7
	Fscore/%	85.8	**86.7**
CW	Recall/%	44.1	55.9
	Precision/%	87.0	63.6
	Fscore/%	58.6	**59.5**

从测试数据的对比结果来看，BiLSTM-CRF 能够召回 CRF 得不到的标签，尤其是 CW，如“海尔洲际酒店”“绵阳米粉”等，可以看出，BiLSTM-CRF 对 CW 的召回，提升效果更好。分析结果发现，通过神经网络模型还能在测试数据中发现一些好的标签，如“假唱”“男神”等，但是标注人员并没有在测试数据中将这些标签标出。

结束语 本文旨在研究关键词自动抽取任务，并将该任务建模为序列标注问题。基于 BiLSTM-CRF 神经网络框架，本文提出了一种新的关键词自动抽取方法。实验结果表明，本文所构建的系统能够获得比基准系统 CRF 更好的效果。

从实验结果可以看出，关键词自动抽取任务仍然具有很大的挑战。目前的方法取得的效果还非常有限，在 CW 层面上的 F 值不足 60%。下一步工作中，我们将重点解决错误分析中部分 SW 无法抽取的情况和 CW 层面上的抽取问题，通过对这两种情况进行研究和对模型进行改进，进一步提升关键词自动抽取的性能。

参 考 文 献

[1] 刘知远. 基于文档主题结构的关键词抽取方法研究[D]. 北京：清华大学，2011.

[2] MARUJO L, WANG L, TRANCOSO I, et al. Automatic keyword extraction on twitter[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers). USA: ACL, 2015: 637-643.

[3] GOLLAPALLI S D, LI X L, YANG P. Incorporating Expert Knowledge into Keyphrase Extraction[C]// Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17). California: AAAI, 2017: 3180-3187.

[4] TURNEY P D. Learning Algorithms for Keyphrase Extraction [J]. Information Retrieval, 2000, 2(4): 303-336.

[5] WU W, ZHANG B, OSTENDORF M. Automatic generation of personalized annotation tags for twitter users[C]// Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics(HLT'10). USA: ACL, 2010: 689-692.

[6] ZHAO W X, JIANG J, HE J, et al. Topical keyphrase extraction from twitter[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies(HLT'11). USA: ACL, 2011: 379-388.

[7] BELLAACHIA A, AL-DHELAAN M. Ne-rank: A novel graph-based keyphrase extraction in twitter[C]// The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology(WI-IAT'12). Washington, DC: IEEE Computer Society, 2012: 372-379.

[8] RILOFF E, LEHNERT W. Information extraction as a basis for high-precision text classification[J]. ACM Transactions on Information Systems(TOIS), 1994, 12(3): 296-333.

[9] WITTEN I H, PAYNTER G W, FRANK E, et al. Kea: practical automatic keyphrase extraction[C]// 4th ACM Conference on Digital Libraries(DL'99). New York: ACM, 1999: 254-255.

[10] MEDELYAN O, PERRONE V, WITTEN I H. Subject metadata support powered by maui[C]// 10th Annual Joint Conference on Digital Libraries(JCDL'10). New York: ACM, 2010: 407-408.

[11] WANG C, LI S J. Corankbayes: Bayesian learning to rank under the co-training framework and its application in keyphrase extraction[C]// 20th ACM International Conference on Information and Knowledge Management (CIKM'11). New York: ACM, 2011: 2241-2244.

[12] FRANK E, PAYNTER G W, WITTEN I H, et al. Domain-specific Keyphrase Extraction[C]// Proceedings of IJCAI. California: AAAI, 1999: 668-673.

[13] HULTH A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge [C] // Proceedings of EMNLP. USA: ACL, 2003: 216-223.

[14] HULTH A, KARLGREN J, JONSSON A, et al. Automatic keyword extraction using domain knowledge[C]// 2nd International Conference on Computational Linguistics and Intelligent Text Processing. Mexico City: Springer-verlag, 2001: 472-482.

(下转第 113 页)

由于随机游走模型是布朗运动理想的数学模型，可以应用于互联网链接分析和金融股票市场，故测试数据使用随机游走模型产生 5 条拥有 1000 个数据点的时间序列，对 5 个时间序列进行相似性比较。算法在同一台计算机上运行，使用 Python 为算法的设计软件，对 DTW 和 MALRDTW 算法进行测试，算法的运行时间及相似比较结果如表 1 所列。表 1 的实验结果说明，DTW 与 MALRDTW 动态时间弯曲距离算法在时间序列上进行相似度度量时结果相近，说明二者之间的准确度近似。但对于算法的运行效率而言，MALRDTW 距离算法远高于 DTW 算法，约提高 96%，这一点在高维度时间序列相似性比较中表现得更为突出。

表 1　DTW 距离算法与 MALRDTW 距离算法的实验结果比较

距离度量方法	DTW 距离算法			MALRDTW 距离算法		
	累积距离	相对距离比	CPU 运行时间/s	累积距离	相对距离比	CPU 运行时间/s
1	10.056	0.0000	0.031	506.346	0.0000	0.001
2	56.209	5.5896	0.032	4320.58	8.5329	0.001
3	87.440	8.6953	0.031	4564.18	9.0140	0.001
4	52.130	5.1840	0.035	4250.00	8.3935	0.004
5	140.500	13.9718	0.035	6977.24	13.7796	0.001

结束语　针对时间序列相似性比较中欧氏距离对序列的异常数据敏感和动态时间弯曲距离时间复杂度为$O(mn)$的问题，提出基于滑动平均与分段线性回归的时间序列相似性算法。算法利用初始可变滑动平均算法对原始时间序列进行预处理，消除了由原始时间序列中的异常数据带来的不利影响，同时使得时间序列更加平滑；在滑动平均序列上提取极值特征点，并以特征点所对应的时间点对原始时间序列进行子序列划分，应用线性回归算法对子序列进行处理；将线性回归的截距和斜率作为原始时间序列的特征序列，实现了数据降维处理。

使用滑动平均与分段线性回归处理后的动态时间弯曲距离算法(MALRDTW)取得了与 DTW 算法相近的相似性比较性能，但是在算法效率上明显优于 DTW 距离算法。

参 考 文 献

[1] 李海林，杨丽彬. 时间序列数据降维和特征表示方法[J]. 控制与决策，2013，28(11)：1718-1722.

[2] Al-NAYMAT G，TAHERI J. Effects of dimensionality reduction techniques on time series similarity measurements[C]// IEEE/ACS International Conference on Computer Systems and Applications. Piscataway：IEEE，2008：188-195.

[3] KEOGH E，RATANAMAHATANA C A. Exact indexing of dynamic time warping[J]. Knowledge and Information Systems，2005，7(3)：358-386.

[4] KEOGH E，PAZZANI M. Derivative dynamic time warping[C]// The First SIAM International Conference on Data Mining. Washington：IEEE，2001：1-11.

[5] 王达，荣冈. 时间序列的模式距离[J]. 浙江大学学报工学版，2004，38(7)：795-798.

[6] DONG X L，GU C K，WANG Z O. Study on Time Series Similarity Measurement Based on Morphology [J]. Journal of Electronics & Information Technology，2007，29(5)：1228-1231.

[7] 张鹏，李学仁，张建业，等. 时间序列的夹角距离及相似性搜索[J]. 模式识别与人工智能，2008，21(6)：763-767.

[8] 陆薛妹，胡轶，方建安. 基于分段极值 DTW 距离的时间序列相似性度量[J]. 微计算机信息，2007，23(27)：204-206.

[9] 李海林，郭崇慧，杨丽彬. 基于分段聚合时间弯曲距离的时间序列挖掘[J]. 山东大学学报(工学版)，2011，41(5)：57-62.

[10] 朱天，白似雪. 基于模式距离度量的时间序列相似性搜索[J]. 微计算机信息，2007，23(30)：216-217.

[11] RABINER L，JUANG B H. Fundamentals of Speech Recognition[J]. Tsinghua University Press，1993，1(1)：353-356.

[12] VULLINGS H J L M，VERHAEGEN M H G，VERBRUGGEN H B. ECG segmentation using time-warping[M]// Advances in Intelligent Data Analysis Reasoning about Data. Springer Berlin Heidelberg，1997：275-285.

[13] BERNDT D J，CLIFFORD J. Using dynamic time warping to find patterns in time series[C]// KDD workshop. Seattle：AAAI Press，1994：359-370.

[14] BERNDT D J，CLIFFORD J. Finding patterns in time series：a dynamic programming approach[C]// Advances in Knowledge Discovery & Data Mining. Washington：American Association for Artificial Intelligence，1996：229-248.

(上接第 96 页)

[15] KIM S N，KAN M Y. Re-examining automatic keyphrase extraction approaches in scientific articles[C]// Proceedings of the ACL-IJCNLP Workshop on Multiword Expressions. USA：ACL，2009：9-16.

[16] LOPEZ P，ROMARY L. HUMB：Automatic key term extraction from scientific articles in GROBID[C]// Proceedings of the 5th International Workshop on Semantic Evaluation. Sweden：ACM，2010：248-251.

[17] JIANG X，HU Y H，LI H. A ranking approach to keyphrase extraction[C]// 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：ACM，2009：756-757.

[18] HUANG Z H，XU W，YU K. Bidirectional LSTM-CRF Models for Sequence Tagging(arXiv)(Version1. 0)[OL]. https://arxiv. org/abs/1508. 01991.

[19] BENGIO Y，DUCHARME R，VINCENT P，et al. A neural probabilistic language model[J]. Journal of Machine Learning Research，2003，3(6)：1137-1155.

[20] COLLOBERT R，WESTON J，BOTTOU L，et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research，2011，12(1)：2493-2537.

[21] MIKOLOV T，YIH W T，ZWEIG G. Linguistic regularities in continuous space word representations[C]// NAACL-HLT. USA：ACL，2013：746-751.

[22] LEVY O，GOLDBERG Y，DAGAN I. Improving distributional similarity with lessons learned from word embeddings[J]. Transactions of the Association for Computational Linguistics，2015，75(3)：211-225.

[23] LAMPLE G，BALLESTEROS M，SUBRAMANIAN S，et al. Neural Architectures for Named Entity Recognition (arXiv)(Version3. 0)[OL]. https://arxiv. org/abs/1603. 01360.

[24] LAFFERTY F，MCCALLUM A，PEREIRA F. Conditional Random Fields：Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of ICML-2001. New York：ACM，2001：282-289.