Journal of Computer Applications
计算机应用,2015,35(5):1314 - 1319
ISSN 1001-9081
CODEN JYIIDU
2015-05-10
http://www.joca.cn

文章编号:1001-9081(2015)05-1314-06
doi:10.11772/j.issn.1001-9081.2015.05.1314

基于多元判别分析的汉语句群自动划分方法

王荣波[1],李　杰[1*],黄孝喜[1],周昌乐[1,2],谌志群[1],王小华[1]
(1.杭州电子科技大学 认知与智能计算研究所,杭州 310018;　2.厦门大学 智能科学与技术系,福建 厦门 361005)
(*通信作者电子邮箱 578099669@qq.com)

摘　要: 针对目前句群划分工作缺乏计算语言学数据支持、忽略篇章衔接词的问题以及当前篇章分析较少研究句群语法单位的现象,提出一种汉语句群自动划分方法。该方法以汉语句群理论为指导,构建汉语句群划分标注评测语料,并且基于多元判别分析(MDA)方法设计了一组评价函数 J,从而实现汉语句群的自动划分。实验结果表明,引入切分片段长度因素和篇章衔接词因素可以改善句群划分性能,并且利用 Skip-Gram Model 比传统的向量空间模型(VSM)有更好的效果,其正确分割率 P_μ 达到 85.37%、错误分割率 *WindowDiff* 降到 24.08%。同时该方法在句群划分任务上有更大的优势,比传统 MDA 方法有更好的句群划分效果。

关键词: 汉语句群划分;多元判别分析;篇章分析;Skip-Gram 模型;篇章衔接
中图分类号: TP391　　**文献标志码:** A

Automatic Chinese sentences group method based on multiple discriminant analysis

WANG Rongbo[1], LI Jie[1*], HUANG Xiaoxi[1], ZHOU Changle[1,2], CHEN Zhiqun[1], WANG Xiaohua[1]
(1. *Institute of Cognitive and Intelligent Computing, Hangzhou Dianzi University, Hangzhou Zhejiang* 310018, *China*;
2. *Department of Intelligent Science and Technology, Xiamen University, Xiamen Fujian* 361005, *China*)

Abstract: In order to solve the problems in Chinese sentence grouping domain, including the lack of computational linguistics data and the joint makers in a discourse, this paper proposed an automatic Chinese sentence grouping method based on Multiple Discriminant Analysis (MDA). Moreover, sentences group was rarely considered as a grammar unit. An annotated evaluation corpus for Chinese sentence group was constructed based on Chinese sentence group theory. And then, a group of evaluation functions J was designed based on the MDA method to realize automatic Chinese sentence grouping. The experimental results show that the length of a segmented unit and one discourse's joint makers contribute to the performance of Chinese sentence group. And the Skip-Gram model has a better effect than the traditional Vector Space Model (VSM). The evaluation parameter P_μ reaches to 85.37% and *WindowDiff* reduces to 24.08% respectively. The proposed method has better grouping performance than that of the original MDA method.

Key words: Chinese sentences grouping; Multiple Discriminant Analysis (MDA); discourse analysis; Skip-Gram model; discourse coherence

0　引言

汉语句群自动划分,就是把段落中在结构和语义上有密切联系的一组句子划分出来。汉语句群划分是句群理论的重要研究方向,在篇章分析、机器翻译等方面有重要的作用。

1)对篇章分析来说,句群是句子和篇章的过渡阶层,对句群的处理是计算机从理解孤立的词义和句义上升到理解语篇整体中心内容的一个重要跃变步骤,实现这一步骤的前提是篇章的句群划分。

2)在机器翻译领域,著名的自然语言处理专家董振东提出:翻译应该在句群层面上进行。但是目前的理论和技术基础还无法做到这一点。多年来相关研究主要侧重在单句的翻译上,忽略了研究句子和句子之间的联系,缺乏句群层次上的词汇消歧、冠词、代词、时态、省略等处理技术,要想进一步提高机器翻译的质量,必须解决句群自动划分问题。

在计算机领域,汉语句群划分工作较为少见,更多的是基于概念层次网络(Hierarchical Network of Concepts,HNC)语境观的篇章分析工作,其更接近规则的方式,缺乏足够的计算语言学数据支持,并不是明确的可计算模型。同时,其他相关的句群划分研究未能考虑篇章衔接词的作用,实验效果也不理想。汉语句群自动划分是篇章分析的重要组成部分,目前篇章分析较多将国外篇章分析理论应用于篇章衔接分析工作,缺少对汉语句群这一级语法单位的研究。

针对以上问题,本文以汉语句群理论为指导,构建了基于句群理论的句群划分语料标注集,提出了一种基于多元判别分析(Multiple Discriminant Analysis, MDA)[1]的汉语句群自

收稿日期:2014-12-05;修回日期:2014-12-24。　基金项目:国家自然科学基金资助项目(61202281,61103101);教育部人文社会科学研究项目青年基金资助项目(10YJCZH052,12YJCZH201)。

作者简介:王荣波(1978 -),男,浙江义乌人,副教授,博士,CCF 会员,主要研究方向:自然语言处理、篇章分析;　李杰(1989 -),男,浙江温州人,硕士研究生,主要研究方向:中文信息处理;　黄孝喜(1979 -),男,浙江温州人,讲师,博士,主要研究方向:自然语言处理、认知逻辑学;周昌乐(1959 -),男,苏州太仓人,教授,博士,主要研究方向:人工智能、中文信息处理;　谌志群(1973 -),男,江西南昌人,副教授,硕士,主要研究方向:中文信息处理、语言网络;　王小华(1961 -),男,浙江温州人,教授,主要研究方向:自然语言处理、模式识别。

动划分方法。具体步骤如下: 1) 通过 Skip-Gram Model[2-3] 训练大规模语料获取词语在低维实数空间的向量表示,在此基础上获得句子的特征向量; 2) 在句子特征向量构成的数据空间中,通过考虑句群内部距离、句群之间距离、切分片段长度以及篇章衔接词四个因素设计基于 MDA 方法评价函数 J,对段落所有可能的句群划分结果进行评价; 3) 在所有评价结果中,寻找评价值最高的句群划分序列,则该序列为该段落的最佳句群划分结果。实验结果表明该方法对句群划分有较好的效果。

1　句群研究现状

1.1　国内外句群研究现状

实际上,国外篇章语言学研究往往侧重于语篇中句际的衔接和连贯,忽视句群这一层次。这是因为英语是分析性的语言,英语有发达的关系词和关系形式,注重语法形式的完整性。韩礼德(Halliday)和哈桑(Hasan)合著的《英语的衔接》中,作者对篇章下了定义,篇章与句子或小句的关系不在于篇幅的长短,而在于衔接[4]。书中给出篇章的定义,但未在篇章和句子之间的层级有所研究。汉语句群理论不同于国外篇章分析理论,它是具有汉语特征的篇章分析理论。英语段落一般只有一个中心意义(controlling idea),常常含有主题句(topic sentence),其他句子则围绕此主题展开句间的联系。而中文段落可以有多个主题,包含多个句群,只要谈论的是相关话题就可以构成一个段落[5]。实际上,在汉语中句际间的衔接和连贯,往往先在句群中实现,然后才是各句群之间的衔接和连贯在篇章中实现。

由于语言学的指导作用,在计算语言学上,国外似乎对句群也并不在意,反而更重视句际的衔接和连贯[6-8]。例如,修辞结构理论(Rhetorical Structure Theory,RST)[7]认为篇章理解的基本单位(Elementary Discourse Unit,EDU)是子句(Clause)。该理论并不关心篇章的下一级单位,而是更关心子句间的关系识别和衔接。又如,篇章词汇化树型连接语法理论(Discourse Lexical Tree Adjunct Grammar,D-LTAG)[8]同样也是篇章连接词驱动的篇章理论。

国内部分语言学研究者认识到句群研究的重要性,相继开展了有关句群特点分析、结构分析及句群教学应用等研究工作,如吴为章等[9]、郝长留[10]以及曹政[11]等。但是,目前在计算语言学领域,对句群的研究还比较少见,更多的是基于不同篇章理论的篇章分析工作[12-20]。

1.2　自动划分研究现状和缺陷

汉语句群自动划分属于篇章分析问题的相关研究。根据篇章分析时所依据的理论不同,国内对句群划分和篇章分析工作可以分为两类:

一类是引用国外的篇章理论完成对汉语篇章分析工作。陈莉萍[12]认为句群理论可以经过修改后作为切实可行的中文篇章分析理论,其认为句群理论和 RST 在研究对象、研究内容、研究方法和呈现形式等方面都极其相似。高芸[13]认为句群研究可以借鉴分段式话语表现理论(Segmented Discourse Representation Theory,SDRT)[14],从而形成一套系统的理论和方法。在隐式篇章关系识别方面,徐凡等[15]、周小佩等[16]在宾州篇章树库(Penn Discourse TreeBank,PDTB)的基础上作了相关研究,该树库是借鉴了 D-LTAG 和 RST 思想构建的篇章语料库。又如张益民等[17]提出了一种混合确定性中文篇章分析方法,它是 RST 分析、主位模式分析、向量空间模型等方法的混合。以上研究是以国外篇章理论对中文篇章衔接关系等方面的篇章分析研究,但是较少涉及在句群这一介于句子与篇章的语法层级的研究。

另一类是基于汉语句群理论的句群自动划分工作。吴晨等[18]从句群本身的构成特点出发,对句群进行了内部语义组合方式的划分,并根据已经取得的"HNC 语言概念空间表示"的研究成果,制定了识别具有以上构成特点句群的相关规则。缪建明等[19]在 HNC 语境观的指导下,通过对领域句类知识和句类知识的整合,形成了带有一定理解含义的语境框架结构。上述研究在 HNC 语境观指导下对句群分析做了一定工作,其更接近规则的方式处理,依赖于一定的领域知识,并不是明确的可计算模型,都缺乏足够的计算语言学数据支持。在计算机领域,汉语句群自动划分的研究较为少见,Chen 等[20]提出了一种基于层次聚类和关键词局部重现度的多重句群自动划分算法,这是对一个多重句群进行层次划分的工作,但是未能考虑到篇章衔接词(句间指代词、句间关联词)带来的影响,实验效果也并不理想。

由此可见,句群研究虽然在篇章理论上有着相当的研究,但是在计算机领域的研究还不够深入。而如何使用具有汉语特色的汉语句群理论为指导,并且引入篇章衔接词因素,提出一种可计算的模型用于汉语句群划分的实现,是本文研究的主要问题。

本文以汉语句群理论为指导,构建了基于句群理论的句群划分语料标注集,提出了基于多元判别分析(MDA)实现汉语句群的自动划分方法; 同时还要考虑句间指代词和关联词的作用,实现篇章段落逻辑结构的自动划分工作。实验结果表明,相比传统的 MDA 方法,本文方法在句群划分工作上有更大优势。

2　汉语句群自动划分模型

2.1　基于 MDA 的句群自动划分方法

1) 句子语义的形式化描述。

句子是句群的基本单位,汉语句群自动划分首先要获取句子的特征向量表示。常用句子向量表示方式是向量空间模型(Vector Space Model, VSM),其通过文本表面词汇数量统计实现,会带来数据稀疏以及维数灾难问题,因此需要更有效的方式挖掘词语的深层语义信息,并对句子语义的形式化描述。

Skip-Gram Model[2-3]是一种基于神经网络语言模型(Neutral Network Language Model, NNLM)[21]的词向量(Distributed Representation)训练模型,其基本思想是通过训练将词语投影到其向量空间,为词语寻找在低维实数空间的向量表示。该词向量能能够用低维向量表征一个词语并更好地表达词语的语义信息。

Skip-Gram 中词向量存在线性可加性(Additive Compositionality),设每个句子 S_i 对应一个特征向量 $\boldsymbol{v}_i$:

$$\boldsymbol{v}_i = \sum_{j=1}^{n} \boldsymbol{C}(w_j) \Big/ \Big\| \sum_{j=1}^{n} \boldsymbol{C}(w_j) \Big\| \tag{1}$$

其中: n 为句子中词语数,$\boldsymbol{C}(w_j)$ 为句子中的词语 w_j 的词向量表示。

2) 句群划分数学模型。

定义一个段落 T 为句子序列 $T=\{S_1, S_2, \cdots, S_n\}$（$n$ 表示段落 T 包含句子的个数）。段落中第 i 个句子 S_i 是一个词语序列 $S_i=\{w_1, w_2, \cdots, w_m\}$（$m$ 表示句子中的词语个数）。定义段落 T 可能的句群划分模式为 $D=\{d_1, d_2, \cdots, d_c\}$，其中 d_i 表示划分出的第 i 个片段（句子或句群），c 表示句群划分后包含的片段 d 的个数。存在一个对应的分割点序列 $B=\{b_0, b_1, \cdots, b_c\}$，其中 b_0 和 b_c 为潜在固定的分割点。如图1所示。

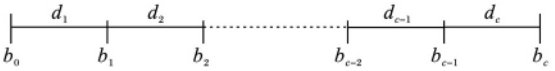

图1 汉语句群划分模式示意图

给定一个段落 T，句群划分模型的关键问题转换成寻求具有最大概率的划分模式。即：

$$\hat{D}=\arg\max_D P(D\mid T) \tag{2}$$

3) 多元判别分析方法。

实际上，直接利用上述句群划分数学模型计算具有最大概率的划分模式比较困难，因此通过设计基于 MDA 方法的评价函数 J，计算所有可能的句群划分模式 D 的评价值，选择最高评价值的句群划分模式为正确划分结果。

MDA 是一种独立于具体领域的文本线性分割统计模型方法[1]。该方法本质是将高维数据投射到低维空间上，关键在于寻找最能分开各类数据的投影方法，以解决多类别分类问题。MDA 基于类别可分离性判定，其基本思想是：数据空间中各类样本位于数据空间的不同区域内，这些区域之间距离越大，类别的可分离性就越大。具体到句群划分任务中，句群是介于句子与篇章的语法单位，可以将每个句群视为一个类别。句群自动划分任务，即是以句子为基本单位的文本线性分割过程，通过 MDA 方法找到最能分开各个句群的划分模式。接下来具体讲解基于 MDA 方法在汉语句群自动划分的实现过程。

4) 基本流程描述。

具体的说，首先通过 Skip-Gram 将段落中每个句子特征向量化，则段落中所有的句子特征向量可以构成一个数据空间（Data Space），如图2所示。将每个句群视为一个类别，每个句子向量为样本向量，句群划分的过程就是数据空间的分割过程：句群内部距离越小（强内聚性），句群外部距离越大（强发散性），句群划分效果越好。句群自动划分过程就是在数据空间中找到其最佳分割的过程。

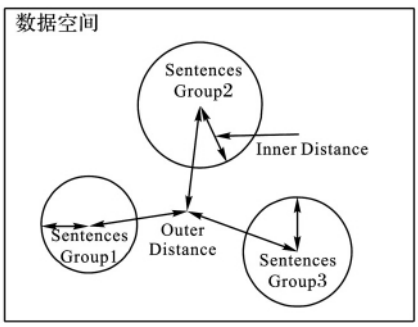

图2 数据空间的最佳划分

为了找到这个最佳划分，通过设计一个基于 MDA 的评价函数 J，评价每个划分模式 D 的评价值，从而在句子特征向量构成的数据空间 $T(\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n)$ 中寻找最优的句群划分结果：

$$\hat{D}=\arg\max_D J(D, T(\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n)) \tag{3}$$

所以，问题的关键在于评价函数 J 的设计。本文评价函数 J 主要考虑4个因素：句群内部距离、句群之间距离、切分片段长度以及篇章衔接词。分别通过句群内离散矩阵 $\boldsymbol{S}_w$、句群间离散矩阵 $\boldsymbol{S}_b$、切分片段长度惩罚因子 S_L、篇章衔接词惩罚因子 S_d 进行计算。在所有可能的句群划分模式中，评价值最高的句群划分模式 D 为最终句群划分结果。此时，

$$\hat{D}=\arg\max_D P(D\mid T)=_{def}\arg\max_D J(D, \boldsymbol{S}_w, \boldsymbol{S}_b, S_L, S_d) \tag{4}$$

2.2 评价函数 J 的设计

句群的划分涉及它的组合关系。《汉语句群》[2] 中提到无论是句群内部句子与句子之间，还是句群外部句子与句群之间，都存在着连贯性，有时体现于语义内容上，有时体现于表达形式上，有时两者兼有。对于句群内部组合提出6种接应方式，即“词语接应”“句式接应”“辞格接应”“音气接应”“语意接应”和“伴随语言接应”。与句群组合相对应的是句群的划分，句群组合的方式自然也就成为句群划分的理论基础。大多数句群的组合理论都采用一定的形式手段，这些手段是辨认和划分句群的可靠标志。

在汉语句群理论的指导下，一方面，在评价函数 J 中，类内离散矩阵 $\boldsymbol{S}_w$ 可用于衡量句群内部的内聚程度；类间离散矩阵 $\boldsymbol{S}_b$ 可用于衡量句群之间的离散程度[1]。其表现的是句群这级语法单位在语义内容上的连贯性。通过类内离散矩阵 $\boldsymbol{S}_w$ 和类间离散矩阵 $\boldsymbol{S}_b$ 可以表现句群这一级语法单位的内聚程度以及句群之间的离散程度。另一方面，《汉语句群》中提出句群的一个重要特点是意义上的相对完整性，也就是说句群中的每个成分，它们所表达的事件或所指称的客体是相互关联的，甚至是同一的，所以本文引入切分片段长度惩罚因子 S_L。另外，衔接是篇章的重要特征，因此要考虑篇章衔接词的影响，句间关联词和句间指代词在句群组合有重要作用。这里对应于句群内部组合方式的“词语接应”，并且表现了表达形式的连贯性，因此本文引入了篇章衔接词惩罚因子 S_d。

1) 切分片段长度惩罚因子 S_L。

在一个划分模式 D 中，分割点十分接近时容易切分出一个单句，将其作为独立的句群，但是段落中的句子常常起到承上或启下的作用，较少作为独立的句群出现。另一方面，句群是由两个或两个以上的句子组成的。在较长的篇章段落中单个句子常常无法表达一个独立完整的中心意思，因此当划分模式 D 连续将单个句子划分为一个片段时，常常将一个句群切分成若干连续的单句，每个句子是一个片段，这种划分序列并不合理，应当进行惩罚。

针对以上现象，引入切分片段长度惩罚因子 S_L：

$$S_L=S_{L\_1}/S_{L\_2} \tag{5}$$

$$S_{L\_1}=-\ln\left(\prod_{i=1}^{c}L_i/L\right) \tag{6}$$

$$S_{L\_2}=1/(1+\mathrm{e}^{-\lambda});\ \lambda=\prod_{i=1}^{k}2/(n_iL) \tag{7}$$

其中：L_i 是片段 d_i 中句子个数，L 为句子的总个数；n_i 是分割点序列 B 中连续将单句划分为一个片段时所用的分割点个数，$n_i\geqslant 2$。例如分割点序列片段（b_0, b_1, b_2）有3个分割点，并且

对应的片段 d_1 和 d_2 的句子长度都为 1,即切分出的片段都是单句,则 $n_i = 3$,k 为这种连续划分片段的个数,特别地当 $k = 0$ 时,$S_{L\_2} = 1$。

S_{L_1} 保证了段落切分的长度越均匀,其值越小,句群划分效果相对越好。S_{L_2} 是 sigmoid 函数,sigmoid 函数是个良好的阈值函数,其值在(0,1) 内,关于(0,0.5) 中心对称,呈 S 型单调递增,保证了连续划分单句的片段个数 k 越少且 k 值越小,则相应的惩罚因子 S_L 越小。

2) 篇章衔接词惩罚因子 S_d。

作为语法的最大单位,句群是由句子依照一定的规则组合而成的。除了在语义内容上,句子间在表达形式上也具有连贯性,这种连贯性通过接应(cohesion) 实现。"词语接应"是其中一种重要的方式,主要包括句间关联词和句间指代词。

《汉语句群》中提到在句群内部接应时句间关联词和指代词的分布有其特定的规律[3]:较少使用成对出现的关联词;单用一个关联词时,一般不出现在句群的第一个句子里;以指代词开头的句子一般不在句群的第一个句子里;某些复句常用的关联词不用来组合句群。因此,当出现以上情况时,应当对其进行惩罚。

设 τ 为划分模式 D 中 d_i 首句存在篇章衔接词的片段个数,其中 $i > 1$。S_d 定义如下:

$$S_d = 1/(1 + e^{-x}) \tag{8}$$

其中:x 为关于 τ 的函数。取 $x = \tau - 2$,当存在 2 个篇章衔接词时,惩罚因子 $S_d = 0.5$。S_d 同样是 sigmoid 函数,是关于 τ 的单调递增函数,τ 越大,对划分序列的罚分越大;相反,τ 越小,对划分序列的罚分越小,句群划分效果越好。

3) 评价函数 J。

设计如下一组不同的评价函数 J_1、J_2、J_3 和 J_4 分别对划分模式 D 进行评价:

$$J_1 = \mathrm{trace}(\boldsymbol{S}_b)/\mathrm{trace}(\boldsymbol{S}_w) \tag{9}$$

$$J_2 = \mathrm{trace}(\boldsymbol{S}_b)/(\mathrm{trace}(\boldsymbol{S}_w) * S_L) \tag{10}$$

$$J_3 = \mathrm{trace}(\boldsymbol{S}_b)/(\mathrm{trace}(\boldsymbol{S}_w) * S_d) \tag{11}$$

$$J_4 = \mathrm{trace}(\boldsymbol{S}_b)/(\mathrm{trace}(\boldsymbol{S}_w) * S_L * S_d) \tag{12}$$

其中:$\boldsymbol{S}_b$ 为句群间离散矩阵,$\boldsymbol{S}_w$ 为句群内离散矩阵。trace($\boldsymbol{S}$) 为矩阵 $\boldsymbol{S}$ 的迹,trace($\boldsymbol{S}_b$) 和 trace($\boldsymbol{S}_w$) 分别为句群之间距离和句群内部距离。J_1 考虑了句群间离散矩阵 $\boldsymbol{S}_b$ 和句群内离散矩阵 $\boldsymbol{S}_w$,J_2 同时考虑了切分片段长度惩罚因子 S_L,J_3 考虑了篇章衔接词因子 S_d,J_4 将 4 个因子全部考虑在内。评价函数 J 的值越大,所对应的分割模型 D 越好。

3　实验测试

3.1　语料数据分析

3.1.1　语料设置

1) 实验测评语料(Testing Data) 来源于《读书》杂志(1979—1983),共计 50 期。实验完成 518 个段落的汉语句群划分测试语料库的构建,标注了该评测语料及其参考标准分割答案。每个段落长度为 4 ~ 20 个句子。518 个段落中共有4 136个句子,切分出 1 343 个片段(句群或句子)。实验时,从语料中随机抽取 200 个段落(Set_1) 用于词向量维度的确定,剩下 318 个段落(Set_2) 用于后续性能评价实验。实验测评语料举例如下:

①但是,不管怎样,说话写文章的逻辑经过作者 20 年的努力探索,已初具规模,已草创了一个独立的体系,为进一步研究奠定了基础。②据知,结合写作研究逻辑的还有不少同志,有的亦已出了成果,或者正在从事著述。③可以预料,说话写文章的逻辑作为人们日常普通思维的工具,将会发展成为一个具有相对独立性的新兴逻辑学科分支。|④我衷心希望作者们继续探索下去,使之进一步充实和完善起来。⑤我也衷心希望在其他逻辑学科如辩证逻辑、数理逻辑等等方面能够有这样探索性的著作尽快地、更多地问世。|⑥各个逻辑学科和分支不是互相排斥的,而是互相促进,互相补充的。⑦看来,逻辑科学,也和其他科学一样,正在朝着分化(分化中有溶合) 的趋势发展。⑧我们不应该有任何门户之见,而应该鼓励和支持各种不同的逻辑学科和分支的发展以及与此相联系的不同逻辑学派的形成。⑨只有这样,才能使逻辑科学日益发展,适应四化的需要,适应现代人的思维的需要。

有益的探索——评《说话写文章的逻辑》王秀贵

第 9 句中"这样"指代上一句,当第 8 句和第 9 句之间存在分割时,显然不合理,应当引入篇章衔接词惩罚因子。在句群内部接应时句间关联词和指代词的分布有其特定的规律,而第 9 句中"只有…才…"是成对出现的关联词,其作用为句内连接。因此建立合适的篇章衔接词表很有必要。句间关联词表 Dic1,共计 54 个词汇,比如"并且""尤其"等;句间指代词表 Dic2,共计 47 个词汇,比如"他""这""其"等。

2) 词向量训练语料(Training Data) 来源于《读书》(1979—1998),共计 237 期。使用词向量训练工具 word2vec 实现。

3) 通过评价函数 J 得到段落的最优句群划分结果后,需要对该句群划分结果正确性进行评价。本文采用 P_μ 评价方法[22] 和 *WindowDiff* 评价方法[23] 进行评价。

P_μ 评价指标统计了两个随机选择的句子被算法正确识别为属于同一片段或者不属于同一片段的概率。算法得到的分割点越接近实际分割点,P_μ 评价值越高。计算公式如下:

$$P_\mu(ref,hyp) = \sum_{1\leqslant i\leqslant j\leqslant N} \gamma_\mu e^{-\mu|i-j|}(\delta_{ref}(i,j) \overline{\oplus} \delta_{hyp}(i,j)) \tag{13}$$

式中:ref 指人工判断的分割模式;hyp 是指算法给出的分割模式;N 为段落的句子数数量;$\delta_{ref}(i,j)$ 是表示从段落中随机选择的第 i 句话与第 j 句话是否属于同一个片段:如果属于人工判断的同一片段则 $\delta_{ref}(i,j) = 1$;否则 $\delta_{ref}(i,j) = 0$。同理,$\delta_{hyp}(i,j)$ 是根据算法对段落划分的情况判断这两个句子是否在同一片段;$\gamma_\mu e^{-\mu|i-j|}$ 根据句子间的距离分配不同的权值,γ_μ 为归一化权值,$\gamma_\mu = 1/\sum e^{-\mu|i-j|}$($i$ 和 j 为自变量)。μ^{-1} 为段落中所有片段句子数的平均值。

不同于 P_μ 评价指标对正确的分割点评价,*WindowDiff* 对不正确的分割点作出惩罚。*WindowDiff* 评价值越低,算法结果越好。计算公式如下:

$$WindowDiff(ref,hyp) = \frac{1}{N-k}\sum_{i=1}^{N-k}(|b(ref_i,ref_{i+k}) - b(hyp_i,hyp_{i+k})| > 0) \tag{14}$$

式中:$b(i,j)$ 表示位置 i 和位置 j 之间分割点的个数,N 表示段落中的词语总数,k 为段落中所有片段词语数量平均值的

一半。

3.1.2 一致性检验

表1通过 *Kappa* 值进行语料一致性检验，标准标注为本实验中使用的参考标注，新的标注为重新由新的标注者对语料的标注结果。其中，“1”和“0”为分割点序列(b_1，b_2，…，b_{c-1})的标注情况，“1”代表待分割点 b_i 为“分割”，“0”代表“不分割”。表格中统计的是518个段落分割点的分布。$N = a + b + c + d$，则观察一致率 $Po = (a + d)/N = 0.9113$，机遇一致率 $Pe = [(a + c)(a + b)/N + (b + d)(c + d)/N]/N = 0.6379$。$Kappa = (Po - Pe)/(1 - Pe) = 0.7550$。$Kappa \geqslant 0.8$ 表示语料高度一致，$Kappa > 0.6$ 表示语料相对一致可靠。

表1 语料一致性检验

标注状态	标准标注		合计
	1	0	
1	698(a)	194(b)	892($a + b$)
0	127(c)	2599(d)	2726($c + d$)
总计	825($a + c$)	2793($b + d$)	3618($a + b + c + d$)

3.2 实验结果与分析

3.2.1 实验结果与结论

表2展示在不同词向量维度下，Skip-Gram 在 J_4 评价函数下的结果。测试语料为 Set_1，在不同词向量维度下，评价指标结果的标准差(std)较小表示结果较为稳定，平均正确分割率 P_μ 为85.22%，错误分割率 *WindowDiff* 为25.41%，都得到较好的句群划分性能。后续实验随机取词向量维数300维，测试语料使用 Set_2。

表2 Skip-Gram Model 在 J_4 评价函数下的结果(语料 Set_1)%

评价指标	维数				*Std*
	100	200	300	400	
P_μ	85.23	85.08	85.22	85.35	0.096
WindowDiff	25.41	25.72	25.45	25.06	0.235

表3对比了评价函数 J 对句群划分的性能的影响。实验结果表明，考虑切分片段长度和篇章衔接词可以提高句群划分性能，同时 Skip-Gram 性能优于传统的 VSM。评价函数 J_4 综合了句群内部距离、句群之间距离、切分片段长度以及篇章衔接词4个因素，相比 J_1、J_2、J_3 得到最佳的实验效果。通过实验表明，评价函数 J_4 为最佳的评价函数。表4展示了本文方法与传统的 MDA 方法(评价函数为 J')对比实验。实验结果表明，本文方法在句群划分任务上有更大的优势。

表3 不同评价函数下的实验结果(语料 Set_2) %

模型	评价指标	函数			
		J_1	J_2	J_3	J_4
VSM	P_μ	77.38	79.85	80.01	82.98
	WindowDiff	45.28	36.51	42.27	29.05
Skip-Gram Model	P_μ	79.07	82.55	81.32	85.37
	WindowDiff	44.65	30.39	41.11	24.08

表4 本文方法与传统 MDA 方法的比较 %

评价指标	MDA(J_4，Skip-Gram)	传统 MDA(J'，Skip-Gram)
P_μ	85.37	81.74
WindowDiff	24.08	28.57

3.2.2 实验结果分析

一方面，衔接是篇章的重要特征，要考虑篇章衔接词的影响与句子表达形式的连贯性。这是句群划分任务引入篇章衔接词惩罚因子 S_d 的必要性。在表3中，评价函数 J_3 考虑篇章衔接词的影响，相比 J_1 评价函数，P_μ 指标提高2% ~ 3%，*WindowDiff* 指标降低3% ~ 4%。评价函数 J_4 同样综合了篇章衔接词的影响，得到了最佳的实验效果。在表4中，由于传统 MDA 未考虑篇章衔接词因素，本文评价方式 J_4 也比传统 MDA 有更好的实验效果。

另一方面，切分片段长度对句群划分性能也有提高作用。在表3中，评价函数 J_2 考虑了切分片段长度的影响，与 J_1 相比，J_2 较大幅度提高了句群划分性能。Skip-Gram 的 P_μ 提高了3.5%(VSM 提高2.5%)，*WindowDiff* 降低14%(VSM 降低8.7%)。

另外，通过表3的对比分析，可以发现 Skip-Gram 模型比传统的 VSM 方法有更好的句群划分性能。原因在于句子作为句群的基本单位，数据稀疏情况较为明显，VSM 方法同样存在数据稀疏问题。

结果表明 J_4 为最佳的评价函数。评价函数 J_4 综合了句群内部距离、句群之间距离、切分片段长度以及篇章衔接词四个因素，取得最佳的实验效果。通过对表3、表4比较分析，可以发现在 Skip-Gram 模型下 J_4 得到最佳的句群划分效果，P_μ 达到85.37%、*WindowDiff* 达到24.08%。

在表4中，传统 MDA 方法中评价函数 $J' = [\text{trace}(\boldsymbol{S}_b)/\text{trace}(\boldsymbol{S}_w)] * S_L$，其中 $S_L = \prod(L_i/L)$ 并且没有引入篇章衔接词惩罚因子 S_d，相比传统 MDA 方法本文方法使得 P_μ 提高了3.6%，*WindowDiff* 降低了4.5%。因此实验表明本文方法在句群划分任务上有更大的优势。

4 结语

本文提出了一种基于 MDA 实现自动完成段落中语义关系紧密的句群划分工作。旨在基于中文篇章分析中的句群理论，实现句群边界的识别。这可以为篇章结构的表示和篇章理解工作提供技术基础，并且对基于篇章的机器翻译有重要的意义。实验构建了518个段落的汉语句群划分测试语料库，语料相对一致可靠，为句群研究提供数据支持。实验得到以下结论：1) 实验结果表明考虑切分片段长度和篇章衔接词可以提高句群划分性能；2) 评价函数 J_4 为最佳的评价函数(P_μ为85.37%、*WindowDiff* 为24.08%)；3) Skip-Gram Model 比传统的 VSM 在句群划分任务上有更好的性能；4) 通过本文方法与传统 MDA 方法的比较分析，本文方法在句群划分任务上有更大的优势。

参考文献：

[1] ZHU J, YE N, LUO H. Text segmentation model based on multiple discriminant analysis[J]. Journal of Software, 2007, 18(3): 555 - 564. (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.)

[2] MIKOLOV T, SUTSKEVER I, CHEN K, *et al*. Distributed representations of words and phrases and their compositionality [C]// NIPS 2013: Proceedings of the Advances in Neural Information Processing Systems 26. Cambridge: MIT Press, 2013: 3111 - 3119.

[3] MIKOLOV T, CHEN K, CORRADO G, *et al*. Efficient estimation

of word representations in vector space[C/OL]. [2014 - 04 - 20]. http://arxiv.org/pdf/1301.3781.pdf.

[4] WANG Y. The analysis of English sentence group[J]. Journal of University of Shanghai for Science an Technology: Social Sciences, 2004, 26(2): 30 - 32. (王跃洪. 英语句群分析[J]. 上海理工大学学报: 社会科学版, 2004, 26(2): 30 - 32.)

[5] LUO T. Discussion on using sentences group as valid basic unit for translation from Chinese to English[J]. Journal of Southeast University: Philosophy and Social Science, 2006, 8(3): 110 - 113. (罗天妮. 论以句群为汉英翻译的有效基本单位[J]. 东南大学学报: 哲学社会科学版, 2006, 8(3): 110 - 113.)

[6] XU F, ZHU Q, ZHOU G. Survey of discourse analysis methods[J]. Journal of Chinese Information Processing, 2013, 27(3): 20 - 32. (徐凡, 朱巧明, 周国栋. 篇章分析技术综述[J]. 中文信息学报, 2013, 27(3): 20 - 32.)

[7] MANN W C, THOMPSION S A. Rhetorical structure theory: a theory of text organization[J]. Text, 1988, 3(8): 243 - 281.

[8] WEBBER B. D-LTAG: extending lexicalized TAG to discourse[J]. Cognitive Science, 2004, 28(5): 751 - 779.

[9] WU W, TIAN X. Chinese sentence group[M]. Beijing: The Commercial Press, 2000: 81 - 88. (吴为章, 田小琳. 汉语句群[M]. 北京: 商务印书馆, 2000: 81 - 88.)

[10] HAO C. Text paragraph knowledge[M]. Beijing: Beijing Press, 1983: 1 - 29. (郝长留. 语段知识[M]. 北京: 北京出版社, 1983: 1 - 29.)

[11] CAO Z. Primary research on sentences groups[M]. Hangzhou: Zhejiang Education Publishing House, 1984: 15 - 17. (曹政. 句群初探[M]. 杭州: 浙江教育出版社, 1984: 15 - 17.)

[12] CHEN L. Rhetorical structure theory and sentences group analysis [J]. Journal of Suzhou University: Philosophy and Social Science, 2008, 29(4): 118 - 121. (陈莉萍. 修辞结构理论与句群研究[J]. 苏州大学学报: 哲学社会科学版, 2008, 29(4): 118 - 121.)

[13] GAO Y. Exploring the rhetorical form of Chinese discourse structure from the angle of SDRT[D]. Chongqing: Southwest University, 2011. (高芸. 从 SDRT 的视角探析汉语话语结构的修辞格式[D]. 重庆: 西南大学, 2011.)

[14] ASHER N, LASEARIDE. Logics of conversation[M]. London: Cambridge University Press, 2003: 6 - 35.

[15] XU F, ZHU Q, ZHOU G. Implicit discourse relation recognition based on tree kernel[J]. Journal of Software, 2013, 24(5): 1022 - 1035. (徐凡, 朱巧明, 周国栋. 基于树核的隐式篇章关系识别[J]. 软件学报, 2013, 24(5): 1022 - 1035.)

[16] ZHOU X, HONG Y, CHE T, *et al*. Implicit discourse relation inference based parallel arguments[J]. Computer Applications and Software, 2012, 29(9): 57 - 61. (周小佩, 洪宇, 车婷婷, 等. 基于平行论元的隐式篇章关系推理研究[J]. 计算机应用与软件, 2012, 29(9): 57 - 61.)

[17] ZHANG Y, LU R, SHEN L. A hybrid method for automatic chinese discourse structure analysis[J]. Journal of Software, 2000, 11(11): 1527 - 1533. (张益民, 陆汝占, 沈李斌. 一种混合型的汉语篇章结构自动分析方法[J]. 软件学报, 2000, 11(11): 1527 - 1533.)

[18] WU C, ZHANG Q. Research on rules for detecting Chinese sentence groups in nature language processing[J]. Computer Engineering, 2007, 33(4): 157 - 159. (吴晨, 张全. 自然语言处理中句群划分及其判定规则研究[J]. 计算机工程, 2007, 33(4): 157 - 159.)

[19] MIAO J, ZHANG Q. The study of sentence group based on the HNC context theory[C]// The Research on Content Computing and Its Applications: 9th Chinese National Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2007: 258 - 263. (缪建明, 张全. 基于 HNC 语境理论的句群处理研究[C]// 内容计算的研究与应用前沿: 第九届全国计算语言学学术会议. 北京: 清华大学出版社, 2007: 258 - 263.)

[20] CHEN Y, SHI X. Automatic partition of Chinese sentence group [J]. Journal of Donghua University: English Edition, 2010, 27(2): 177 - 180.

[21] BENGIO Y, SCHWENK H, SENECAL J S, *et al*. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(2): 1137 - 1155.

[22] BEEFERMAN D, BERGER A, LAFFERTY J. Statistical models for text segmentation[J]. Machine Learning, 1999, 34(1/2/3): 177 - 210.

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

(上接第 1313 页)

[9] DING X, LIU G, MENG K, *et al*. Research and improvement of filter algorithm of malicious information based on one-class SVM [J]. Computer Science, 2013, 40(Z2): 86 - 90. (丁霄云, 刘功申, 孟魁, 等. 基于一类 SVM 的不良信息过滤算法改进 [J]. 计算机科学, 2013, 40(Z2): 86 - 90.)

[10] CHEN T, XU R, WU M, *et al*. A sentiment classification approach based on sentiment sentence framework [J]. Journal of Chinese Information Processing, 2013, 27(5): 67 - 74. (陈涛, 徐睿峰, 吴明芬, 等. 一种基于情感句模的文本情感分类方法 [J]. 中文信息学报, 2013, 27(5): 67 - 74.)

[11] PREM M, WOJCIECH G, RICHARD D. Sentiment analysis of blogs by combining lexical knowledge with text classification [C] // Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009: 1275 - 1283.

[12] WU Q, LIU Y, SHEN H, *et al*. A unified framework for cross-domain sentiment classification [J]. Journal of Computer Research and Development, 2013, 50(8): 1683 - 1689. (吴琼, 刘悦, 沈华伟, 等. 面向跨领域情感分类的统一框架 [J]. 计算机研究与发展, 2013, 50(8): 1683 - 1689.)

[13] XU L, LIN H, PAN Y, *et al*. Constructing the affective lexicon ontology [J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 180 - 185. (徐琳宏, 林鸿飞, 潘宇, 等. 情感词汇本体的构造 [J]. 情报学报, 2008, 27(2): 180 - 185.)

[14] MA H, LIU Y, WANG L, *et al*. An analysis and comparison of three methods for document semantic orientation recognition [J]. New Technology of Library and Information Service, 2007(4): 43 - 47. (马海兵, 刘永丹, 王兰成, 等. 三种文档语义倾向性识别方法的分析与比较 [J]. 现代图书情报技术, 2007(4): 43 - 47.)