

基于基本要素的文摘内容连贯性评测模型

刘德喜^{1),2)} 姬东鸿³⁾

¹⁾(江西财经大学信息管理学院 南昌 330013)

²⁾(江西财经大学数据与知识工程江西省高校重点实验室 南昌 330013)

³⁾(武汉大学计算机学院 武汉 430079)

摘 要 文摘的自动化面临诸多困难,一个重要的原因是文摘的内容缺乏有效的自动评测方法.文中提出了基于基本要素(BE)关系网格的文摘内容连贯性评测模型.模型以 BE 为内容单元,以 BE 中的“关系”为内容单元的语法角色,通过 BE 关系在 BE 关系网格中的转移概率来表达文摘内容的连贯性.在 DUC2005 数据集上的评测结果显示,模型评测结果与人工评测结果的 Pearson 相关系数为 0.408,比 Lapata 2005 年提出的实体网格模型得到的结果提高了约 66%.

关键词 自动文摘;内容连贯性;基本要素;关系网格

中图法分类号 TP18

Evaluation Model of Summary Coherence Based on Basic Element

LIU De-Xi^{1),2)} JI Dong-Hong³⁾

¹⁾(School of Information Management, Jiangxi University of Finance & Economics, Nanchang 330013)

²⁾(Jiangxi Key Laboratory of Data and Knowledge Engineering, Jiangxi University of Finance and Economics, Nanchang 330013)

³⁾(School of Computer, Wuhan University, Wuhan 430079)

Abstract One of the key problems in automatic summarization is the absence of effective auto-evaluation method. A BE-Relation-Grid based evaluation model for "summary coherence" is proposed. BE(Basic Element) is viewed as the content unit and the "relation" part in BE as grammar role of the content unit. Then, the content coherence is scaled by BE relation transition probability in the BE-Relation-Grid. Experiment results on DUC2005 dataset show that the Pearson correlation coefficient of the evaluation results between this model and manual ones is 0.408, which increased by about 66% comparing with the result of entity grid model presented by Lapata in 2005.

Keywords automatic summarization; content coherence; basic element; relation grid

1 引 言

随着网络的日益普及,在线信息急剧增加,如何有效地获取和描述这些文本信息显得越来越重要.尽管用户通过搜索引擎可以快速获得丰富的文档,

但要获取其中内容则需要消耗大量时间去阅读每一篇文档.自动文本文摘(Automatic Text Summarization)能够为用户提供一个原文档的压缩版本,旨在减轻用户的阅读压力,是 NLP(Natural Language Processing)的一个研究热点.

文摘的自动化面临诸多困难,一个主要的原因

是对文摘的内容缺乏有效的评测标准. 当前评测文摘内容最主要的方法是考察文摘的内容覆盖率, 这需要有理想文摘作为参照. 然而, 不同的人由于兴趣和背景的差异, 对相同的文本做摘要, 得到的摘要内容也各不相同, 因此理想文摘又很难统一. 事实上, 文摘评测的困难以及如何解决的建议一直伴随着自动文摘的整个发展进程^[1-5]. 另一方面, 由于文摘是对原文档内容的浓缩, 难免对原文档进行内容上的删减, 这种删减后的短文在可读性方面会受到很大的影响. 然而, 目前对可读性的评测主要由人工完成, 由于需要消耗大量的人力物力, 使得人工评测方法不能普及. 同时, 这种人工评测的结果会受到评价员个人素质的影响, 缺乏客观性. 如果能对这种可读性进行自动评测, 对促进自动文摘的发展会大有帮助.

文摘的评测主要包括两个方面, 一是文摘内容覆盖率的评测, 即文摘内容覆盖了多少原文档的中心内容, 或覆盖了多少用户所希望获取的内容. 另一方面的评测是文摘内容的可读性, 即文摘内容是否流畅. 按文档理解会议 DUC2005^[6] 的评测标准, 文摘内容的可读性包括五个方面: 语法、无冗余、指代清楚、聚焦、结构与连贯性. 本文主要研究文摘内容连贯性的自动评测模型, 其结构安排如下: 第 2 节介绍对文摘自动评测的相关研究及开展本研究工作的原因, 并简要介绍基本要素 BE 的概念; 第 3 节讨论基于 BE 关系网格的文摘内容连贯性自动评测模型; 第 4 节以 DUC2005 中的理想文摘为训练集、机器文摘为测试集, 对模型的性能进行评测; 最后是结论与展望.

2 相关研究

目前文摘的自动评测方法主要是判断机器文摘对标准文摘的信息覆盖程度, 是对文摘内容完整性的一种评测. Donaway^[5] 对基于抽取的文摘提出了基于句子排序和基于内容的评测方法. Saggion^[7] 等提出了三种基于文摘内容相似度的自动评测方法, 分别是基于余弦相似度、基于单元覆盖和基于最长公共子串的方法. Lin 和 Hovy^[8] 给出了基于 n -gram 同现统计的方法 ROUGE, 于 2004 年, Lin^[9] 又给出了基于最长公共子串、基于指定距离的句子内词对的共现统计等方法, 并证明该评测方法与人工评测具有很好的一致性. 由于子串长度的确定、词对间距离的确定比较武断, Hovy^[10-11] 提出通过统计理想文

摘与机器文摘间 BE 共现率的方法. 因为 BE 的获取是以句法分析为基础, 这种方法比武断确定待比较串的长度或词对的距离更值得信赖, 基于 BE 的评测已成功应用于 DUC2005 和 DUC2006. 鉴于理想文摘可能有多个, Nenkova^[12] 提出了一种在手工构建金字塔型文摘内容单元 (SCU) 的基础上, 对文摘内容进行评测的方法, 该方法已在 DUC2005、DUC2006 中得到检验. 对评测方法的评价, 目前基本上都采用自动评测结果与人工评测结果的相关性来衡量, 相关性高, 则说明该评测方法较好.

国内而言, 与对自动文摘系统的研究相比, 对文摘自动评测方面的研究工作开展得相对较少. 1995 年北京大学计算语言学研究所承担了“八六三”计划智能计算机专家组办公室下达的任务, 对三个中文自动文摘系统进行评测. 在此次评测实践的基础上, 俞士汶教授^[13] 进一步考虑建立机械文摘自动评测系统, 并在“八六三”计划资助项目“机器翻译与机器文摘的自动评价”的支持下, 建立了一个机械文摘质量自动评测模型系统及专家文摘辅助写作系统, 并进行了自动评测的实验. 这种评测方法主要针对基于句子抽取的文摘系统, 通过机器文摘与专家文摘的句子匹配率来反映机器文摘系统的性能. 上海交通大学的沈洲^[14] 实践了一种参照 Turing 测试的思想进行自动文摘系统评价的方法, 该方法是一种人工评测方法. 张姝^[15] 提出利用文本余弦相似度评价自动文摘系统的方法, 研究了不同权重选取策略对评测结果的影响, 并且用 DUC 测试集对评价方法进行了测试. 华中师范大学胡珀提出一种基于代表熵的不依赖理想文摘的评测方法^[16]. 总的来看, 对中文自动文摘系统的自动评测还很不成熟, 以至于在 2003 年度“八六三”计划中文自动文摘的评测中仍然用人工的方式.

对自动文摘评测的研究大都集中在内容的召回率和文摘的语法上^[8, 17-18], 然而, 对文摘的结构, 也就是其连贯性方面, 目前的评测还是采用人工方式^[6]. 在文档内容连贯性自动评价方面, 学者们也做了大量的工作. Higgins 等^[19] 设计了一个用于评测学生论文整体连贯性的系统, 该系统通过手工标注训练语料, 用于学习哪些类型的片段会导致文档不连贯. 其它的方法主要集中在描述局部连贯性方面. Miltasakaki 和 Kukich^[20] 对学生论文通过手工标注其中的实体转移信息, 得出结论: 实体转移类型的分布与人工对连贯性的评分结果相关. Foltz 等^[21] 提出不需要手工操作的局部连贯性模型: 如果相邻句

子的意义有较程度的重复,则相应的文本内容可以认为是连贯的.于是,通过构建词汇向量空间来计算相邻句子在向量空间的距离,并将其作为相邻句间的语义相关性. Foltz 指出,这种模型与人工判断有很好的相关性,可以用于分析文档的话题结构.

上述工作大都需要人工对文档进行标注,并且只用到了语法或语义其中的一种.文献[22]提到的方法自动对文档进行浅层分析,该模型结合语法和语义两种信息.首先通过对文档自身进行浅层分析,获得其中的名词实体,再计算各实体所充当的角色在句子间的转移概率,得到文档语法上的连贯性.对于语义的连贯性,该文献用句子间的相似性来定量描述语义连贯性.然而,上述方法在考察语法连贯性时,只考察了名词实体及其所充当的两种语法角色:主语(subj)或宾语(obj).然而,除了名词实体外,还有其它很多对语法连贯性有强烈影响的实词,特别是动词,并且这些实体在句中的角色也远不止主语或宾语两种.

基本要素^[10] (Basic Element, BE)描述的是基本要素中心词(head)及其修饰(modifier)之间的关系(relation),表示为一个三元组“中心-修饰-关系”(head|modifier|relation),其中“中心词”表示主要的语法要素,它通常是名词、动词、形容词或副词短语.图1是句子“The United Nations imposed sanctions on Libya in 1992 because of their refusal to surrender the suspects”中基本要素的一个例子.

head	modifier	relation	
imposed	united nations	subj	(BE-F)
imposed	sanctions	obj	(BE-F)
sanctions	libya	on	(BE-F)
libya	1992	in	(BE-F)
refusal	their	gen	(BE-F)
libya	refusal	because of	(BE-F)
refusal	surrende	compl	(BE-F)
surrender	united nations	subj	(BE-F)
surrender	suspects	obj	(BE-F)

图 1 BE 的一个例子

BE 旨在表示文本中高信息量的一元(1-gram)、二元(2-gram)或通过组合得到的更长的内容单元(n -gram).由于 BE 是依据句子的结构信息得到的,所以这个三元组中不仅有语义(词)的信息,还含有句子的结构信息.一个更重要的方面是,这些单元可以自动构造而不需要人工进行.

BE 可由多种方法获得,通常先采用语法剖析器生成语法树,再利用一些裁剪规则从语法树中提取

合法的 BE.本文采用的是南加州大学发布的 BE 包^[11].

BE 是根据对句子的语法分析得到的句子基本要素及其间的关系,能够更精确地把握句子的结构. BE 中头和修饰本身可视为其语义,而其关系因为来源于语法分析,又是句子语法的体现,通过考察 BE 在句子间的转换情况可兼顾语法和语义,所以,本文利用句子间 BE 的转移概率来描述文摘的连贯性.

3 文摘内容连贯性评测模型

文摘是由句子构成的,因此其整体的连贯性也可通过局部(句子间)的连贯性来体现.本文的主要焦点仍放在文本的局部连贯性上,也就是考察相邻句子间的连贯性.中心理论 CT (Centering Theory^[23-24])对局部内容连贯性的建模影响重大. CT 的一个基础就是对实体的引入和讨论,该理论认为,谈论相同实体的相邻片段(如句子)比谈论不同实体的相邻片段具有更高的连贯性.因此,连贯性分析就可利用实体在片段间的转移模式来进行分析.本文提出的连贯性模型就是基于这样一个假设:在连贯的话题中,实体是按一定的模式在相邻片段间转移的. BE 不仅描述有句子中的实体,同时也给出了修饰关系,因此,选择 BE 作为话题的代表是比较合理的.为消除公用词的影响,在构建 BE 集合时,利用 WordNet 的公用词表,将原始 BE 中的公用词去掉.

由于 BE 是由三部分组成:中心词(BEH)、修饰(BEM)及二者之间的关系(R),在计算句子间的转移概率时,需要确定内容单元的粒度.为对比粒度选择对模型的影响,分别以 BE, BEH, BEHM (BEH 与 BEM 的并集)为内容单元.此外,由于主要考虑句中的实体,所以需要对 BE 进行选择,选择的条件由 BEH 与 BEM 之间的关系来确定. BE 中心词与修饰间存在多种关系,然而,中心理论认为,话题中的中心实体通常在语法位置上充当主语或宾语.为比较关系的选择对模型的影响,分别用以下 3 种方法来选择 BE.

(R1) 只保留关系为“subj (subject)”、“obj (object)”的 BE;

(R2) 保留关系为“subj”、“obj”、“conj”, “nn”的 BE. 保留关系“conj (conjunction)”的原因在于,对于多个宾语或主语的情况,其间的关系是用“conj”描述的;保留关系“nn (noun compound)”的原因在于,它表示的是名词词组,可能与其它名词一起充当施

事或受事。

(R3) 进一步扩大保留的范围,保留关系为“subj”、“obj”、“conj”、“nn”、“obj2”、“abbrev”、“mod”的 BE. 这些 BE 基本包含了文摘内容中的绝大部分实体。

考虑到有些 BE 包含多个词,而在后面的句子中对其进行缩写,但其代表的实体对象是相同的,如“falkland islands”与“falkland”,一种简单的办法是将相应 BE 拆成多个词进行处理。

将整个文档表示为一个网格,称其为 BE 网格,其中的列对应于内容单元(BE, BEH, 或 BEHM),

行对应于句子. 网格的列代表一个内容单元出现或不出现在句子序列(s_1, \dots, s_n)中. 其详细描述如下: 网格中的单元 $r_{i,j}$ 代表第 i 个内容单元在第 j 个句子 s_j 中呈现的关系(relation). 如 subject(subj)、object(obj)、conjunction(conj)、noun compound(nn)、其它关系(用“x”表示)、该内容单元在本句未出现(用“-”表示)等. 例如,对于图 2(a)中的文档(来自 DUC2005 数据集中的文件“D324.M.250.E.C”),通过 BE 分析得到图 2(b). 以 BEHM 为内容单元,并用第 2 种方法(R2)选择 BE,去掉共用词并构建 BE 网格如图 2(c).

Since the 1982 war over the Falkland Islands, relations between Argentina and Britain have steadily improved. Full diplomatic relations resumed in 1990, and senior British and Argentine officials have visited in London and Buenos Aires...

(a) BE 网格的例子:原文档内容

----- sentence 1 -----	----- sentence 2 -----
war 1982 num	relations full mod
war falkland islands over	relations diplomatic mod
falkland islands relations conj	resumed relations subj
relations argentina between	resumed 1990 in
relations britain conj	british senior mod
improved steadily amod	officials argentine mod
since improved compl	british officials conj
improved war subj	visited british subj
	visited london in
	london buenos aires conj

(b) BE 网格的例子:BE 分析结果

	falkland islands	relation	war	british	london	prince	visit	...
s_1	conj	conj, x	subj, x	—	—	—	—	
s_2	—	subj, x	—	conj, subj, x	conj, x	—	—	

(c) BE 网格的例子:BE 网格

图 2

相邻句子间的连贯性可用句子所含 BE 间的关系转换与某一特定模式相符程度来衡量,本文用概率模型描述这种转换. 对于包含基本要素 $\beta_1, \beta_2, \dots, \beta_m$ (m 为文摘 S 中经过筛选得到的 BE 总数)的文本 $S(s_1, s_2, \dots, s_n)$ (n 为文摘中句子总数),其连贯性定义为 BE 在句子中的联合概率分布:

$$P_{coh} = P(\beta_1, \beta_2, \dots, \beta_m; s_1, s_2, \dots, s_n) \quad (1)$$

为简化问题,假设各 BE 的使用与文档中的其它 BE 无关,则上式可简化为

$$P_{coh} = \prod_{i=1}^m P(\beta_i; s_1, s_2, \dots, s_n) \quad (2)$$

对于每一个基本要素 β_j ,其在文本中各句间分布可得用 BE 关系网格进行计算:

$$\begin{aligned} P(\beta_j; s_1, s_2, \dots, s_n) &= P(r_{1,j} \cdots r_{n,j}) \\ &= P(r_{1,j})P(r_{2,j} | r_{1,j})P(r_{3,j} | r_{1,j}r_{2,j}) \cdots \end{aligned}$$

$$\begin{aligned} P(r_{n,j} | r_{1,j} \cdots r_{n-1,j}) \\ = \prod_{i=1}^n P(r_{i,j} | r_{1,j} \cdots r_{(i-1),j}) \end{aligned} \quad (3)$$

其中, $r_{i,j}$ 表示第 i 个基本要素 β_i 在第 j 个句子 s_j 中呈现在关系.

假设本句的语法只与其相邻的上一句的语法相关,而与更前面的句子无关,则式(3)可简化为

$$P(\beta_j; s_1, s_2, \dots, s_n) = \prod_{i=1}^n P(r_{i,j} | r_{(i-1),j}) \quad (4)$$

例如对于图 2,有 $P(\text{relation}; s_1, s_2, \dots, s_4) = P(\text{conj})P(\text{subj} | \text{conj})P(- | \text{subj})P(\text{subj} | -)$. 此处需要进一步解释的是,实体“british”在第 2 句中存在两种关系“conj”和“subj”,由于充当主语显得更为重要,所以此处强制选择“subj”. 关系的优先顺序依次为“subj”>“obj”>“nn”>“conj”>“x”.

为了对不同长度和不同 BE 个数的文本的连贯

性进行比较,需对上式所得的结果进行标准化.同时,为避免小概率对整个结果的影响,我们对结果取对数

$$P_{coh} = \frac{1}{m \cdot n} \sum_{j=1}^m \sum_{i=1}^n \log P(r_{i,j} | r_{(i-1),j}) \quad (5)$$

其中, m 表示整个 BE 关系网格中 BE 的数量, n 表示文摘中句子的个数.

通过在理想文摘集上进行训练,可以得到各种关系间的转移概率,于是可计算出文档的连贯性.

4 实验结果

在对一篇文摘进行连贯性分析之前,需要知道 BE 关系网格中各关系间的转移概率. 在 DUC2005 的数据集中,不仅有机器文摘,还有供参考的理想文摘,而理想文摘是通过人工生成的,具有很好的连贯性,可以作为 BE 关系转移概率的训练数据集. 另外,DUC 数据中还包括对机器文摘连贯性的人工评测结果,可以为文摘连贯性的自动评测结果提供参照. 因此,对连贯性自动评测模型的实验,就建立在

DUC2005 的数据集上.

4.1 转移概率估计

对于 DUC2005 中的所有理想文摘(共计 311 篇)中的相邻句子,首先进行 BE 分析,得到本句所含的所有 BE 三元组. 转移概率的估计采用极大似然估计. 设 β_{pre}, β_{suc} 分别为文摘 S 中相邻句子 s_{pre}, s_{suc} 中的 BE,则关系 $r_j \rightarrow r_i$ 的转移概率定义为“由 r_j 转移到 r_i 的次数与由 r_j 转移到所有关系的总次数之比:

$$P(r_i | r_j) = \frac{P(r_i, r_j)}{P(r_j)} = \frac{\sum_S \sum_{r_i \in \beta_{suc}, r_j \in \beta_{pre}} (r_j, r_i)}{\sum_S \sum_{r_j \in \beta_{pre}} r_j} \quad (6)$$

由于 BE 关系网格所采用的关系非常有限,因此训练数据集中的数据稀疏问题并不严重. 但考虑到有些关系的转移在训练语料中可能缺失,采用加 1 法进行平滑,即在利用上式进行统计前,所有关系间的转移次数被初始化为 1. 表 1 是在 DUC2005 理想文摘上训练的结果.

表 1 通过 DUC2005 数据集训练得到的转移概率

	—	conj	obj	nn	x	subj
—	0.000028	0.095859	0.147613	0.091099	0.516825	0.148576
conj	0.996336	0.002537	0.000282	0.000282	0.000282	0.000282
obj	0.988344	0.000188	0.009964	0.000564	0.000564	0.000376
nn	0.986111	0.000296	0.000887	0.011525	0.000296	0.000887
x	0.991979	0.000157	0.000315	0.000524	0.006762	0.000262
subj	0.991185	0.000188	0.000938	0.000563	0.000563	0.006564

表 1 中“—”到“—”的转换次数本该为 0,但为避免数据稀疏所引起的某些转移概率为 0,采用对每种转移的转移次数初值置 1 的方法后,“—”到“—”也获得了一次转移.

4.2 文摘连贯性模型评价

选择 DUC2005 中的机器文摘作为测试数据,共 32 个参赛系统,每个参赛系统为每个原文档簇生成一篇文摘,共计 32×50 篇机器文摘,从中随机选取 100 篇作为测试集. DUC 评测中,除了通过机器

自动评测文摘的信息含量外,还包括手工对文摘质量的各项指标进行评价. 文摘“结构和连贯性”指标的手工评测共分 5 个级别,分值越高,说明连贯性越好. 表 2 是文摘连贯性模型的评测结果与手工评测结果之间的 Pearson 相关性,其中 Hum 为人工评测、模型命名规则为“内容单元+BE 选择方法”,如 H1 是以 BEH 为内容单元并通过方法 R1 选择 BE、HM2 以 BEHM 为内容单元并通过方法 R2 选择 BE、BE3 以 BE 为内容单元并通过方法 R3 选择 BE.

表 2 连贯性模型评测结果与人工评测结果的 Pearson 相关性 (** $p < 0.01$ (2-tailed); * $p < 0.05$ (2-tailed), $N=100$)

		Hum	H1	HM1	BE1	H2	HM2	BE2	H3	HM3
H1	Pearson Correlation	.305**								
	Sig. 2-tailed	.002								
HM1	Pearson Correlation	.271**	.883**							
	Sig. 2-tailed	.006	.000							
BE1	Pearson Correlation	.308**	.584**	.667**						
	Sig. 2-tailed	.002	.000	.000						
H2	Pearson Correlation	.344**	.947**	.862**	.545**					
	Sig. 2-tailed	.000	.000	.000	.000					

		(续 表)								
		Hum	H1	HM1	BE1	H2	HM2	BE2	H3	HM3
HM2	Pearson Correlation	.320**	.782**	.926**	.575**	.881**				
	Sig. 2-tailed	.001	.000	.000	.000	.000				
BE2	Pearson Correlation	.408**	.451**	.514**	.692**	.640**	.699**			
	Sig. 2-tailed	.000	.000	.000	.000	.000	.000			
H3	Pearson Correlation	.265**	.743**	.913**	.533**	.823**	.968**	.608**		
	Sig. 2-tailed	.008	.000	.000	.000	.000	.000	.000		
HM3	Pearson Correlation	.273**	.754**	.910**	.537**	.838**	.970**	.625**	.998**	
	Sig. 2-tailed	.006	.000	.000	.000	.000	.000	.000	.000	
BE3	Pearson Correlation	.220*	.788**	.936**	.475**	.839**	.955**	.505**	.983**	.979**
	Sig. 2-tailed	.028	.000	.000	.000	.000	.000	.000	.000	.000

文献[25]考察了 DUC2002 中不同人对相同文摘所做的评测结果,发现人工评测结果间也有不小的差异.通过“留一重采样法(leave-one-out resampling)”,得到人工评测结果之间的相关性平均值为 0.768,这可以作为机器评测结果与人工评测结果相关性的上限.由于 DUC2005 数据集中,每个机器文

摘只有一个手工评测结果,不能在此数据集上计算人工评测间的相关性,因此,我们也把 0.768 作为上限.

作为参照,此处给出文献[25]提出的“实体关系网格模型”的评测结果,如表 3.

表 3 文献[25]的结果

	Humans	Egrid	Overlap	LSA	HStO	Lesk	JCon	Lin
Egrid	.246*							
Overlap	.120	-.341**						
LSA	.230*	.042	.013					
HStO	.322**	.071	.093	.037				
Lesk	.125	.27	-.032	.098	.380**			
JCon	-.290**	-.392**	.485**	.035	.625**	.270*		
Lin	.073	.074	-.107	.053	.776**	.421**	.526**	
Resnik	.207	-.003	.052	-.063	.746**	.410**	.606**	.809**
* $p < .05$ (2-tailed) ** $p < .01$ (2-tailed)								

对比表 2 和表 3 可以看出,基于 BE 关系网络的评测模型与人工评测的相关性 0.768 这个上限还有一定的距离,但较文献[25]给出的基于实体关系网络的连贯性评测模型 Egrid(表 3 中黑体部分)的结果 0.246,已有很大的改善.以 BE 为内容单元,保留关系为“subj”、“obj”、“conj”、“nn”的 BE 时,基于 BE 关系网络的文摘连贯性模型评测结果与人工评测结果的相关系数为 0.408,提高了近 66 个百分点.

表 2 同时也显示,用 BE 选择方法 R1 时,由于所用的 BE 关系太少,不能很好地描述句子间的语法转换关系,但选用的 BE 关系太多时(R3),提高了模型的复杂度,使得模型过度拟合了训练集而在测试时表现欠佳.

5 结论与展望

本文讨论了一种基于 BE 关系网络的文摘内容连贯性评测模型.该模型基于这样的假设:在一个连

贯的文摘中,某句的出现以一定的规律依赖于其前面的句子.由于句子是由内容单元构成的,因此可将内容单元视为句子的语义,内容单元之间的关系视为句子的结构,这样,上述的假设可简化为:内容单元在某句中充当的语法角色以一定的规律依赖于其在前面句子中所充当的语法角色,这种依赖规律可以用马尔可夫模型来描述.

以句子为行、文摘中包含的内容单元为列,可以构建出关系网格,因为内容单元来自 BE,因此称这种关系网络为 BE 关系网格,网格中各单元的值为相应 BE 在句子中所充当的语法角色(此处用 BE 三元组中的“关系”部分).整个文摘的连贯性可用 BE 在相邻句中所充当的语法角色的转移概率来衡量,转移概率可在理想文摘集上进行训练得到.

由于很多 BE 并没体现句子的中心内容之所在,因此,需要按 BE 关系对 BE 进行筛选.本文对比了以 BEH、BEHM、BE 为内容单元、不同的 BE 筛选方法所得模型的性能.训练集选用 DUC2005 数

据集中的人工文摘,测试集用其中的机器文摘,模型的评测方法是考察模型对机器文摘内容连贯性的评测结果与人工评测结果之间的相关性.评测结果显示,以 BE 为内容单元,保留关系为“subj”、“obj”、“conj”、“nn”的 BE 时,模型评测结果与人工评测结果的 Pearson 相关系数为 0.408,比文献中给出的实体网格模型得到的结果提高了约 66%,这说明文摘的连贯性可用文摘中句子的转移概率模型描述,并且,基于 BE 关系网络的连贯性评测模型能够更好地抓住句子的语义信息和结构信息.

对文摘可读性的评测除连贯性外,还有其它几项指标,因此,如何对这些指标进行自动评测将是我们下一步研究的主要内容.

参 考 文 献

- [1] Rath G J, Resnick A, Savage R. The formation of abstracts by the selection of sentences, Part 1: Sentence selection by man and machines. *American Documentation*, 1961, 2(12): 139-208
- [2] Minel J L, Nugier S, Piat G. How to appreciate the quality of automatic text summarization?//*Proceedings of the ACL/ECL'97 Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain, 1997: 25-30
- [3] Jing H, Barzilay R, McKeown K et al. Summarization evaluation methods: Experiments and analysis//*Proceedings of the AAAI Symposium on Intelligent Summarization*. Menlo Park, CA, USA, 1998: 60-68
- [4] Goldstein J, Kantrowitz M, Mittal V et al. Summarizing text documents: Sentence selection and evaluation metrics//*Proceedings of the SIGIR-99*. Berkeley, CA, 1999: 121-128
- [5] Donaway R, Drummney K, Mather L. A comparison of rankings produced by summarization evaluation measures//*Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*. Seattle, 2000: 69-78
- [6] Over P. An introduction to DUC-2004: Intrinsic evaluation of generic news text summarization systems//*Proceedings of the Document Understanding Conference (DUC-2004)*. Boston, USA, 2004
- [7] Saggion H, Radev D, Teufel S et al. Meta-evaluation of summaries in a cross-lingual environment using content-based metrics//*Proceedings of the 19th International Conference on Computational Linguistics*. Taipei, China, 2002: 849-855
- [8] Lin C Y, Hovy E. Automatic evaluation of summaries using *N*-gram co-occurrence statistics//*Proceedings of the Human Language Technology Conference*. Edmonton, 2003: 71-78
- [9] Lin C Y. ROUGE: A package for automatic evaluation of summaries//*Proceedings of the ACL 2004 Workshop on Text Summarization*. Spain, 2004: 74-81

- [10] Hovy E, Lin C Y, Zhou L. Evaluating DUC 2005 using basic elements//*Proceedings of the Document Understanding Conference (DUC-2005)*. Vancouver, B. C., Canada, 2005
- [11] Hovy E, Lin C Y, Zhou L et al. Basic elements. Technical report, University of Southern California, 2005
- [12] Nenkova A, Passonneau R. Evaluating content selection in summarization: The pyramid method//*Proceedings of the HLT/NAACL 2004*. Boston, Massachusetts, USA, 2004: 145-152
- [13] Yu Shi-Wen, Duan Hui-Ming, Tian Jian-Qiu. The principle and realization of automatic evaluation for machinery digest: Intelligent computer interface and application progress//*Proceedings of the 3rd China Computer Intelligent Interface and Intelligent Application Conferences*. Zhangjiajie, China, 1997: 230-233(in Chinese)
(俞士汶,段慧明,田剪秋.机械文摘自动评测的原理及实现:智能计算机接口与应用进展//第3届中国计算机智能接口与智能应用学术会议论文集.湖南张家界,1997:230-233)
- [14] Shen Zhou, Wang Yong-Cheng, Xu Yi-Zhen et al. Study and practice of evaluation method for automatic summarizing system. *Journal of the China Society for Scientific and Technical Information*, 2001, 20(1): 66-72(in Chinese)
(沈洲,王永成,许一震等.自动文摘系统评价方法的研究与实践[J].情报学报,2001,20(1):66-72)
- [15] Zhang Shu, Zhao Tie-Jun, Zhao Hua et al. Application and analysis of content-similarity-based automatic evaluation for summarization systems. *Chinese High Technology Letters*, 2006, 16(3): 241-245(in Chinese)
(张姝,赵铁军,赵华等.基于内容相似度的文摘自动评测方法及其有效性分析.高技术通讯,2006,16(3):241-245)
- [16] Hu P, He T, Ji D et al. A study of Chinese text summarization using adaptive clustering of paragraphs//*Proceedings of the 4th International Conference on Computer and Information Technology (CIT'04)*. Wuhan, 2004: 1159-1164
- [17] Bangalore S, Rambow O, Whittaker S. Evaluation metrics for generation//*Proceedings of the INLG*. Mitzpe Ramon, Israel, 2000: 1-8
- [18] Papineni K, Roukos S, Ward T et al. BLUE: A method for automatic evaluation of machine translation//*Proceedings of the ACL*. Philadelphia, PA, USA, 2002: 311-318
- [19] Higgins D, Burstein J, Marcu D et al. Evaluating multiple aspects of coherence in student essays//*Proceedings of the NAACL*. Boston, Massachusetts, USA, 2004: 185-192
- [20] Miltasakaki E, Kukich K. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 2004, 10(1): 25-55
- [21] Foltz P, Kintsch W, Landauer T. Textual coherence using latent semantic analysis. *Discourse Processes*, 1998, 25(2&3): 285-307
- [22] Mani I. Summarization evaluation: An overview//*Proceedings of the NTCIR Workshop, Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization*. Tokyo, Japan, 2001

[23] Grosz B, Joshi A, Weinstein S. Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, 1995, 21(2): 203-225

[24] Walker M, Joshi A, Prince E. Centering Theory in Discourse. Oxford: Clarendon Press, 1998

[25] Lapata M, Barzilay R. Automatic evaluation of text coherence: Models and representations//Proceedings of the 19th International Joint Conference on Artificial Intelligence. Edinburgh, Scotland, UK, 2005



LIU De-Xi, born in 1975, Ph. D. , associate professor. His research interests include natural language processing, automatic summarization, text retrieval.

Ji Dong-Hong, born in 1966, professor, Ph. D. supervisor. His research interests include computational Linguistics, natural language processing.

Background

This work is supported by National Nature Science Foundation of China named by "A Study on Document Re-Ranking Techniques in Information Retrieval" (60703008) and "Sentence Ordering for Chinese Text Information Fusion" (60773011). To measure the performance of document re-ranking and sentence ordering techniques, a proper evalua-

tion method should be selected. However, most of the existing automatic evaluation methods for information retrieval and automatic summarization are based on recall ratio of content. So, the work in this paper aims at exploring an auto-evaluation method of content's readability.