文章编号：1003-0077(2018)12-0084-08

多特征融合的汉越双语新闻摘要方法

叶 雷，余正涛，高盛祥，刘书龙，张亚飞

（昆明理工大学 信息工程与自动化学院，云南 昆明 650500）

摘 要：为了获取同一事件的汉越双语新闻的自动摘要，该文提出了一种多特征融合的汉越双语新闻摘要方法。关于同一事件的新闻文本，其句子间具有一定的关联关系，利用这些关联关系有助于生成摘要。根据该思想，首先计算句子间的新闻要素共现程度及句子间的相似度；然后将这两种特征融入句子无向图，并利用图排序算法对句子进行排序；之后结合句子的位置特征对排序结果进行调序；最后挑选重要句子并去除冗余生成摘要。在汉越双语新闻文档集上进行了摘要实验，结果表明该方法取得了较好的结果，具有有效性。

关键词：双语新闻；多特征；句子无向图；自动摘要

中图分类号：TP391 **文献标识码**：A

A Bilingual News Summarization in Chinese and Vietnamese Based on Multiple Features

YE Lei, YU Zhengtao, GAO Shengxiang, LIU Shulong, ZHANG Yafei

(School of Information Engineering and Antomation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: In order to generate a summary for a news event reported in both Chinese and Vietnamese, a multi-feature fusion method for bilingual news summarization is proposed. It employs the cross-lingual correlations between sentences in the news text. Firstly, this method analyzes the co-occurrence degree of news elements and the similarity between sentences. Then, these two features are integrated into an undirected graph and a ranking algorithm is used to sort sentences. Finally, important sentences are selected and the redundancy is removed to generate a summary. Experiment on the Chinese and Vietnamese bilingual news archive shows that the proposed method achieved good results.

Keywords: bilingual news; multi-feature; undirected sentence graph; automatic summarization

0 引言

随着互联网技术的发展，网络上每天都会生成大量的文本数据，从这些数据中获取有用的信息变得越来越难。自动摘要技术利用计算机对文档进行处理，生成包含原文档核心内容的摘要，实现对文档的压缩，是解决信息爆炸问题的有效方法。随着“一带一路”倡议的提出，中越两国的交流变得愈发密切。关于一些重要的新闻事件，两国媒体会发布大量的汉语新闻和越南语新闻。若能利用自动摘要技术处理这些双语新闻，我们便能快速地获取这些海量新闻的主要内容，这对于我国与越南的经济交流、文化交流等有着重要意义。

按照生成摘要的方式，自动摘要技术可以分为抽取式（extractive）摘要和抽象式（abstractive）摘要。前者主要对原文档的句子进行重要性评估，再从中选取重要语句构成摘要；后者则是在理解原文档的基础上，重新组织语言生成摘要。由于越南语的自然语言生成技术还有一定的局限性，因此本文主要研究抽取式摘要的生成。抽取式摘要按照方法的不同可以分为基于特征统计的方法、基于机器学习技术的方法和基于图模型的方法。

（1）基于特征统计的方法使用词频、句子位置、

收稿日期：2018-09-29 **定稿日期**：2018-10-29

基金项目：国家自然科学基金（61472168，61761026，61732005，61672271，61762056）；云南省高新技术产业专项（201606）；云南省科技创新人才基金（2014HE001）；云南省自然科学基金（2018FB104）

是否包含关键词等这类特征对句子的重要程度进行衡量，然后通过一定的策略选取重要句子构成摘要。例如，Luhn 利用了最直观的思想，即词频越高的词汇越有可能描述文档的主要内容[1]，因此利用句中词汇的频率给句子打分，选择得分高的句子生成摘要。另外，也有方法根据原文档的特点，融入句子位置[2]、句子长度、句子与标题的相似度[3]等特征来更好地衡量句子的重要性。这类方法应用于写作规范、结构清晰的文档时能取得较好的结果。

（2）随着机器学习技术的发展，也逐渐出现了一些基于机器学习技术的自动摘要方法。例如，有研究者利用朴素贝叶斯分类模型[4]判断文档里的每个句子是否为摘要句，也有研究者利用决策树[5]、隐马尔科夫模型[6]等算法来生成摘要。这类方法适用于有足够多的训练语料的情况，而且在处理科技文献、新闻文档等结构化文档时能取得较好的结果。

（3）基于图模型的方法得到了广泛的应用，这类方法的一般思想是把文档分解为若干单元（词或句子），然后以这些单元为顶点、以单元间的关联为边建立图模型，通过图排序算法计算得到各个顶点的得分，再通过一定的策略选择得分高的顶点构成摘要。例如，文献[7]在处理文档时，以文档中的句子作为顶点、句子间的相似度作为边来构建句子图，之后在句子图上使用 TextRank 算法对句子进行排序，选择排序靠前的句子构成摘要。这类方法具有一定的扩展性，可以方便地融入一些特征。例如，文献[8]在为医学文献生成摘要时，用句子含有的医学本体（ontology）来表征每个顶点，通过融入领域知识来提升文档摘要的准确性。另外，句子间的余弦相似度、语义相似性等特征[9-10]也能用于衡量句子间的关联强度，以提升自动摘要的效果。

上述的自动摘要方法都是应用于单语环境，近年来，研究者们逐渐开始探索跨语言或多语言环境下的自动摘要方法。例如，文献[11]提出了一种跨语言自动摘要方法，旨在为阿拉伯文的新闻文档生成英文摘要。实验使用了相关的双语新闻文档集，首先通过机器翻译把阿拉伯文文档翻译为英文文档，然后从翻译后的文档中抽取摘要，之后计算这份摘要与英文文档集中句子的相似度，最后从英文文档集中挑选出相似度足够高的句子作为阿拉伯文文档集的摘要。文献[12]提出了一种多语言自动摘要方法，旨在为相关的中英文报道生成两份摘要，分别代表中文报道独有的观点和英文报道独有的观点。该方法也是使用机器翻译的方法，把中文文档翻译为英文、把英文文档翻译为中文，然后在两种单语环境下生成摘要。现有的跨语言、多语言环境下的自动摘要方法，都利用了机器翻译技术。在机器翻译效果较好时，能够取得较好的自动摘要结果。

我们的目标是为相关的汉越双语新闻生成一份双语摘要，处理的对象是汉越双语新闻文档。由于汉语和越南语之间的机器翻译效果还不理想，因此无法直接借鉴已有的方法。关于同一事件的新闻文本，不论这些文本是同种语言还是不同语言，其句子之间具有一定的关联关系，利用这些关联关系有助于生成自动摘要。因此，本文提出了多特征融合的双语新闻摘要方法，通过一定的方法定量分析新闻句子间的关联关系，并将这些关联关系融入图模型，提升自动摘要的效果。

1　新闻文本的特点分析

关于同一事件，往往会有很多新闻对其进行报道。由于新闻体裁要求用最准确、简洁的文字对事件进行描述，故不同的新闻文本在写作时往往具有一些相同的特点，下面以表 1 为例对新闻文本的写作特点进行说明。

表 1　两篇关于同一事件的新闻

标题：2017 年赴越中国游客突破 400 万人次 正文：越南统计总局 27 日称，2017 年截至目前，越南已接待中国游客超过 400 万人次，比去年同期增长 48.6%。今年越南已接待外国游客 1 290 万人次，同比增长 29.1%。其中韩国游客数量增长比例最大，达到 56.4%，其次是中国游客，再次是增长 32.3%的俄罗斯游客。
标题：赴越中国游客突破 400 万人次 正文：越南统计总局 12 月 27 日称，2017 年截至目前，越南已接待中国游客超过 400 万人次，比去年同期增长 48.6%。最近一年，越南方向出游人次同比增长 198%，成为东南亚地区出游人次增速最快的一匹“黑马”，芽庄、河内、下龙湾、岘港等目的地深受中国游客欢迎。

比较两篇新闻可以发现，关于同一新闻事件的不同新闻文本，往往会有如下一些写作特点：

（1）多篇新闻文本，虽然会从各个相同的或不同的角度对新闻事件进行描述，但在描述的过程中

会出现相同的新闻要素，如时间、地点、参与人、组织机构等；

（2）多篇新闻文本，会引用相似的、甚至是相同的句子对新闻事件进行描述；

（3）新闻文本会在标题、正文第一段、段落第一句等位置，简明扼要地对新闻事件进行描述或表达新闻媒体的观点。

通过以上分析我们认为，如果能在汉越双语新闻的自动摘要任务中利用这些新闻文本的写作特点，就能更好地生成双语新闻的摘要。

2 融合多特征的汉越双语新闻摘要方法

为了获取关于同一事件的汉越双语新闻的主要内容，我们利用新闻文本的写作特点，提出了一种融合多特征的汉语双语新闻摘要方法，整体框架如图1所示。

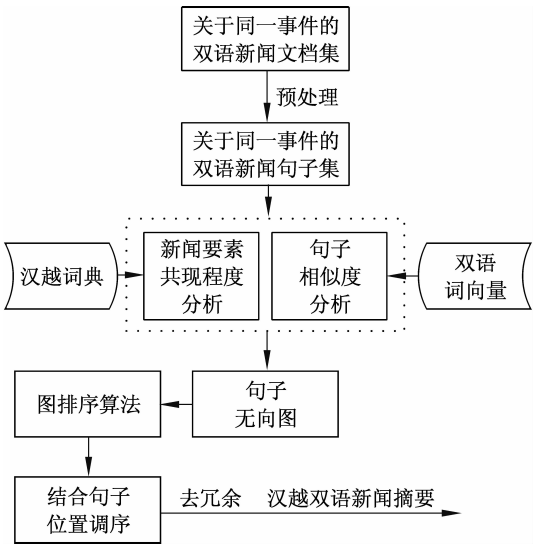


图1 融合多特征的汉越双语新闻摘要方法

本方法的流程是：①对双语新闻文档集进行预处理，建立以句子为顶点的无向图；②根据新闻文本的写作特点，用句子间的新闻要素共现程度以及句子间的相似度来衡量句子间关联关系的强弱，并以此作为顶点间边的权重；③在句子图上利用图排序算法计算句子的重要性并进行排序；④结合句子的位置特征对排序结果进行调序；⑤挑选出重要句子并去除冗余，生成汉越双语新闻文本的摘要。

2.1 句子间的新闻要素共现程度分析

新闻要素包含了事件发生的时间、地点、参与人和涉及到的组织机构等信息。为了用新闻要素共现程度来衡量句子间的关联关系强弱，我们借鉴文献[13]的方法对句子间的要素共现程度进行定量分析。首先，使用句子所包含的新闻要素表征句子；然后，使用句子间的新闻要素共现次数来衡量共现程度，具体步骤如下。

第一步，抽取句子中的命名实体作为新闻要素并表征句子。

抽取句子中的命名实体作为新闻要素，得到的汉语新闻要素集合记为 $E^{cn}=\{e_1^{cn},e_2^{cn},\cdots,e_n^{cn}\}$，越南语新闻要素集合记为 $E^{vi}=\{e_1^{vi},e_2^{vi},\cdots,e_m^{vi}\}$。得到两个新闻要素集合之后，任意一个汉语句子 s_i^{cn} 和越南语句子 s_j^{vi} 可以表示为：$s_i^{cn}=\{e_1^{cn},e_2^{cn},\cdots,e_{n_1}^{cn}\}$，$s_j^{vi}=\{e_1^{vi},e_2^{vi},\cdots,e_{n_2}^{vi}\}$，其中 $e_i^{cn}\in E^{cn}$，$e_j^{vi}\in E^{vi}$。

第二步，对齐汉语新闻要素和越南语新闻要素。

为了衡量汉语句子和越南语句子之间的新闻要素共现程度，需要对抽取得到的汉语要素集合 E^{cn} 和越南语要素集合 E^{vi} 进行人工对齐。借助汉越双语词典，人工对齐具有互译关系的汉语新闻要素和越南语新闻要素，得到对齐的汉越新闻要素集合 $E^{cv}=\{(e_1^{cn},e_i^{vi}),(e_2^{cn},e_2^{vi}),\cdots,(e_k^{cn},e_k^{vi})\}$。

第三步，计算句子间的新闻要素共现程度。

对任意句子 $s_i=\{e_1,e_2,\cdots,e_{n_1}\}$、$s_j=\{e_1,e_2,\cdots,e_{n_2}\}$，如果表征 s_i 和 s_j 的集合有交集，则 s_i 和 s_j 之间具有要素共现关系。其中，若 s_i 和 s_j 是同一语种的句子，则直接做交集运算即可判断，若 s_i 和 s_j 是不同语种的句子，则需使用对齐集合 E^{cv} 中的要素重新表征句子 s_i 和 s_j 之后，再做交集运算进行判断。

考虑到最终生成的交集有大有小，它可能只包含一个新闻要素，也可能包含多个新闻要素。句子间的新闻要素共现程度，在交集包含多个新闻要素时，理所应当地比交集只包含一个要素时强。此外，包含新闻要素较多的句子与其他句子具有要素共现关系的概率更大，而句子间的新闻要素共现程度不应该受句子本身所包含的新闻要素数量的影响。根据上述思想，使用式(1)计算任意两个句子间的新闻要素共现程度。

$$R_e(s_i,s_j)=\frac{2\times \mathrm{Count}(s_i\cap s_j)}{\mathrm{Count}(s_i)+\mathrm{Count}(s_j)} \quad (1)$$

其中，$\mathrm{Count}(s_i\cap s_j)$ 表示句子 s_i 和 s_j 的交集中新闻要素的数量，$\mathrm{Count}(s_i)$ 表示句子 s_i 所包含的新闻要素的数量。

2.2 句子间的相似度分析

根据新闻文本的写作特点，我们还使用句子间

的相似度来衡量句子间的关联关系强弱。句子相似度计算是自然语言处理领域中的一项重要任务，根据不同的句子相似度定义方法，可以分为语义(semantic)相似度和主题(topic)相似度。以“他喜欢吃苹果”和“他不喜欢吃苹果”两个短句为例进行说明，由于两个句子所表达的情感极性不同，所以两个句子的语义相似度较低，但是由于两个句子谈论的内容是相关的，所以两个句子的主题相似度较高。我们根据任务需要使用主题相似度，即只要两个新闻句子谈论的是相关的内容，就认为二者具有较高的相似度。为了计算不同语种句子间的主题相似度，我们使用文献[14]提出的方法训练汉越双语词向量，使用双语词向量表征句子，并计算相似度，具体做法如下。

第一步：训练汉越双语词向量。

利用维基百科语料训练中文词向量 $\boldsymbol{\Sigma}$ 和越南语词向量 $\boldsymbol{\Omega}$，然后使用文献[14]提出的方法把两份单语词向量投影到同一向量空间，得到汉越双语词向量。投影后的中文词向量记为 $\boldsymbol{\Sigma}^*$，投影后的越南语词向量记为 $\boldsymbol{\Omega}^*$。

第二步：利用词向量表征句子，得到句子的向量表示。

由于计算的是句子之间的主题相似度，所以在表征句子时，需要剔除那些与新闻事件无关的、不重要的词，比如介词、连词和冠词等，具体步骤如下。

首先，对句子进行分词并标注词性，选择动词、名词、形容词和副词来表征句子。

然后，用挑选出的词表征句子，例如，汉语句子 s_i^{cn} 和越南语句子 s_j^{vi} 可以分别表示为：$s_i^{\mathrm{cn}} = \{w_1^{\mathrm{cn}}, w_2^{\mathrm{cn}}, \cdots, w_k^{\mathrm{cn}}\}$，$s_j^{\mathrm{vi}} = \{w_1^{\mathrm{vi}}, w_2^{\mathrm{vi}}, \cdots, w_l^{\mathrm{vi}}\}$。

之后，利用训练好的词向量生成句子的向量表示。Kusner 等提出了利用词向量计算文档相似度的方法[15]，他们在论文中指出，简单地利用句子中词的词向量求平均即可获得句子的向量表示，且这种方法在计算句子相似度时具有一定的效果。因此，我们可以做如下表示：$s_i^{\mathrm{cn}} = \frac{1}{k}\sum_{m=1}^{k} v_m^{\mathrm{cn}}$，$s_j^{\mathrm{vi}} = \frac{1}{l}\sum_{n=1}^{l} v_n^{\mathrm{vi}}$。其中，$v_m^{\mathrm{cn}} \in \boldsymbol{\Sigma}^*$，$v_n^{\mathrm{vi}} \in \boldsymbol{\Omega}^*$。

第三步：利用句子的向量表示计算句子之间的相似度。

汉语句子 s_i^{cn} 和越南语句子 s_j^{vi} 的相似度可以用对应向量的余弦值计算，如式(2)所示。

$$R_{\mathrm{sim}}(s_i^{\mathrm{cn}}, s_j^{\mathrm{vi}}) = \frac{1}{2} \times \left(1 + \frac{s_i^{\mathrm{cn}} \cdot s_j^{\mathrm{vi}}}{\| s_i^{\mathrm{cn}} \| \times \| s_j^{\mathrm{vi}} \|}\right) \quad (2)$$

上述步骤以双语句子间的相似度计算为例进行说明，单语句子间的相似度计算过程与之相似。

2.3　句子关联无向图的建立

在得到句子间的新闻要素共现程度和句子间的相似度之后，就可以建立以句子为顶点、以句子间的关联关系为边的无向图，建立好的句子无向图如图 2 所示。

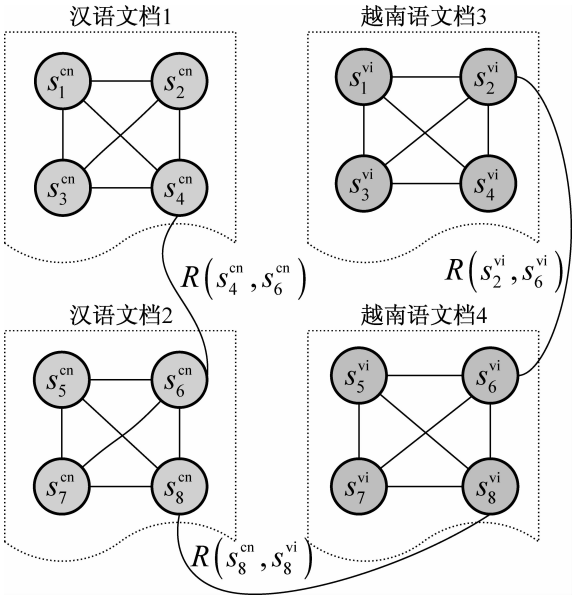


图 2　双语新闻句子无向图示例

图中的文档是关于同一事件的汉语新闻文档和越南语新闻文档。对任意两个句子 s_1 和 s_2，我们在 2.1 节中对句子间的新闻要素共现程度 $R_{\mathrm{e}}(s_1, s_2)$ 做了定量分析，在 2.2 节中对句子间的相似度 $R_{\mathrm{sim}}(s_1, s_2)$ 做了定量分析，结合二者如式(3)所示。

$$R(s_1, s_2) = \alpha R_{\mathrm{e}}(s_1, s_2) + \beta R_{\mathrm{sim}}(s_1, s_2) \quad (3)$$

其中，$R(s_1, s_2)$ 表示句子 s_1 和 s_2 的关联强度，式中 α 和 β 是权重参数，两个参数满足 $0 < \alpha, \beta < 1$ 且 $\alpha + \beta = 1$。对于句子无向图中的所有顶点，两两之间利用式(3)计算关联强度，则可以得到句子无向图的关联强度矩阵，如式(4)所示。

$$\boldsymbol{M} = \begin{bmatrix} R_{1,1} & R_{1,2} & \cdots & R_{1,n} \\ R_{2,1} & R_{2,2} & \cdots & R_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ R_{n,1} & R_{n,2} & \cdots & R_{n,n} \end{bmatrix} \quad (4)$$

其中，$R_{i,j}$ 就是句子 s_i 和 s_j 的关联强度，即 $R_{i,j} = R(s_i, s_j)$。为了简化后续的计算，如果两个句子之间的关联强度 $R_{i,j} < 0.2$，则在建立句子关联无

向图时不连接这两个顶点且把关联强度 $R_{i,j}$ 置为 0。

2.4 句子关联无向图顶点权重的计算

在建立好句子无向图并计算得到相应的关联强度矩阵 $\boldsymbol{M}$ 后，利用 TextRank 算法在句子无向图上进行排序计算，得到各个顶点的权重得分。TextRank 算法把 PageRank 算法的思想扩展到了带权重的无向图模型上，其计算如式(5)所示。

$$\mathrm{Score}(s_i)=(1-d)+d\times\sum_{j=1,j\neq i}^{n}\frac{R_{i,j}}{\sum_{k=1,k\neq j}^{n}R_{j,k}}\mathrm{Score}(s_j) \quad (5)$$

其中，$\mathrm{Score}(s_i)$ 表示句子 s_i 的权重得分，d 表示阻尼系数，一般设置为 0.85，$R_{i,j}$ 是关联强度矩阵 $\boldsymbol{M}$ 中的值。算法迭代多次并收敛之后，就能得到每个句子的重要程度。

2.5 句子的位置重要性分析

在上述计算句子重要性的过程中，只考虑了新闻句子间的要素共现程度及相似度，二者分析的是句子与句子间的关系对句子重要性的影响，没有考虑到句子在文本中的位置也反映了句子的重要性。已有研究表明：在一定类型的文档中，句子重要性与句子位置具有一定的关系。例如有研究者指出，标题后的句子更有可能表达文档的中心思想，且重要句子更可能出现在文档的首段或尾段，以及段落的首句或尾句[16]，且这类基于位置评价句子重要性的方法，对新闻文本、科技文献等写作规范的文本效果相对较好。根据新闻文本的写作特点并结合已有的研究结果，我们提出以下调序公式，如式(6)所示。

$$S_{\mathrm{reo}}(s_i)=\begin{cases}1.5\times\mathrm{Score}(s_i), & s_i\text{ 为标题句}\\ 1.4\times\mathrm{Score}(s_i), & s_i\text{ 为第一段的句子}\\ 1.2\times\mathrm{Score}(s_i), & s_i\text{ 为非第一段的第一句}\\ 1.0\times\mathrm{Score}(s_i), & s_i\text{ 为其他句子}\end{cases} \quad (6)$$

其中，$\mathrm{Score}(s_i)$ 是排序算法得到的句子 s_i 的得分，即式(5)的最终结果，$S_{\mathrm{reo}}(s_i)$ 是调序后的句子 s_i 的得分。

2.6 去除冗余句子生成摘要

上述的排序和调序过程，为新闻文档集中的每个句子都分配了重要性得分，得分越高的句子越好地描述了文档集的主要内容。但是，由于新闻文档集中存在很多相似、甚至是重复的句子，因此不能直接按照得分高低抽取句子构成摘要。需要去除冗余句子提高摘要的可读性，具体做法如下。

第一步：设调序后的句子集合为 C，集合中的句子按照得分从高到低排序，序号为 1 至 $|C|$。

第二步：选择集合 C 中的第一个句子 s_1，对于 $i=2$ 至 $i=|C|$，利用公式(3)计算 $R(s_1,s_i)$，如果 $R(s_1,s_i)$ 的值大于阈值 θ，则从集合 C 中删除句子 s_i。

第三步：把句子 s_1 加入摘要并从集合 C 中删除。对集合 C 中的句子重新排序，序号为 1 至 $|C|$。

第四步：重复第二步和第三步，直到获得满足要求的摘要。

3 实验及分析

3.1 实验语料

实验语料包括两部分，一部分用于训练双语词向量，另一部分用于验证我们提出的汉越双语新闻摘要方法，具体信息分别叙述如下。

3.1.1 维基百科语料

考虑到训练单语词向量的目的，是为了把两份向量进行投影，投影后的两份词向量构成一个第三方向量空间，使得语义相近的词汇(不论是中文词汇或是越南语词汇)在空间中的位置也尽量相近。最终我们使用维基百科作为词向量的训练语料，它有两个优点：一是维基百科方便获取且规模较大；二是从双语语料的内容一致性来说，汉越维基百科所讨论的内容是天然相关的，即几乎每个越南语维基百科页面，都有相应的汉语维基百科页面。语料的内容越是一致，则语义相近的词越多，越有利于单语词向量的投影。

下载得到的维基百科语料包含一些待编辑词条的页面，这些页面词数很少，几乎不含有有用的语义信息，无法用于训练，需要对其删减，具体信息如表 2 所示。

表 2 维基百科语料的具体信息

	删减前		删减后	
	文档数	字数	文档数	字数
越南语	1 656 469	152 149 241	287 534	124 197 165
汉语	2 974 751	198 438 813	292 315	170 438 813

3.1.2　双语新闻语料

目前还没有公开的汉越双语新闻语料，因此我们从中国新闻网、新华网、新浪新闻等国内新闻网站，以及越南每日快讯、越南通讯社网、中华网越南版等越南新闻网站收集新闻，每个新闻保留其标题、正文、发布时间等。人工整理收集来的新闻文本，挑选出三个在汉越双方都有较多报道的新闻事件，同时根据关键词从谷歌检索、补充一定量的相关新闻报道构成汉越双语新闻语料。针对每个事件，从相关的新闻文本中人工抽取 6 个句子（汉越句子各 3 个）作为参考摘要。把双语新闻文档和人工抽取的摘要作为实验数据，具体信息如表 3 所示。

表 3　汉越双语新闻数据的具体信息

新闻事件	语言	文档数量	句子数量
阮富仲访华	汉语	25	441
	越南语	25	424
湄公河放水	汉语	25	435
	越南语	25	419
中越防长会晤	汉语	25	408
	越南语	25	379

3.2　评价指标

采用自动摘要任务中常用的 ROUGE 值作为评价指标[17]，ROUGE 是一种基于召回率的相似性度量方法，它通过比较候选摘要与参考摘要中共现的 n 元组 n-gram 来评价候选摘要的质量。ROUGE 值越高说明候选摘要的质量越好，计算方法如式(7)所示。

$$\text{ROUGE-}n=\frac{\sum_{s\in R}\sum_{n\text{-gram}\in s}\text{Count}_{\text{match}}(n\text{-gram})}{\sum_{s\in R}\sum_{n\text{-gram}\in s}\text{Count}(n\text{-gram})} \tag{7}$$

其中，n 表示 n 元组的长度，R 表示构成参考摘要的句子的集合，s 表示参考摘要中的句子，Count(n-gram)表示句子 s 中 n 元组的数目，$\text{Count}_{\text{match}}$($n$-gram)表示候选摘要句与参考摘要句 s 共同包含的 n 元组的数目。通过式(7)可以发现 ROUGE-n 反映的是参考摘要句的 n 元组的召回率。实验中我们使用 ROUGE-1 和 ROUGE-2 来评价摘要结果的好坏。在计算 ROUGE 值时，汉语摘要句和越南语摘要句分开计算，然后再取平均值。

3.3　实验设计与结果分析

本文包含三个实验，实验 1 训练双语词向量并验证其有效性；实验 2 通过对比选择最佳的 α、β 参数以及阈值 θ；实验 3 通过比较验证所提方法的有效性。

3.3.1　训练双语词向量并验证其有效性

首先使用 Word2Vec 工具训练单语词向量，训练之前需要对维基语料进行预处理。中文语料的预处理包括分词、去除标点与特殊符号，以及繁简体转换。越南语语料的预处理包括分词、去除标点及特殊符号。越南语语料的处理使用 Vitk 工具包①。

经过多次训练比较，本文的训练参数设置如下：上下文窗口长度为 10，词向量维度设为 100，低频词阈值设为 10，采用 skip-gram 模型进行训练，迭代次数为 50 次，其余参数使用默认值。训练完毕后使用文献[14]提供的代码②训练双语词向量。从训练完的双语词向量中随机选择了几个名词、动词及形容词，并计算它们在另一种语言中前 5 个相近的词，结果如表 4 所示。

表 4　双语词向量效果示例

输入	中国	越南
近似词	Trung Quốc(中国)	Việt Nam(越南)
	Trung Hoa(中国)	VN(越南)
	Đài Loan(台湾)	Hồ Chí Minh(胡志明)
	Đại lục(大陆)	Hà Nội(河内)
	Triều Tiên(朝鲜)	Thùa Thiên Huế(承天顺化省)
输入	thăm(访问)	quan trọng(重要)
近似词	访问	代表性
	出访	重要性
	会面	意义
	到访	主要
	谈话	极其重要

从结果可以看出，对于名词、动词这类具有明确语义信息的词来说，训练得到的双语词向量能取得较好的结果，形容词的效果相对较差，但也能匹配到较为相关的词。因此我们认为，双语词向量可以用于句子间的相似度计算。

3.3.2　通过对比选择最佳参数

本文提出的摘要方法含有参数 α、β 和 θ。θ 用于

① Vitk 工具包：https://github.com/phuonglh/vn.vitk

② 代码地址：https://github.com/mfaruqui/crosslingual-cca

去冗余过程中过滤关联强度过高的相似句子，α和β用于确定句子间的要素共现程度和相似度对句子关联强度的贡献比例。

先用以下方法确定θ的取值。

首先，在$\alpha=1$、$\beta=0$和$\alpha=0$、$\beta=1$两种情况下生成摘要；然后，将θ从1逐步减少到0（每次减少0.1），在这个过程中人工统计不同θ值下摘要中高度相似的句子的数量；最后，选择最大的θ值使得生成的摘要中几乎没有高度相似的句子。$\alpha=1$、$\beta=0$时$\theta=0.6$，$\alpha=0$、$\beta=1$时$\theta=0.7$。最终选择$\theta=0.65$作为去冗余时的阈值。

在确定$\theta=0.65$后，通过对比生成摘要的ROUGE值选择最佳的α和β参数。具体结果如图3所示。从图中可以看到在$\alpha=0.4$、$\beta=0.6$时，提出的方法取得了最好的结果。

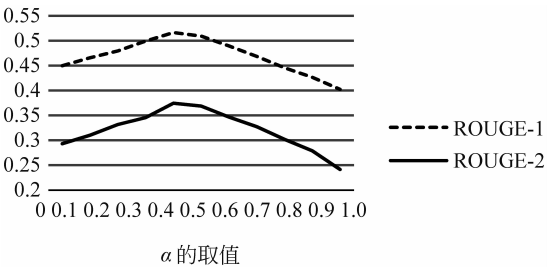


图3 α和β参数对ROUGE值的影响

3.3.3 验证汉越双语新闻摘要方法的有效性

为了验证所提方法的有效性，将如下几个方法进行对比。

① Multi-Feature：本文所提出的方法。参数选择为$\alpha=0.4$、$\beta=0.6$、$\theta=0.65$。

② Multi-Feature-e：本文所提出的方法。参数选择为$\alpha=1$、$\beta=0$、$\theta=0.65$。

③ Multi-Feature-sim：本文所提出的方法。参数选择为$\alpha=0$、$\beta=1$、$\theta=0.65$。

④ TextRank：该方法在计算句子间边的权重时，原本使用的是句子间词的重叠数量，仅适用于单语文档。为了使该方法能够适用于双语新闻摘要任务，使用句子间的新闻要素共现程度来代替句子间词的重叠数，用于计算句子间边的权重。此外，该方法本身用于单文档摘要，没有考虑去冗余，故此处我们再为其加上去冗余的步骤。该方法与Multi-Feature-e相比，没有利用句子的位置对句子重要性进行调序。实验结果如表5所示。

比较Multi-Feature-e和TextRank可以看出，对排序后的结果再进行调序，取得了较大提高。

表5 在三个新闻事件上的实验结果

	ROUGE-1	ROUGE-2
Multi-Feature	0.516 3	0.374 5
Multi-Feature-e	0.402 2	0.241 5
Multi-Feature-sim	0.449 6	0.293 1
TextRank	0.318 4	0.154 7

我们认为这验证了把句子位置作为特征的有效性。另外，之所以取得较多提高的原因，我们认为是新闻摘要本身的特点造成的结果，因为新闻报道的标题本身就对新闻文本做了简练的概括，调序时针对这一特点对新闻标题赋予了较高的权重。

比较Multi-Feature-e和Multi-Feature-sim可以看出，引入词向量计算句子之间的关联强度，相比仅使用词共现来计算句子之间的关联强度更为有效，我们认为这是因为词向量不仅能计算共现词之间的相似度，还能计算那些相关词之间的相似度；另外，在衡量双语句子间的关系时，词向量的效果比双语词典要好。

比较Multi-Feature与其他方法可以发现，在汉越双语新闻摘要任务上，本文所提出的方法取得了较好的结果，具有有效性。

4 总结

为了生成汉越双语新闻的摘要，本文提出了一种融合多特征的汉越双语新闻摘要方法。该方法根据新闻文本的写作特点，分析了句子间的新闻要素共现程度、句子间的相似度以及句子的位置重要性，并把这三个特征融合到模型中。实验证明，所提出的方法在汉越双语新闻摘要任务上取得了较好的结果。由于在分析句子间的相似度时，仅利用特定类型词的词向量加权来衡量句子相似度，有一定的局限性。在下一步工作中，考虑使用新的方法来衡量句子间的相似度，以提升自动摘要效果。

参考文献

[1] Luhn H P. The automatic creation of literature abstracts[J]. IBM Journal of Research and Development, 1958, 2(2): 159-165.

[2] Baxendale P B. Machine-made index for technical literature: An experiment[J]. IBM Journal of Research and Development, 1958, 2(4): 354-361.

[3] Edmundson H P. New methods in automatic extracting[J]. Journal of the ACM, 1969, 16(2):264-285.

[4] Kupiec J, Pedersen J, Chen F. A trainable document summarizer[C]//Proceedings of the 18th annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1995:68-73.

[5] Lin C Y. Training a selection function for extraction [C]//Proceedings of the . Eighth International Conference on Information and Knowledge Management. ACM, 1999:55-62.

[6] Conroy J M, O'Leary D P. Text summarization via hidden Markov models[C]//Proceedings of the 24th amnual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2001:406-407.

[7] Mihalcea R. Graph-based ranking algorithms for sentence extraction, applied to text summarization[C]//Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, 2004:20.

[8] Morales L P. Concept-graph based biomedical automatic summarization using ontologies[C]//Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing. Association for Computational Linguistics, 2008:53-56.

[9] Ferreira R, Freitas F, Cabral L D S. A four dimension graph model for automatic text summarization[C]//Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence. IEEE, 2013: 389-396.

[10] Ferreira R, Lins R D, Freitas F. A new sentence similarity method based on a three-layer sentence representation[C]//Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence. IEEE Computer Society, 2014:110-117.

[11] Evans D K, McKeown K, Klavans J L. Similarity-based multilingual multi-document summarization [R]. Technical Report CUCS-014-05, cOUNBIA uNIVERSITY.

[12] Wan X, et al. Summarizing the differences in multilingual news[C]//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011: 735-744.

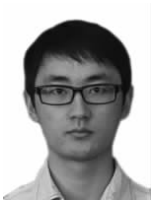
[13] 刘书龙. 汉越双语新闻观点句抽取及分析方法研究[D]. 昆明：昆明理工大学硕士学位论文，2017.

[14] Faruqui M, Dyer C. Improving vector space word representations using multilingual correlation[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014: 462-471.

[15] Kusner M J, et al. From word embeddings to document distances[C]//Proceedings of the 32rd International Conference on Machine Learning, 2015: 957-966.

[16] Lin C Y, Hovy E. Identifying topics by position [C]//Proceedings of the 5th conference on Applied Natural Language Processing. Association for Computational Linguistics, 1997: 283-290.

[17] Lin C Y, Hovy E. Automatic evaluation of summaries using N-gram co-occurrence statistics[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics, 2003:71-78.



叶雷(1992—)，硕士研究生，主要研究领域为信息检索、自然语言处理。
Email:yelei@sina.com

高盛祥(1977—)，通信作者，博士，主要研究领域为信息检索、机器翻译。
E-mail:gaoshengxiang.yn@foxmail.com

余正涛(1970—)，博士，博士生导师，主要研究领域为自然语言处理、信息检索、机器翻译。
E-mail:ztyu@hotmail.com