

基于指代消解的汉语句群自动划分方法

王荣波¹, 孙小雪¹, 黄孝喜¹, 刘和平²

(1. 杭州电子科技大学 计算机学院, 浙江 杭州 310018;

2. 浙江大学 软件学院, 浙江 杭州 310000)

摘要:汉语句群自动划分是将篇章划分成包含不同主题的文本片段,在信息提取、文摘生成、语篇理解及其他多个领域有着极为重要的应用。指代消解是识别篇章中先行词和照应词关联起来的过程,消解不同表达是自然语言理解的基础之一。针对目前的句群划分工作的重点在于划分出主题之间的边界而较少利用其本身指代关系来进行语言理解,或者因指代模糊而得到错误的划分结果的问题,提出了一种基于指代消解的句群自动划分方法。该方法从对篇章的指代情况消解出发,利用适合中文的多层过滤指代消解方法得到指代链信息,以消除不同名词代表相同实体、代词指代不明的问题。结合指代链信息,并同时考虑篇章衔接词因素,设计并进行了基于多元判别分析(Multiple Discriminate Analysis,MDA)的一组评价函数J评价句群划分验证实验。实验结果表明,所提出的方法能够有效地进行句群自动划分,统计正确分割平均P_μ提高了7%左右。

关键词:句群划分;指代消解;多层过滤;多元判别分析

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2017)08-0061-05

doi:10.3969/j.issn.1673-629X.2017.08.013

An Automatic Partition Method for Chinese Sentences Group with Coreference Resolution

WANG Rong-bo¹, SUN Xiao-xue¹, HUANG Xiao-xi¹, LIU He-ping²

(1. School of Computer, Hangzhou Dianzi University, Hangzhou 310018, China;

2. School of Software, Zhejiang University, Hangzhou 310000, China)

Abstract:Automatic Chinese sentence grouping is to divide the text into texts fragments with different theme and plays an important role in information extraction, summary generation, sentence comprehension and other fields. Coreference resolution is a procedure of recognizing antecedent and anaphora and associating them in the chapter. Resolution of the different expression is one of the basis of natural language understanding. Currently, focus of automatic Chinese sentences grouping is recognizing boundaries of different topics. Instead, the coreference relations of passage are rarely used for language comprehension, and inaccurate results are usually existed due to vagueness resolution. So an automatic Chinese sentences grouping method based on coreference resolution is proposed, which starts with resolution of the passages and get link of resolution with multi-layer filter resolution method to eliminate different terms referred to the same entity or to unknown. Besides, the cohesive markers of passages are taken into account. A group of evaluation functions are designed to evaluate sentences grouping and the experimental results show that it has improved the Chinese sentences grouping work, by which P_μ has increased about 7%.

Key words: sentences grouping; coreference resolution; multi-pass sieve; MDA

1 概述

在中文信息处理技术的发展过程中,人们发现传统的中文语法单位“词语”、“句子”能够承载的信息量太小,而“段落”、“篇章”承载的信息量又太大。根据汉语本身的意合特点,语义相关的内容通常会出现在

同一片段内,要完全理解一个句子的含义往往需要充分利用其上下文信息^[1],因而将篇章段落划分为不同的句群是篇章理解的重中之重。自然语言中还存在大量的指代现象,篇章理解的另外一个工作就是指代消解,指代消解可以有效避免“一词多义”和“多词同义”

收稿日期:2016-09-14

修回日期:2016-12-15

网络出版时间:2017-07-05

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(10YJCZH052,12YJCZH201)

作者简介:王荣波(1978-),男,副教授,CCF会员(E200017318M),研究方向为自然语言处理、篇章分析。

网络出版地址: <http://jns.cnki.net/kcms/detail/61.1450.TP.20170705.1651.062.html>

的问题。指代消解连接了指代词和先行语,明确了代词以及有歧义的名词指向,句群为其内的句子提供了可靠的上下文语境,句群划分结合指代消解在篇章分析、机器翻译、自动文摘领域有重要作用^[2-3]。

汉语句群自动划分是将篇章划分成包含不同主题的文本片段,指代消解是将篇章中的先行词和照应词关联起来的过程,消解不同表达是自然语言理解的基础之一。目前汉语句群的自动划分方法研究主要分为两种:基于规则的汉语句群划分方法和基于文本信息的句群划分方法。研究者对句群这一语法单位的相关研究比较少,也不够深入,相比较而言,他们更加注重句子、段落这种存在天然分割点的语法单位,或者是在研究句群划分时忽略了语言本身的指代结构、关联词等问题,从而得到不够准确的句群划分。

张全等^[4]根据汉语篇章句群本身的语义关联性和接应、组合规律制定了句群划分的相关规则;在概念层次网络(HNC)语境观的指导下,通过对领域句类知识的研究,阐述了一种新型的句群处理方法^[5]。韦向峰等^[6]根据 HNC 理论,认为句群领域分析是句群分析的关键,通过研究自动获取句群的领域或语境信息得到句群。但是上述基于 HNC 概念的研究工作会受到相对固定的领域知识或者判定规则的限制。

句子完整含义的理解需要有较为全面的上下文。陈怡疆等^[1]认为,如果上下文信息量太少,那么很多有用的信息就会丢失,将得不到句子全部的含义,但是如果信息量太大,又会造成搜索空间过大和数据稀疏问题,因而表示这个合适的大小不是句子或者段落,而是句群,是包含一个意义完整的主题的一组句子。他们提出了一种利用局部重现度较高的词作为特征的层次聚类算法,将篇章表示成一棵句群树,叶子节点为单个句子,内部节点就是一个多重句群,但是并未考虑篇章指代词的作用。李杰等^[7]提出一种基于多元判别分析的汉语句群自动划分方法,是一种明确可计算的模型。算法通过 Skip-Gram Model 获取句子的特征向量,与传统 VSM 相比,减少了数据稀疏,再考虑句群内部距离、句群间距离、切片片段长度和篇章衔接词等因素,设计基于 MDA 方法的评价函数 J ,通过比较 J 的值获得句群划分结果,但仅仅考虑了句首指代词。

针对现有的句群划分缺少指代消解的情况,在已有基于多元判别分析(MDA)的句群划分方法的基础上,通过引入指代消解来优化汉语句群的自动划分。基本步骤为:利用适合中文的多层过滤指代消解模型获取中文语料指代消解的结果^[8];通过 Skip-Gram Model 获取句子的特征向量;设计明确可计算的基于 MDA 的评价函数 J ,加入指代因素、考虑关联词的作用,实现对段落的划分并对所有的划分结果进行评价;

评价价值最高的句群划分序列为该段落的最佳句群划分结果。实验结果表明,加入指代消解后指代链信息提高了句群划分的效果,与传统 MDA 方法的结果对比, P_u 提升约 9%,WindowDiff 降低约 1%;与未加入指代消解的相同方法相比 P_u 提升约 7%。

2 基于指代消解的汉语句群自动划分方法

2.1 指代消解的处理

中文指代消解的研究发展较为缓慢,主流方法主要有三类:基于无监督的方法、基于有监督的方法和基于规则的层次过滤的方法。因为基于无监督的指代消解方法不依赖标注好的语料库,所以一度盛行。随着中文语料库的发展,基于有监督的指代消解方法以其较高的消解准确率取得一席之地。然而,基于有监督的指代消解方法在提取的特征向量中存在一些消解正确率较低的特征,该类特征会覆盖消解正确率较高的特征,从而影响模型的消解正确率。基于规则的层次过滤模型不需要标注好的语料库,而且模型的各个层次按照消解精度从高到低排列,不会出现消解正确率低的特征覆盖消解正确率高的特征的现象,因此该方法会获得更好的消解效果,也比较适合中文的指代消解^[9]。

按照基于规则的层次过滤指代消解的思想,该模块的系统框架分为三部分:预处理、待消解项识别、指代消解处理^[10-11]。

(1)预处理:对语料进行分词,词性标注,命名实体识别和句法分析,句法分析结果由 Stanford Parser 处理得到。根据相应的语言学规则从句法分析结果抽取出候选待消解项,包括名词、名词短语和代词。

(2)待消解项识别:待消解项识别的精度对整个指代消解模型的精度产生了极大影响,并且丢失待消解项比错分指代链更影响消解模型的精度。待消解项识别分为两部分:扩充阶段,提取所有的名词和名词短语,尽量保证不会丢失待消解项;过滤阶段,去除一些无需消解的停用词,没有意义的时间,数词,金钱等词汇,过滤重复词,在保证一定召回率的同时,提高待消解项识别的正确率^[12]。

(3)指代消解处理:字符串完全匹配,别名匹配和同位语对名词短语的指代消解贡献达到了 97%^[13],而代词指代消解是篇章指代消解的一个关键。因此,设置四个层次,将各个过滤层次按照消解正确率从高到低排列,名词短语和代词通过层次过滤寻找其先行语。各个层次过滤模块如表 1 所示。

●完全字符串匹配。

若两个字符串完全相同,则认为这两个名词短语指向同一个实体。该层的准确率最高。

表 1 指代消解各层过滤模块

层次	处理数据类型	特征
1	名词、名词短语	全匹配特征
2	名词、名词短语	别名特征
3	名词、名词短语	同位语特征
4	名词、名词短语、代词	代词

●别名匹配。

若一个字符串是另外一个的子串或抽取子串,则说明它们之间有别名关系,是指向的同一个实体。例如:“普京”是“弗拉基米尔·弗拉基米罗维奇·普京”的子串,“中国”是“中华人民共和国”的抽取子串。

●同位语。

若两个短语之间有同位语关系,则说明他们指向相同。同位语的定义是一个名词(或其他形式)对另一个名词或代词进行解释或补充说明,这个名词(或其他形式)就是同位语。

●代词匹配层。

代词指代是指代的重点和难点。这层是解决代词和名词或名词短语之间是否具有指代关系,主要通过判断单复数匹配关系、性别是否一致、有无生命,还有根据命名实体结果分为组织、地点、人名、杂项等的匹配。

基于指代消解的汉语句群划分方法整体框架如图 1 所示。

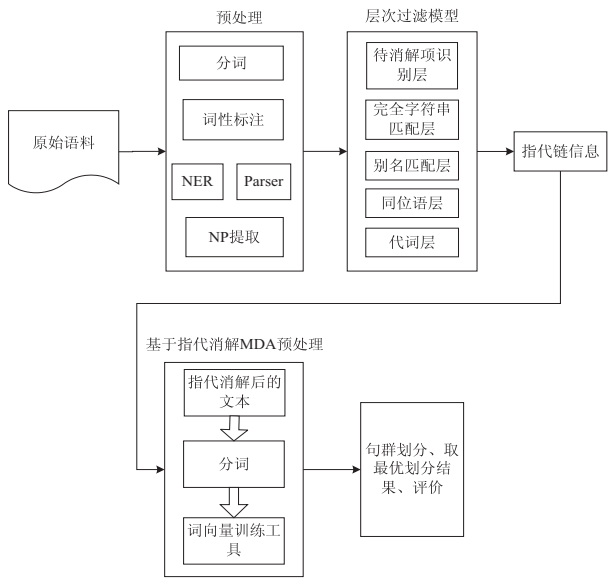


图 1 基于指代消解的汉语句群划分方法整体框架图

2.2 句群划分模型

句群,顾名思义就是若干句子的组合,它们描述同一个中心,意义完整,句子的组合有一定的逻辑顺序^[14]。句群划分主要依据语言本身的特点和组合规律。句群划分实例如图 2 所示。句群 1 中的“它”是一个指代词,指代白杨树,通过指代关系的确认可以很好

地消解词语的二义性,对以后衡量类内距离有重要作用;句群 2 揭示了其组合规律,用“难道”开头的四个反问句表达了对北方军民的赞颂,是一种递进关系,第④句中存在衔接词“但是”,代表转折关系,如果切分出来必然不合理,需要对这种切分结果进行惩罚。

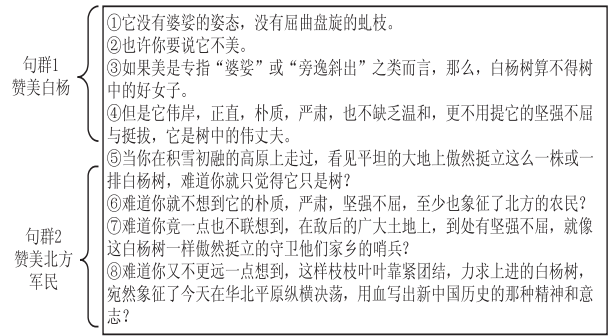


图 2 句群划分实例

根据汉语表达习惯,一个句子可以独立地表达一个完整的意思,相似的内容一般出现在同一片段内,段落是一个意义完整性的天然分割点。但一个段落中可能包含不同的主题,所以句群的划分以句子为基本单位进行,在一个段落中划分出不同主题的句子群。

MDA 是一种独立于具体领域的文本线性分割统计模型方法,可以通过定义评价函数实现对句群划分的全局评价^[15]。具体是对句子向量构成的数据空间进行划分,考虑句群内部距离、句群之间距离、切分片段长度、指代因素以及篇章衔接词因素,设计基于 MDA 的评价函数 J ,使函数 J 值取得最大的划分即为最优划分结果。

设最优划分结果为 D ,则:

$$D = \operatorname{argmax}_D P(D | T) = \operatorname{def} \operatorname{argmax}_D J(D, S_w, S_b, S_L, S_d, S_e) \quad (1)$$

其中, S_w 为类内离散矩阵; S_b 为类间离散矩阵; S_L 为切分片段长度惩罚因子; S_e 为指代因子; S_d 为篇章衔接词惩罚因子。

(1)句群内部距离与句群间距离。

句群内部的紧凑性和句群间的离散性是重要特点。类内离散矩阵可用于衡量句群内部的内聚程度。

(2)切分片段长度因素。

当划分模式切分出连续的单句时,需要对结果进行惩罚。

(3)指代因素。

消除代词的指代不明和实体的不同名词短语表达问题是计算机理解自然语言的基础。这里将指代消解后的指代链信息加入评价函数 J 。

(4)篇章衔接词因素。

句子之间在表达形式上也会显示出其连贯性。建立篇章衔接词表 Dict,包含“而”、“并且”等词。

3 实验测试

3.1 实验语料与测评

(1)语料设置。

目前还没有一个公开、通用的中文句群划分评测语料,为了验证指代消解对句群划分的影响,取与文献[7]相同的实验语料—《读书》杂志(1979-1983),共50期,人工标注了其划分结果,分割片段的平均句子数为3,段落的平均句子数为9,文献作者通过计算Kappa值说明了语料的相对一致可靠性。

首先对原语料进行指代消解处理,得到指代链信息,对位于同一指代链上的名词、名词短语或者代词进行一定规则的替换。之后进行句群自动划分的处理,分词后使用词向量训练工具 word2vec 获取词语在低维空间中的向量表示,再对形成的数据空间进行划分,通过评价函数 J 得到最优划分结果。

(2)测评指标。

传统的评价方式(准确率和召回率)主要是考虑绝对匹配的情况,而在句群划分中,这一评价方式不再适合。为此,采用文本分割中常用的 P_u ^[16] 和 WindowDiff^[17] 评价方法。

P_u 通过计算任意两个句子是否被算法正确划分为同一片段的概率,分割点距离正确的分割点越近, P_u 评价价值越高。计算公式如下:

$$P_u(\text{ref}, \text{hyp}) = \sum_{1 \leq i \leq j \leq N} \gamma_u e^{-\mu |i-j|} (\delta_{\text{ref}}(i, j) \oplus \delta_{\text{hyp}}(i, j))$$

(2)

其中, hyp 为机器自动划分模式;ref 为人工划分模式; N 为划分段落中的句子总数。人工划分模式下,当第 i 个句子和第 j 个句子同属一个句群时, $\delta_{\text{ref}}(i, j)$ 为1,否则为0;类似,算法自动划分模式下,当第 i 个句子和第 j 个句子同属一个句群时, $\delta_{\text{hyp}}(i, j)$ 为1,反之则为0。 \oplus 为异或符号; γ_u 取 $1/\sum e^{-\mu |i-j|}$, 随着第 i 个句子和第 j 个句子间的距离得到不同的权值。

WindowDiff 对不正确的分割点做出惩罚,即“正错误”和“负错误”。“正错误”是指在实验中多做了分割,“负错误”是指在实验中遗漏了分割。WindowDiff 值越小,说明分割结果越好。计算公式如下:

$$\text{Window Diff}(\text{ref}, \text{hyp}) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(\text{ref}_i, \text{ref}_{i+k}) - b(\text{hyp}_i, \text{hyp}_{i+k})| > 0)$$

(3)

其中, $b(i, j)$ 为相应划分模式下位置 i 和位置 j 直接的切分点的数量; k 为平均切分片段句子数的 $1/2$ 。

3.2 实验结果及分析

(1)实验结果。

指代消解性能见表2。其中, P (正确率)= 正确识别的个体数/识别出的个体总数; R (召回率)= 正确

识别的个体总数/测试集中存在的个体总数; F = 准确率 * 召回率 * 2/(准确率+召回率)。

表 2 基于层次过滤的指代消解性能

评价指标	数值
P	64.3
R	71.5
F	67.7

表3展示了对文本进行指代消解后的句群划分在不同维度下评价函数 J 的实验结果,统计正确分割的平均 P_u 值为91.26%,统计错误分割的平均 WindowDiff 值为27.26%,从100~300维, P_u 值略有提升、WindowDiff 值下降,而在400维, P_u 下降、WindowDiff 上升。

表 3 不同维度下评价函数 J 的实验结果 %

评价指标	维数			
	100	200	300	400
P_u	87.96	92.68	93.66	90.79
WindowDiff	27.56	27.34	27.12	27.40

表4展示了加入指代消解和未加入指代消解的基于MDA的汉语句群自动划分方法的比较结果, P_u 提升约7%, WindowDiff 提升约2%。

表 4 加入和未加入指代消解的基于 MDA 的汉语句群自动划分方法对比 %

评价指标	文中方法	原基于 MDA 的句群划分方法
P_u	91.26	84.18
WindowDiff	27.36	24.93

表5展示了文中方法与传统MDA方法的结果对比, P_u 提升9%, WindowDiff 降低1%。其中传统MDA方法的评价函数 J 通过衡量类内离散矩阵、类间离散矩阵和切分片段长度得到。实验结果表明,指代因素 S_c 和篇章衔接词因素 S_d 起到了一定的作用。

表 5 文中方法与传统 MDA 方法的比较 %

评价指标	文中方法	传统 MDA(J')方法
P_u	91.26	81.74
WindowDiff	27.36	28.57

(2)实验分析。

加入指代消解后,显著提高了句群划分的效果,统计平均正确分割 P_u 有一定程度的提升,统计错误的平均分割 WindowDiff 有所下降。对句群划分加入指代消解的处理消除了代词指代不明、不同名字实则相同实体的情况,是篇章理解的重要因素,在后续衡量句群内部的紧凑性和句群之间的离散性中发挥了重要作用。汉语篇章表述中,代词指代是文本中数量较多的指代形式,而另外三种指代形式则出现较少,所以代词

指代对句群划分的贡献度最大,而因为完全字符串匹配、别名匹配、同位语匹配这三层准确率达到 97% 左右,因此也很好地涵盖了其他形式的指代情况。

通过 Skip-Gram Model 训练大规模语料获取词语在低维实数空间向量表示,通过挖掘深层语义信息获取文本表面的联系,通过表 3 说明并不是维度越高越好, P_0 值与维度并不是线性关系。

由表 4 知,加入指代消解较未加入指代消解的 P_0 值提升明显,说明加入指代消解后划分句群的算法得到的切割点较接近实际的切割点,而 WindowDiff 值也较未加入指代消解的大,WindowDiff 是对“正错误”和“负错误”的衡量,说明分割算法在这方面是有缺陷的。

4 结束语

为了在篇章理解的基础上优化汉语句群自动划分,提出一种基于指代消解的句群自动划分方法。该方法在 MDA 句群划分法的基础上,从语料名词、名词短语、代词的指代消解出发,进而实现汉语句群的自动划分。基于该方法构建了自动划分系统,并实现了基于指代消解的句群划分。实验结果表明,与传统 MDA 方法对比, P_0 提升约 9%,WindowDiff 降低约 1%;与未加入指代消解进行对比, P_0 提升约 7%。表明该方法有效可行。

参考文献:

[1] 陈怡疆,史晓东,周昌乐. Automatic partition of Chinese sentence group[J]. Journal of Donghua University: English Edition, 2010, 27(2): 177-180.

[2] 刘福君. 基于指代消解的自动文摘研究[D]. 合肥:安徽大学, 2012.

[3] 石晶. 文本分割综述[J]. 计算机工程与应用, 2006, 42(35): 155-159.

[4] 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.

[5] 韦向峰,缪建明,张全,等. 基于概念基元的句群情景框架抽取研究[J]. 微计算机应用, 2010, 31(4): 21-24.

[6] 韦向峰,缪建明,张全. 汉语句群领域的自动抽取研究[J]. 计算机工程与应用, 2009, 45(4): 11-15.

[7] 王荣波,李杰,黄孝喜,等. 基于多元判别分析的汉语句群自动划分方法[J]. 计算机应用, 2015, 35(5): 1314-1319.

[8] 周炫余,刘娟,卢笑. 篇章中指代消解研究综述[J]. 武汉大学学报:理学版, 2014, 60(1): 24-36.

[9] 周炫余,刘娟,罗飞,等. 中文指代消解模型的对比研究[J]. 计算机科学, 2016, 43(2): 31-34.

[10] Raghunathan K, Lee H, Rangarajan S, et al. A multi-pass sieve for coreference resolution[C]//Conference on empirical methods in natural language processing. Mit Stata Center, Massachusetts, USA: A Meeting of Sigdat, A Special Interest Group of the ACL, 2010: 492-501.

[11] Lee H, Peirsman Y, Chang A, et al. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task[C]//Proceedings of the fifteenth conference on computational natural language learning: shared task. [s. l.]: Association for Computational Linguistics, 2011: 28-34.

[12] 孔芳,朱巧明,周国栋. 中英文指代消解中待消解项识别的研究[J]. 计算机研究与发展, 2012, 49(5): 1072-1085.

[13] 高俊伟,孔芳,朱巧明,等. 基于 SVM 的中文名词短语指代消解研究[J]. 计算机科学, 2012, 39(10): 231-234.

[14] 梅汉成. 现代汉语句群研究概述[J]. 盐城师范学院学报:人文社会科学版, 1996(3): 35-37.

[15] 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.

[16] Beeferman D, Berger A, Lafferty J. Statistical models for text segmentation[J]. Machine Learning, 1999, 34(1-3): 177-210.

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

(上接第 60 页)

lin; Springer-Verlag, 1998: 105-119.

[7] Grahne G, Zhu J. High performance mining of maximal frequent itemset[EB/OL]. [2014-07-06]. <http://www.docin.com/p-773109811.html>.

[8] 路松峰,卢正鼎. 快速开采最大频繁项目集[J]. 软件学报, 2001, 12(2): 293-297.

[9] 宋余庆,朱玉全,孙志挥,等. 基于 FP-tree 的最大频繁项目集挖掘及更新算法[J]. 软件学报, 2003, 14(9): 1586-1592.

[10] 吉根林,杨明,宋余庆,等. 最大频繁项目集的快速更新

[J]. 计算机学报, 2005, 28(1): 128-135.

[11] 钱雪忠,惠亮. 关联规则中基于降维的最大频繁模式挖掘算法[J]. 计算机应用, 2011, 31(5): 1339-1344.

[12] 杨鹏坤,彭慧,周晓锋,等. 改进的基于频繁模式树的最大频繁项集挖掘算法—FP-MFIA[J]. 计算机应用, 2015, 35(3): 775-778.

[13] Tan Pangning. 数据挖掘导论:英文[M]. 北京:人民邮电出版社, 2006.

[14] 秦亮曦,史忠植. SFP-Max—基于排序 FP-树的最大频繁模式挖掘算法[J]. 计算机研究与发展, 2005, 42(2): 217-223.