基于领域本体的文本分割方法研究

刘 耀[1] 帅远华[2] 龚幸伟[1] 黄 毅[1]
(中国科学技术信息研究所 北京 100038)[1] (北京大学 北京 100080)[2]

摘 要 文本分割在信息检索、摘要生成、问答系统、信息抽取等领域发挥着重要作用。在总结现有的国内外文本分割方法的基础上,提出了一种基于领域本体对文本进行线性分割的方法。该方法利用初始概念自动获取结构化语义概念集合,并根据获取的概念、属性及属性词在文本中出现的频次、位置和关系等因素为段落赋予语义标签,挖掘文本的子主题信息,将拥有相同语义标注信息的段落划分为相同语义段落,实现了文本不同子主题之间的分割。实验结果表明,该方法对于特定领域的文本分割的准确率、召回率以及F值分别达到了85%,90%和88%,分割效果能够满足实际应用需求,并优于现有的无需训练语料的文本分割方法。

关键词 文本分割,领域本体,语义标注,语义段落
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.01.021

Study on Text Segmentation Based on Domain Ontology

LIU Yao[1] SHUAI Yuan-hua[2] GONG Xing-wei[1] HUANG Yi[1]
(Institute of Scientific and Technical Information of China, Beijing 100038, China)[1]
(Peking University, Beijing 100080, China)[2]

Abstract Text segmentation plays an important role in information retrieval, abstract generation, question-answering system, information extraction and so on. This paper put forward a new text segmentation method based on domain ontology after analyzing and summarizing existing methods at home and abroad. The method first uses initial concept to automatically obtain structured semantic concepts set, which are then used to affix semantic labels to paragraphs in text based on the frequency of occurrence, position and relationship of concepts and properties. Paragraphs with the same semantic annotation information are grouped into one semantic paragraph, which helps discover the sub-topics information and meanwhile realize topic segmentation for texts. The experimental result shows that the precision, recall and F-measure of this method can achieve 85%, 90% and 88% respectively, which performs better than most existing methods and satisfies the real application needs.

Keywords Text segmentation, Domain ontology, Semantic annotation, Semantic paragraph

1 引言

网络的快速发展使信息资源以爆炸性速度增长,海量的文本资源为文本处理和分析带来了巨大挑战,文本分割则是解决该问题的重要步骤之一。文本分割是按照主题对文本进行划分,寻找不同主题之间的边界,将语义上相关的自然段或句子划分到同一个主题段落,其分割对象是静态文本、语音流或者网络动态数据。

文本分割被广泛运用于自然语言处理的多个任务中,如信息检索、摘要生成、问答系统、信息抽取等。在信息检索领域,文本分割技术可以帮助用户缩小检索范围,快速找到搜索结果,提高搜索准确性。对于摘要生成,文本自然段的分析往往不能准确表达整个文本主题思想,而文本分割则可以将文本划分为多个子主题,选取每个子主题下的代表句子形成整个文本的主题,基于子主题的文本摘要能更为全面地覆盖整个文本的信息。文本分割也可以运用于问答系统,问答系统对文本结构分析有更高的要求,对文本主题的划分将大大减少无关信息的干扰,提高问答系统的性能。

现有的大多数文本分割方法运用词汇的重复信息,但很少考虑文本的语义以及概念之间的关系。领域本体作为描述相关领域知识的概念体系,被广泛运用于知识工程、图书情报以及人工智能等领域。对于特定领域的文本,领域本体的概念结构能够较为完整地涵盖文本的父主题与子主题之间的关系。因此,本文提出了一种基于领域本体的文本分割方法,其主要思路是利用文本标题或段落中的主题概念自动获取本体的概念关系以及属性特征对文本进行语义标注,将拥有不同

到稿日期:2017-05-08 返修日期:2017-09-18

刘 耀(1972—),男,博士,研究员,CCF高级会员,主要研究方向为自然语言处理、知识组织与知识工程,E-mail:liuy@istic.ac.cn(通信作者);**帅远华**(1993—),男,硕士,主要研究方向为自然语言处理、视频摘要;**龚幸伟**(1986—),男,硕士,主要研究方向为自然语言处理、语义爬虫;**黄 毅**(1986—),男,硕士,主要研究方向为本体与知识工程。

语义标签的段落划分为不同语义段落，从而实现文本主题的划分。其优点在于能自动获取概念语义结构，并在分割文本的同时挖掘文本的子主题信息以及语义关系。

2　研究进展

文本分割的思想最早来自对文本语篇结构的建模研究，它是文本结构分析和构造的首要步骤。本文将目前国内外主要的文本分割技术分为 3 类：基于词汇聚集、基于语言特征和基于概率统计思想。文本分割技术主要涉及两个问题：语义段落边界的识别以及主题数目的确定[1]。要找到最合理的语义段落的边界，就要实现分割单元内部具有最大的语义相关性而分割单元之间具有最小的语义相关性。

2.1　基于词汇聚集的文本分割方法

词汇聚集方法的主要思想来源于 Halliday 和 Hasan 的研究[2]，该研究提到词或短语的重复出现反映了文本片段的词汇聚集程度，同一主题内的词汇聚集度明显大于主题边界处的词汇聚集度。因此，基于词汇聚集的文本分割假定相同、相似或语义相关的词汇倾向于出现在同一语义段落内，通过计算相同或相似词汇的聚集程度来判断语义段落的边界，从而划分文本主题。

20 世纪 90 年代，国外学者开始研究文本分割技术，其中大部分是基于词汇聚集的思想。TextTiling 方法[3]利用窗口滑动的办法对文本进行分割，通过词频空间向量计算每相邻两个窗口的相似度值，对相似度值进行平滑处理后计算波谷的深度值，深度值超过阈值的为分割点。TextTiling 计算复杂度小，但是没有考虑文本之间的语义关联，分割正确率较低。词汇链方法[4]采用 Roget 综合词典来识别词汇的同义或聚集关系，将句子内重复出现的词构成词汇链，并根据大量的词汇链信息确定文档的分割位置。词汇链的文本分割算法除了考虑词汇的重复，还考虑了词汇的多种语义关联，如同义词、具体和一般的关系，以及整体和局部的关系。Dotplotting 方法[5]利用点阵图来反映文本不同部分词汇的重复出现情况，通过在词重复的位置绘制点来生成点图，点图中的区域密度反映了整篇文档中词汇的分布情况，点图中密度大的区域即为一个语义段落，对角线外部区域密度最小的点则为语义边界。Dotplotting 算法的不足之处在于只能支持给定分割单元数目条件下的文本分割。

此外，Roman Kern 和 Michael Granitzer[6] 提出了一种 TSF 的文本线性分割方法，该方法通过文本块外部相似度（即两个文本块之间的相似度）以及文本块内部句子相似度来寻找主题边界，分割错误率 P_k 值低于 TextTiling 和 C99 方法的分割错误率。文献[7-9]则基于层次聚类思想对文本进行分割，该类方法通过计算句子间的相似度值，利用聚类算法将句子逐层合并，最后寻找最优分割点。但是该方法通常需要事先确定主题数目，才能在最后的层次聚类树中决定分割结果。

2.2　基于语言特征的文本分割方法

基于语言特征的分割则是基于特定的语言现象，比如提示短语、韵律特征、停顿标记、指代、句法和词汇的形态变化，通过研究它们与主题片段首尾的关系确定语义段落边界。国外学者基于语言特征对文本进行分割，研究人员利用语言学的特征信息进行文本分割，其中包括新词出现、线索词、命名实体、代词的使用、同义词重复等。这种方法一般适用于特定文本类或者语音流的处理。

Reynar[10] 将文本分割视为一种标注任务，利用最大熵分类模型来计算这种潜在分割点被标注为语义段落边界的概率，同时考虑了其他的文本特征，例如领域线索词、线索短语、词频、命名实体的重复、代词的使用等多种线索。Kan[11] 提出了一种发现语篇结构的新方法，其分割片段基本取决于从文本中抽取的有用主题信息，这些信息包括专有名词、普通名词、人称代词以及所有格代词；然后利用这些词的出现频次及关联为自然段分配权重，权重若为正值则意味着一个语义段落的开始。Kauchak[12] 则提出一种基于特征的记叙文分割方法，将文本分割视为一种分类问题。该方法综合考虑记叙文的多方面特征（例如 WordNet 拥有同一父类的词、命名实体经常出现在分割边界），以及全名、命名实体链、代词、数字、对话等特征，然后通过统计分类器计算候选分割点属于语义边界的概率。其分割正确率比基于 TextTiling 和 PLSA 模型的方法高 24%以上。

2.3　基于概率统计的文本分割方法

基于词汇聚集和语言特征的文本分割算法通常利用了概率和统计的思想，但是基于概率统计方法的文本分割主要是指通过建立概率统计模型对文本进行主题划分，包括潜在语义分析（Latent Semantics Analysis，LSA）、概率潜在语义分析（Probabilistic Latent Semantics Analysis，PLSA）、潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）、隐马尔科夫模型（Hidden Markov Model，HMM）等方法。

文本中存在同义词和一词多义的现象。LSA 利用奇异值分解（SVD）将高维度的词汇与文档共现矩阵映射到低维度的潜在语义空间，从而使看似不相关的词与词之间建立深层次的语义关联。Choi[13] 于 2001 年利用 LSA 代替 C99 中的余弦向量函数来计算句子间的相似度，通过聚类寻找主题边界，取得了较好的分割结果。PLSA 主题模型的基础思想和 LSA 相同，但是使用了概率模型，比 LSA 有更坚实的数学基础及便于利用的数据生成模型。Brants 等人[14] 基于 PLSA 识别文本主题边界，并比较相邻文本块的距离以及通过相邻块之间的相似度值选择分割点，实验结果表明其分割错误率 P_k 低于 LSA 分割方法的分割错误率。

LDA 是一种非监督机器学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息。文献[15-18]均将 LDA 主题模型运用到文本分割中，其分割效果优于其他分割方法。LDA 主题模型认为每个文本都由特定的主题分布构成，每个主题又由潜在的词汇分布组成。文献[16]使用改进的动态规划（Dynamic Programming）算法改进了基于 LDA 的文本分割的计算量。文献[18]运用 MCMC 技术决定合适的主题边界，MCMC 的优点是既可以识别弱主题边界也可以识别强主题边界。

此外，Eisenstein 等人[19] 以及 Lan Du 等人[20] 利用贝叶斯方法进行非监督的主题分割，其主要思想是通过贝叶斯方法计算词汇的紧凑度，最后通过最大概率的观察序列产生文

本分割片段。文献[20]的实验结果表明,该方法在多个评测指标上优于现有其他方法。

3 基于本体结构的文本分割方法

3.1 文本分割思路

对于不同的应用场景以及应用目的,文本分割的方法、分割颗粒度以及分割技术会有所不同。传统的 TextTiling 方法的计算复杂度小,但是分割准确性较低;文献[13,21]等文本分割方法则需要预设主题数量或者待比较文本块的大小;其次,一些文本分割方法在计算文本相似度时只考虑了词汇的重复信息,而没有考虑词汇之间的语义关联;概率统计方法的分割效果较好,但是计算复杂度高,需要大量语料进行训练。此外,现有的大多数文本分割方法是针对通用领域的,并且分割颗粒度通常为句子级别。然而在实际应用中,文本分割颗粒度不需要细分到句子级别,一个自然段落通常只描述一个主题。

对于特定领域的文本,领域本体的概念结构能够较为完整地涵盖文本的父主题与子主题。因此,本文针对特殊领域提出了一种基于段落颗粒度的文本分割方法,其主要思想是利用本体的概念、属性以及概念间的关系对文本段落进行语义标注,将拥有相同语义标签的相邻段落划分为同一语义段落。该方法的优点在于先对文本进行语义段落划分,同时可以获得语义段落的主题。文本分割流程如下:

(1)输入领域文本。

(2)确定文本标题,无标题情况执行步骤(3),有标题情况执行步骤(4)。

(3)文本无标题的情况:1)文本有来源路径,根据来源找回文本标题/主题;2)文本无来源路径,根据式(1)取得分最高的前 n 个概念词,并自动获取概念语义结构,然后对段落进行语义标注。

(4)领域文本预处理:将文本按自然段划分,并对文本标题和正文进行分词、词性标注、去停用词处理,保留标题中的名词。

(5)根据文本标题词或段落主题词获取本体三层概念结构。

(6)利用已构建的概念结构对文本段落进行语义标注,语义标注的具体过程见 3.2.2 节。

(7)输出段落及段落语义标注信息。

(8)对于相邻的段落,若具有相同语义标注信息并且语义标注得分超过阈值 Y,则将其划分为同一语义段落。

由于概念层次的自动获取需要输入初始概念,因此对于无标题的文本段落,本文借鉴了文献[22]中高效的短文本主题词抽取方法,其计算公式为:

$$W(w_i)=tf(w_i)\times df(w_i)\times(1+g(w_i)) \tag{1}$$

$$tf(w_i)=\frac{f_j(w_i)}{n(d_j)} \tag{2}$$

$$df(w_i)=-\frac{n(w_i)}{N}\times\log\frac{n(w_i)}{N}-(1-\frac{n(w_i)}{N}\log(1-\frac{n(w_i)}{N}) \tag{3}$$

$$g(w_i)=\frac{n(w_i)}{n(w_i)+1} \tag{4}$$

其中,$tf(w_i)$是文档 d_j 中的词 w_i 的相对词频,由式(2)计算求得;$f_j(w_i)$是 w_i 在文档 d_i 中出现的次数;$n(d_j)$是文档中实词的个数;$df(w_i)$是词汇 w_i 的权重因子,由式(3)求得;$n(w_i)$是出现词汇 w_i 的段落个数;N 是文本段落总数;$g(w_i)$代表词汇 w_i 的主题表现力,由式(4)计算求得。对于没有标题的文本段落,采用式(1)计算词汇对主题的贡献得分,取得分最高的前 n 个主题词作为初始概念获取其概念层次结构。

3.2 具体步骤与方法

3.2.1 概念层次的获取

本体(Ontology)的概念最初来源于哲学领域,用于对世界任何领域内的真实存在做出系统的解释或说明。其后,本体的概念被引入到计算机领域,近年来被广泛运用于计算机的诸多领域,如知识工程、情报工作、信息系统以及人工智能等。

本体是相关领域内不同主体之间相互交流的一种语义基础,是对某领域知识的通用理解,其描述了领域内共同认可的概念以及概念之间的关系。目前,本体构建的思路主要分为本体论工程方法以及叙词表转换为本体方法两大类。本文采用了文献[23-24]中受限文本的本体自学习机制,从叙词表、百科类网站、图书目录以及搜索引擎获取本体的三层概念结构(见图 1)实现了本体概念体系的自动构建。

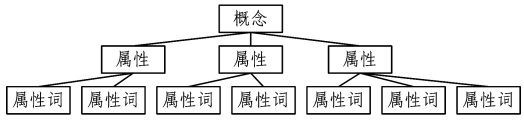

图 1 概念语义结构

Fig. 1 Semantic structure of concept

本体概念层次的自动构建流程如下:

(1)标题中的关键词(名词)或者段落获取的前 n 个概念词作为本体概念结构的第一层。

(2)本体概念结构第二层的获取方式如下:

1)从主题词表获取第一层概念的下位概念;

2)通过百度百科和维基百科检索第一层概念词,获取搜索结果的一级目录;

3)在当当网、亚马逊平台检索概念词,获得图书目录,过滤掉“目录”“前言”“概要”“简介”“第 1 章”“第三节”等词。

(3)对通过以上过程获取的字符串进行分词、词频统计、词性标注,保留名词,得到名词-频次集合。

(4)对上一步骤得到的各类资源的名词进行合并、去重,将结果作为备选关联词集合 $W=\{w_1,w_2,\cdots\}$。

(5)将关联词集合 W 与第(3)步得到的名词-频次集合输入关联词得分计算公式,通过计算得到 W 中每个词的关联词得分 $score(w_i)$。

$$score(w_i)=\mu\sum_{j=1}^{m}(\sum_{k=1}^{n}weight_j\times(\frac{w_i.length}{term_{kj}.length})\times tf_j(term_{kj})) \tag{5}$$

其中,$score(w_i)$为备选关联词集合 W 中词 w_i 的关联词得分,用以提取关联词的资源类别数量;提取第二层属性时,由于包含百科类、图书目录类、网页标题类及主题词表 4 类资源,因此 m 值为 4;n 为第 j 类资源中包含有 w_i 的词项数量;$weight_j$ 为第 j 类资源所对应的权重;$w_i.length$ 为 w_i 的词长;

$term_{kj}$ 为第 j 类资源中第 k 个包含有 w_i 的词项，$term_{kj}.length$ 为其词长；$tf_j(term_{kj})$ 为 $term_{kj}$ 在第 j 类资源中的词频；μ 为调节因子，仅对关联词得分进行一定倍数的放大，以便比较。通过实验计算和分析，将百科类、图书目录类、网页标题类及主题词表 4 类资源的权重分别设置为 0.52，0.15，0.13 和 0.20。

(6)将关联词得分高于阈值 Q 的名词保留，作为第二层候选词。通过实验分析，Q 的经验值为 1.5。

(7)计算第二层候选词与第一层概念词的归一化 Google 距离[25]，过滤掉距离过大的词，其余作为第二层概念词的输出。如两个词 x 和 y 之间的归一化 Google 距离为：

$$NGD(x,y)=\frac{\max\{\log f(x),\log f(y)\}-\log f(x,y)}{\log M-\min\{\log f(x),\log f(y)\}} \tag{6}$$

其中，M 表示 Google 索引的网页总数，$f(x)$ 和 $f(y)$ 分别表示搜索词 x 和 y 的命中数量，$f(x,y)$ 表示同时出现 x 和 y 的网页数量。若 x 和 y 从未一起出现在同一网页上，而只是单独出现，则它们之间的归一化 Google 距离无穷大；若 x 和 y 总是同时出现，则它们之间的归一化 Google 距离为 0。

(8)本体第三层概念属性词获取方式为：通过“第一层关键词＋第二层属性”构建检索词，在百度、必应和 Google 搜索检索词获取网页，提取网页的主要内容，标注词性并保留名词，选择 TF-IDF 值较大的词作为候选概念。利用归一化 Google 距离对非领域内概念进行过滤，将过滤后得到的词汇作为本体第三层概念属性词。

3.2.2　语义标注过程

利用本体概念结构对文本进行标注的语义描述过程被称为语义标注。语义标注可视为输入和输出的过程，输入为领域待分割文本，输出为文本自然段落及其语义信息。

本文利用自动获取的结构化语义资源对文本进行语义标注，语义标注过程借鉴文献[26]的思路，即通过对待分割文本进行分词、词性标注、去停用词等处理，仅保留标题中的名词作为概念，并将文本按其自然段进行划分。算法自动获取概念的语义层次结构，然后对切分后的文本进行标注，将拥有相同语义标注信息的相邻段落划分为同一语义段落。语义标注的具体特征规则如下：

(1)若标题中出现概念实体，则其相较于未出现于标题中的概念实体更能表达文本的主题；

(2)若标题中出现概念的属性，则其相较于未出现于标题中的概念属性更能表达文本的主题；

(3)若标题中出现概念实体及其属性，则相较于未出现于标题中的概念及其属性，其更能全面表达文本主题；

(4)若正文段落中出现概念实体，则概念实体出现的次数越多，越能表达该段落的主题；

(5)若正文段落中出现概念的属性，则属性出现的次数越多，越能表达段落的主题；

(6)若正文段落中出现属性的属性词，则属性词出现的次数越多，越能表达段落的主题；

(7)若正文段落中出现属性及其属性词，则相较于未同时出现于正文的属性和属性词，其更能全面表达段落的主题。

根据上述语义标注规则，段落文本语义标注得分的计算方式如下：

$$A=\begin{cases}0.5, & \text{概念出现在标题中}\\ 0, & \text{概念未出现在标题中}\end{cases} \tag{7}$$

$$B=\begin{cases}0.5, & \text{属性出现在标题中}\\ 0, & \text{属性未出现在标题中}\end{cases} \tag{8}$$

$$C=\frac{class_{count}}{words_{count}} \tag{9}$$

$$D=\frac{property\_count}{words\_count} \tag{10}$$

$$E=\begin{cases}\theta\cdot\dfrac{propertywords\_count}{words\_count}, & \dfrac{propertywords\_count}{words\_count}\geqslant P\\ \mu\cdot\dfrac{propertywords\_count}{words\_count}, & \dfrac{propertyword\_count}{words\_count}<P\end{cases} \tag{11}$$

$$ParaSemanticScore=\alpha\cdot A+\beta\cdot B+\gamma\cdot C+\delta\cdot D+E \tag{12}$$

其中，$class\_count$，$property\_count$，$propertywords\_count$ 分别代表结构化语义资源中的概念、属性及属性词，$words\_count$ 表示文本段落的总词数，$\alpha,\beta,\gamma,\delta,\theta,\mu$ 为权重因子，P 为阈值。式(11)引入了加分和罚分机制，语义得分计算式(12)综合考虑了概念、属性以及属性词出现的位置和次数等因素。段落语义得分最高的概念和属性组最能表现该段落的主题。

3.2.3　语义段落划分

利用语义资源对文本进行语义标注，计算每个段落的语义标注分数，得分最高的“概念-属性”组即为段落的主题。语义标注后的输出结果为文本自然段以及其语义标注信息，段落语义信息包括段落的概念及属性。对于包含部分语义标注的段落集合，进行文本分割的思路如下：

(1)对于细颗粒度的主题分割，选取相邻的拥有相同“概念-属性”语义信息的段落，若段落语义标注分数都超过阈值 Y，则将其划分为同一语义段落。

(2)对于粗颗粒度的主题分割，选取相邻的拥有相同“概念”语义信息的段落，若段落语义标注分数都超过阈值 Y，则将其划分为同一语义段落。

经过语义标注后的文本语义段落可以通过切分标注、计算归一化 Google 距离筛选出同领域内的概念，从而补充本体概念结构的第三层属性词，实现语义结构的进化。

4　实验结果与分析

文本分割的评测标准比较主观，主要是因为人们对主题边界的位置以及文本分割颗粒度往往没有一致的观点；另外，不同的应用场景对文本分割的准确性有不同的要求。因此，一部分研究人员将不同主题的文本相连，以确定文本分割边界；另外一部分研究人员则以大多数人的意见为分割标准。本文采用前一种方式构造待分割文本。

最早使用的文本分割评价指标是准确率(Precision)和召回率(Recall)。准确率是指分割得到的正确分割点数目占总分割点数目的百分比；召回率则是指得到的正确分割点数目

占标准分割点数目的百分比。准确率和召回率评价具有片面性，而 F-measure 评价方法则综合了两个指标，其评价公式为：$F=2*$准确率$*$召回率/(准确率$+$召回率)。现在常用的文本分割评测方法有 P_k 评价方法，其公式为：$P_k(ref,hyp)=\frac{1}{N-k}\sum_{i=1}^{N-k}(\delta_{ref}(i,i+k)\text{XOR}\ \delta_{hyp}(i,i+k)>0)$。Pevzner 和 Hearst[27] 改进了 P_k 评价方法，提出了 Windowdiff 评价方法，其定义为：$WindowDiff(ref,hyp)=\frac{1}{N-k}\sum_{i=1}^{N-k}(|b(ref_i,ref_{i+k})-b(hyp_i,hyp_{i+k})|>0)$。

实验选用化学工业领域文本，构造了 10 篇文本作为评测语料集。文本大主题为化工领域的下位概念，小主题为概念的下位概念，即属性。每个文本中包含的段落数目不固定，考虑到通常情况下自然文本的长度，每篇文本的子主题数量为 6～11，平均自然段数目为 27.3。

由于本文的文本分割方法基于段落颗粒度，因此采用准确率、召回率以及 F 值作为评价指标，若分割后的语义段落拥有相同的“概念-属性”语义信息，则分割正确，否则错误。通过研究分析，将段落语义分数计算公式中的权重和阈值分别设置为 $\alpha=1,\beta=1,\gamma=0.8,\delta=0.7,\theta=1.5,\mu=0.6,P=0.024$。实验发现，对于文本存在标题和不存在标题两种情况，当阈值 Y 分别为 0.549 和 0.054 时，文本分割算法具有较高的准确率、召回率以及 F 值，所提方法与其他文本分割方法的对比结果如表 1 所列。

表 1 文本分割实验结果对比

Table 1 Results comparison of different text segmentation methods

	基于知网的文本分割	基于 GA 的文本分割	基于多元判别分析的文本分割	本文方法
准确率	0.54	0.71	0.47	0.85
召回率	0.57	—	0.48	0.90
F 值	0.55	—	0.48	0.88

表 1 中的数据显示，本文的文本分割方法的平均准确率、召回率和 F 值分别达到了 0.85，0.90 和 0.88，基本满足实际应用中的文本分割需求，并且优于现有的大多数中文文本分割方法[1)]，如基于知网的文本分割方法[28]、基于多元判别分析的文本分割方法[29]和基于 GA 的文本分割方法[30]。通过观察分割结果可以发现，本体语义结构覆盖范围越全面，越能挖掘段落的语义信息，分割结果越准确。此外，自然段的长度影响分割结果，长度过短的段落通常缺失属性和属性词，从而造成语义标注的主题颗粒度较大。

结束语 本文归纳并总结了现有的文本分割方法，将其分为基于词汇聚集、基于语言特征以及概率统计思想三大类。通过总结现有分割方法的长处及不足，提出了一种基于领域本体的文本分割方法，其优点在于自动获取语义资源，并利用其中的概念和概念关系识别文本包含的语义以及语义关系，分割过程不需要预设文本块的大小以及分割主题的数量，在划分语义段落时可以获得语义段落的语义信息。实验结果表明，本文方法有效提高了特定领域文本分割的召回率与准确率。

目前，针对中文文本分割的研究以及应用还相对较少，本文所提出的基于领域本体的文本分割方法对于领域知识库检索系统具有实际应用价值，可以减少检索信息冗余。该方法的不足之处在于语义标注的好坏及标注的颗粒度受语义资源的丰富与完整性影响；此外，概念的语义结构获取时间有待缩短。

1) 其他文本分割方法的实验数据来源于文献[28-30]

参 考 文 献

[1] CHOI F Y Y. Advances in domain independent linear text segmentation [C]// NAACL 2000. 2000:26-33.

[2] HALLIDAY, KIRWOOD M A, HASAN R. Cohesion in English [M]. Routledge, 2014.

[3] HEARST M A. TextTiling: segmenting text into multi-paragraph subtopic passages [M]. MIT Press, 1997.

[4] REYNAR J C. An automatic method of finding topic boundaries [C]// Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. 1994:331-333.

[5] REYNAR, JEFFREY C. An Automatic Method of Finding Topic Boundaries [J]. Computer Science, 1994, 14(101):331-333.

[6] KERN R, GRANITZER M. Efficient linear text segmentation based on information retrieval techniques [C] // International Conference on Management of Emergent Digital Ecosystems. ACM, 2009:25.

[7] WU J W, TSENG J C R, TSAI W N. An Efficient Linear Text Segmentation Algorithm Using Hierarchical Agglomerative Clustering[C]// Seventh International Conference on Computational Intelligence and Security. IEEE Computer Society, 2011: 1081-1085.

[8] KAZANTSEVA A, SZPAKOWICZ S. Linear text segmentation using affinity propagation [C] // Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011:284-293.

[9] BAYOMI M, LEVACHER K, GHORAB M R, et al. OntoSeg: A Novel Approach to Text Segmentation Using Ontological Similarity [C] // IEEE International Conference on Data Mining Workshop. IEEE, 2016:1274-1283.

[10] REYNAR J C. Statistical Models for Topic Segmentation [C] // Proc. of Annual Meeting of the Association for Computational Linguistics, 1999. 1999:357-364.

[11] KAN M Y, KLAVANS J L, MCKEOWN K R. Linear Segmentation and Segment Significance[C]// WVLC-6. 1998:197-205.

[12] KAUCHAK D, CHEN F. Feature-based segmentation of narrative documents[C]// ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing. Association for Computational Linguistics, 2005:32-39.

[13] CHOI F Y Y, WIEMER-HASTINGS P, MOORE J. Latent Semantic Analysis for Text Segmentation[J]. Proceedings of Emnlp, 2001, 4(3):109-117.

[14] BRANTS T, CHEN F, TSOCHANTARIDIS I. Topic-based document segmentation with probabilistic latent semantic analysis[C]// Eleventh International Conference on Information and Knowledge Management. ACM, 2002:211-218.

(下转第 156 页)

fold Analysis of Face Pictures and Regression on Aging Features [C]// IEEE International Conference on Multimedia and Expo. IEEE, 2007: 1383-1386.

[10] GUO L L, DING S F. Research Progress on Deep Learning[J]. Computer Science, 2015, 42(5): 28-33. (in Chinese)
郭丽丽, 丁世飞. 深度学习研究进展[J]. 计算机科学, 2015, 42(5): 28-33.

[11] PAN Q X, DONG H B, HAN Q L, et al. A computing method for attribute importance based on BP neural network[J]. Journal of University of Science and Technology of China, 2017(1): 18-25. (in Chinese)
潘庆先, 董红斌, 韩启龙, 等. 一种基于 BP 神经网络的属性重要性计算方法[J]. 中国科学技术大学学报, 2017(1): 18-25.

[12] LEVI G, HASSNCER T. Age and gender classification using convolutional neural networks[C]// Computer Vision and Pattern Recognition Workshops. IEEE, 2015: 34-42.

[13] DONG Y, LIU Y, LIAN S. Automatic age estimation based on deep learning algorithm[J]. Neurocomputing, 2016, 187: 4-10.

[14] ZHUANG F Z, LUO P, HE Q. Survey on Transfer Learning Research[J]. Journal of Software, 2015, 26(1): 26-39. (in Chinese)
庄福振, 罗平, 何清. 迁移学习研究进展[J]. 软件学报, 2015, 26(1): 26-39.

[15] LI Y D, HAO Z B, LEI H. Survey of convolutional neural network[J]. Journal of Computer Applications, 2016, 36(9): 2508-2515. (in Chinese)
李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述[J]. 计算机应用, 2016, 36(9): 2508-2515.

[16] OJALA T, PIETIKÄLNEN M, MÄENPÄÄ T. Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2000, 1842(7): 404-420.

[17] AHONEN T, HADID A, PIETIKÄLNEN M. Face Recognition with Local Binary Patterns[M]. IEEE Computer Society, 2006.

[18] ZHOU Z H, CHEN S F. Neural Network Ensemble[J]. Chinese Journal of Computers, 2002, 25(1): 1-8. (in Chinese)
周志华, 陈世福. 神经网络集成[J]. 计算机学报, 2002, 25(1): 1-8.

[19] KROGH A, VEDLEBSBY J. Neural network ensembles, cross validation and active learning[C]// International Conference on Neural Information Processing Systems. MIT Press, 1994: 231-238.

[20] SHAN C. Learning local features for age estimation on real-life faces[C]// ACM International Workshop on Multimodal Pervasive Video Analysis. ACM, 2010: 23-28.

(上接第 132 页)

[15] MISRA H, JOSE J M, CAPPE O. Text segmentation via topic modeling: an analytical study[C]// DBLP. 2009: 1553-1556.

[16] SUN Q, LI R, LUO D, et al. Text segmentation with LDA-based Fisher kernel[C]// Proceedings of the, Meeting of the Association for Computational Linguistics on Human Language Technologles: Short Papers. 2008: 269-272.

[17] RIEDL M, BIEMANN C. TopicTiling: a text segmentation algorithm based on LDA[C]// Student Research Workshop. Association for Computational Linguistics, 2012: 37-42.

[18] YU K, LI Z, GUAN G, et al. Unsupervised text segmentation using LDA and MCMC[C]// Tenth Australasian Data Mining Conference. Australian Computer Society, Inc. 2012: 21-26.

[19] EISENSTEIN J, BARZILAY R. Bayesian unsupervised topic segmentation[C]// Conference on Empirical Methods in Natural Language Processing (EMNLP 2008). DBLP, 2008: 334-343.

[20] DU L, BUNTINE W, JOHNSON M. Topic Segmentation with a Structured Topic Model[C]// Naacl-Hlt. 2013: 190-200.

[21] KERN R, GRANITZER M. Efficient linear text segmentation based on information retrieval techniques[C]// International Conference on Management of Emergent Digital Ecosystems. ACM, 2009: 25.

[22] CHANG P, MA H. Efficient short text subject extraction method [J]. Computer Engineering and Applications, 2011, 47(20): 126-128. (in Chinese)
常鹏, 马辉. 高效的短文本主题词抽取方法[J]. 计算机工程与应用, 2011, 47(20): 126-128.

[23] LIU Y, SUI Z F, HU Y W, et al. Domain Ontology automatic construction research [J]. Journal of Beijing University of Posts and Telecommunications, 2006, 29(s2): 65-69. (in Chinese)
刘耀, 穗志方, 胡永伟, 等. 领域 Ontology 自动构建研究[J]. 北京邮电大学学报, 2006, 29(s2): 65-69.

[24] GONG X W, LIU Y. Research on Construction of Integrated Semantic Crawler [J]. ICIC Express Letters, Part B: Applications, 2016, 7(7): 1591-1598.

[25] CILIBRASI R L, VITANYI P M B. The Google Similarity Distance[J]. IEEE Transactions on Knowledge & Data Engineering, 2004, 19(3): 370-383.

[26] LIU Y, SHI H Q, ZHENG D J. Study on semantic annotation for professional literature[J]. ICIC Express Letters (Part B), 2014, 5(5): 1383-1389.

[27] PEVZNER, HEARST, MARTI A. A critique and improvement of an evaluation metric for text segmentation[J]. Computational Linguistics, 2002, 28(1): 19-36.

[28] ZHU H J, ZHANG G P, CAI D F, et al. Application of Knowledge Network in Text Segmentation Algorithm [C]// International Conference on Information Processing. 2007. (in Chinese)
朱海军, 张桂平, 蔡东风, 等. 知网在文本分割算法中的应用[C]// 中文信息处理国际会议. 2007.

[29] ZHU J B, YE N, LUO H T. A text segmentation model based on multiple discriminant analysis [J]. Journal of Software, 2007, 18(3): 555-564. (in Chinese)
朱靖波, 叶娜, 罗海涛. 基于多元判别分析的文本分割模型[J]. 软件学报, 2007, 18(3): 555-564.

[30] ZHONG B B, LIU Y C, XU Z M. Study on Parameter Optimization in Text Sub-topic Segmentation Based on GA [J]. Computer Engineering and Applications, 2005, 41(21): 97-99. (in Chinese)
钟彬彬, 刘远超, 徐志明. 基于 GA 的文本子主题切分中的参数优化研究[J]. 计算机工程与应用, 2005, 41(21): 97-99.