北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.004

基于加权词汇衔接的文档级机器翻译自动评价

贡正仙† 李良友

苏州大学计算机科学与技术学院，苏州 215006，†E-mail: zhxgong@suda.edu.cn

摘要 在文档词汇衔接评价 LC 方法的基础上，提出基于权重的 LC，即 WLC，该方法通过在文档词图上运行 PageRank 算法获得词汇权重。根据词性信息使得 PageRank 算法偏向特定的词汇，并提出 PWLC 方法。实验表明，在文档级别上，所提的两种方法与人工评价的相关度都优于 LC，融合两种方法后，BLEU 和 TER 在文档级别上的评价性能有显著的提高。

关键词 词汇衔接；文档级评价；机器翻译；自动评价; PageRank

中图分类号 TP391

Document-Level Automatic Machine Translation Evaluation Based on Weighted Lexical Cohesion

GONG Zhengxian†, LI Liangyou

Department of Computer Science and Technology, Soochow University, Suzhou 215006; †E-mail: zhxgong@suda.edu.cn

Abstract Based on LC method, weighted LC (WLC) method is proposed, which assigns weights for words by PageRank algorithm running on word graph of documents. Furthermore, a new method named PWLC is also proposed, which biases PageRank algorithm to words with specific POS tags. The authors show how to combine the evaluation of lexical cohesion with other mainstream automatic evaluation metrics, in order to help these methods to evaluate translation quality at document level. Compared with LC, experiments show WLC and PWLC have higher Spearman correlation at document-level evaluation. Combined with others metrics, such as BLEU and TER, they both show better performance of evaluation at document level.

Key words lexical cohesion; document-level evaluation; machine translation; automatic evaluation; PageRank

机器翻译系统在篇章级别上的输出准确性对系统用户来说极为重要，因为相比独立的句子，他们更加关心一段文本的整体意思[1]。一个翻译系统，如果不考虑篇章上下文和句子间的恰当链接，而只是简单地把独立的句子放在一起，不管把句子翻译的多好，都不能保证输出文本的连贯性。但目前针对篇章质量的自动评价方法还很少，如主流的评价方法 BLEU、METEOR 和 TER 等，注重的是系统级别或者是句子级别的评价。这种评价方式一个很大的缺点就是忽视了篇章的上下文和结构信息。因此，在此类评价方法上进行优化的机器翻译系统也不太可能产生像人工翻译那样自然的文本。

Beaugrande 等[2]认为篇章具有 7 个基本特征：衔接性、连贯性、意图性、可接受性、信息性、情景性和跨篇章性，其中衔接性和连贯性被认为是区分一段文字是否构成篇章或者文本的两个基本特征。本文主要研究句子间的语言学特征：衔接性(cohesion)和连贯性(coherence)，并将它们加入到已有的评价中，以产生更好的文档级别的评分。在机器翻译评价框架 FEMTI [3]中，连贯性被定义为“读者能够描述出每一个句子或一组句子在一整篇文本中的角色的程度”。连贯性的度量必须依赖于衔接

863 计划(2012AA011102)和国家自然科学基金(61305088)资助
收稿日期: 2013–06–18；修回日期: 2013–09–22；网络出版时间: 2013-11-11 10:25

性, 即文本中存在的要表达的意思的关系[4]。衔接性通过句子间的语法和词汇元素的相互链接实现。

目前针对篇章的自动评价方法的研究非常有限, 文献[5−6]提出基于语篇表述理论(discourse representation theory, DRT)[7]的自动评价方法。DRT为篇章里的语义依赖关系提供了表述语言, 它使用语篇表述结构(discourse representation structure, DRS)描述上下文的语义联系。DRS有两个关键的组成部分: 一组存在于语篇中的实体和一组表示篇章中实体关系的DRS条件。例如, 句子"A farmer owns a donkey. "的DRS表达为[x, y: farmer(x), donkey(y), owns(x, y)]。因此两篇文档的DRS表述的匹配情况可以被用来作为文档的评价。

除了机器翻译评价之外, 文章自动评分程序, 如E-rater[8], 也采用丰富的文档特征, 包括文法、用法、风格、背景、主要观点、支持观点、结论方式等。然而, 上面描述的方法都依赖于语言学特征, 特征的解析过程可能会因为译文中的文法错误而影响评价的性能。因此这些方法的准确性和可靠性也不可避免地会依不同的评价数据而产生波动。

与上述方法不同, 文献[9]使用一种与文法错误无关的词汇衔接对文档进行评价, 并将它们与其它的主流方法进行融合。词汇衔接是衔接手段中最重要的一种形式, 占据英语衔接手段中近一半的数量[4]。与高度依赖于文本句法准确度的文法衔接相反, 词汇衔接几乎不被文法错误影响, 它只需要依赖于现有多数语言都存在的词典。文献[9]的词汇衔接手段被定义为一篇文档中出现一次或重复多次的实义词, 包括同义词、近义词、上位词、副本和搭配。副本是指文档中的相同词汇或经过词形还原后被认为是相同的词汇。文献[9]的实验结果表明, 词汇衔接和传统的自动评价方法的联合使用能显著地提高与人工评测的相关性。

1 词汇衔接评价—LC的介绍

文献[9]通过下式计算一篇译文的词汇衔接性:

$$\mathrm{LC}=\mathrm{lcd}\big/\mathrm{cw}, \tag{1}$$

其中, lcd是文中词汇衔接手段的数量, cw是文中实义词的数量。LC越高则表示实义词中词汇衔接手段的比例越高。lcd的识别和计算过程可以用如下所示的算法来更清晰地表述。

输入: 文档 d

输出: 文档 d 的词汇衔接手段集合

for 文档 d 中的每一个实义词 w do

  for 词汇链集合 L_d 中的每一个词汇链 L do

    for L 中的每一个单词 w' do

      if $w=w'$ or $\mathrm{stem}(w)=\mathrm{stem}(w')$ then 将 w 加入 L

      else if w 和 w' 属于同一词集 then 将 w 加入 L

      else if w 和 w' 在 WordNet 中距离为 1 then 将 w 加入 L

      else if $\mathrm{sim}(w,w')\geqslant 0.96$ then 将 w 加入 L

      end if

    end for

  end for

  if w 未加入到任何词汇链中 then

    创建新的空词汇链 L', 将 w 加入 L'

    将 L' 加入 L_d

  end if

end for

删除 L_d 中长度为1的词汇链

L_d 中的单词则为词汇衔接手段

通过上述算法, 我们可以看到文献[9]通过LC来衡量文本的衔接性时, 它平等地对待每一个实义词。因为不同类型的实义词携带的信息量不同, 我们很自然地想到用不同的权重区别对待这些词汇, 从而更准确地衡量译文的衔接性。因此本文在LC的基础上提出两种新的方法WLC和PWLC。

2 带权重的词汇衔接评价——WLC和PWLC

与LC不同, 本文的方法不是简单地进行计数, 而是根据词的权重进行计算, 称为基于权重的词汇衔接, 即WLC以及依赖于词性的PWLC(pos-WLC)。本文使用PageRank算法进行权重的计算, 在运行算法之前, 首先需要构建文档的词图。

2.1 词图

在自动文摘技术的研究中, 用图来表示文档已经得到广泛的认同[10−11], 本文采用文献[11]描述的简单图模型来表示文档上下文中词之间的连接关系, 它是运行PageRank的基础。本文中, 词之间连接的方向是这样确定的: 使用一个宽度为 W 的滑动窗口, 在图中添加窗口内的第一个词指向窗口中其它词的连接, 该窗口每次滑动一个词的距离。

图1是窗口大小为3的有向词图的一个示例,

本文中词图的构建只使用实义词。可以看出，除了句子内部的上下文关系，词图也通过共有词汇表达句子间的关系。

2.2 PageRank

在介绍 PageRank[12]之前，本文先给出一些数学符号。用 $G=(V,E)$ 表示文档的词图，其中节点集 $V=\{w_1,w_2,\cdots,w_N\}$。如果从 w_i 到 w_j 有一条连接，那么 $(w_i,w_j)\in$ 边集 E。在一个词图中，每一个节点表示一个词，每一条边表示词之间的关联。(w_i,w_j) 边的权重表示为 $e(w_i,w_j)$，节点 w_i 的出度表示为 $O(w_i)=\sum_{j:w_i\to w_j} e(w_i,w_j)$。

PageRank 是一种有名的排序算法，它使用链接信息为网页分配一个全局的重要性分数。PageRank 的基本思想是如果有其他重要的节点指向一个节点，那么这个节点也是重要的。这种方法可以看成是节点间的投票或者推荐。在 PageRank 中，一个词 w_i 的分数 $R(w_i)$ 被定义为

$$R(w_i)=\lambda\sum_{j:w_j\to w_i}\frac{e(w_j,w_i)}{O(w_j)}R(w_j)+(1-\lambda)\frac{1}{|V|},\quad(2)$$

其中，λ 是阻尼因子，值在 0 和 1 之间，$|V|$ 是节点数。该阻尼因子表明，每一个节点都有一个随机跳到其它节点的概率 $(1-\lambda)$。通过不断迭代运行式(2)直到收敛可以得到 PageRank 的分数。式(2)中的最后一部分可以看成一个平滑因子，它使得图满足非周期和不可约的特性，以此保证 PageRank 向一个唯一的固定的分布收敛。在 PageRank 中，对图中的所有节点，这一部分设置成一个相同的值 $\frac{1}{|V|}$，表明随机跳到所有节点的概率都相同。

由于 PageRank 是迭代算法，本文将其最大迭代次数设置为 100，或者所有节点的值变化之和小于 0.00001 时停止算法。

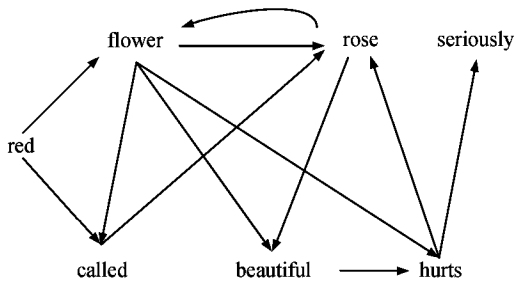

图 1　词图示例

Fig. 1　Example of word graph

2.3 WLC 和 PWLC 的计算

本文利用 PageRank 对词图中的词分配权重，然后根据文档的词汇衔接情况采用下式来计算 WLC:

$$\mathrm{WLC}=\frac{\sum_{w\in\mathrm{lcd}}R(w)}{\sum_{w\in\mathrm{cw}}R(w)},\quad(3)$$

其中，lcd 是词汇衔接中的词，cw 是文档中的实义词。与 LC 不同，WLC 会突出文档中的重要词汇对文档衔接性的影响。

此外，式(2)的最后一部分可以被设置成不同的值。假设要给某些节点分配比较大的概率，那么最后 PageRank 得分就会偏向这些节点。而 PWLC 的基本思想就是这样运行 PageRank 算法，使得某些词汇具有较高的权重。

在本文中，根据每一个词 w_i 的词性可以分配一个词性相关的值 $p_{\mathrm{pos}}(w_i)$，作为该词的随机跳转概率，且满足 $\sum_{w_i\in V}p_{\mathrm{pos}}(w_i)=1$。这样，PageRank 的公式就可以改写成式(4)的形式:

$$R(w_i)=\lambda\sum_{j:w_j\to w_i}\frac{e(w_j,w_i)}{O(w_j)}R(w_j)+(1-\lambda)p_{\mathrm{pos}}(w_i)。\quad(4)$$

根据词性分配不同权重的基本思想：一篇文档中的词的词性分布具有不平衡性，且通常名词占据文档词的大部分(参见 3.2 节实验部分的统计)，这些词对文档的理解有重要的影响；而且通常一篇文档中的名词代表文档的主要对象，会经常反复出现，因此其衔接情况相对能反映出文档的衔接性。

通过实验观察，按照文档词性的分布情况，本文定义词性权重如表 1 所示。由于本文使用 WordNet 获取每个词的所有可能词性，因此每个词的词性相关的权重 $\mathrm{weight}_{\mathrm{pos}}(w_i)$ 是所有词性权重的均值。词 w 在词图中的跳转概率定义为

$$p_{\mathrm{pos}}(w)=\frac{\mathrm{weight}_{\mathrm{pos}}(w)}{\sum_{w'\in d}\mathrm{weight}_{\mathrm{pos}}(w')},\quad(5)$$

其中 d 表示文档。

以图 1 为例，“rose”在 WordNet 中有名词、动词

表 1　不同词性的权重

Table 1　Weights of different POS

词性	权重
名词	0.50
动词	0.28
形容词	0.15
副词	0.05
其他	0.02

和形容词 3 种词性, 因此它的词性权重为(0.5 + 0.28 + 0.15)/3 = 0.31。计算每个词的权重后, 进行归一化, 最终形成图 2。图中每个词都加上了基于词性的跳转概率。每个词分配一个不同的跳转概率后, PageRank 会偏重概率较大的节点, 之后按照 WLC 的方法计算的词汇衔接评价即为 PWLC, 词汇衔接与传统评价方法的联合框架。

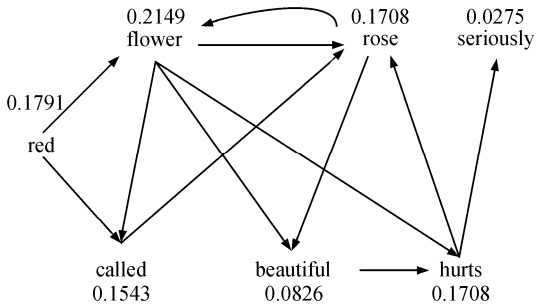

图 2　基于词性的跳转概率示例

Fig. 2　Example of word graph with transfer probability based on POS

从前文可以看出, LC, WLC 以及 PWLC 都可以针对一篇文档, 给出一个 0 到 1 之间的评价结果, 反映文档的词汇衔接情况。所以它们都可以作为独立的评价方法用于文档的评价, 同时也可以与其它已有的评价方法进行组合。

本文使用下式将词汇衔接加入到已有评价中[9]:

$$H = \alpha m_{\text{doc}} + (1-\alpha) m_{\text{seg}}, \tag{6}$$

其中 α 是 0 到 1 之间的实数, m_{doc} 是直接针对文档的评价, 在本文中指词汇衔接评价, m_{seg} 是已有的评价方法给出的文档评分, 本文将在实验中选取几个代表性的传统方法来观察融合的效果。

3　实验

通常, 评价的粒度分为句子级、文档级和系统级。由于本文的评价数据含有的系统太少, 产生的相关系数可信度不高, 所以本文所有的实验均未给出系统级的评价。此外, 由于衔接性评价是针对文档级进行, 不适用于句子级别的评价, 因此本文给出的实验结果都是针对文档级的相关系数。

3.1　实验准备

本文在实验中共使用两个数据集: MTC Part2(LDC2003T17) 和 MTC Part4(LDC2006T04), 简记为 MTC2 和 MTC4。其中 MTC2 主要出现在前 3 个实验中, 用来评测本文方法的优劣; MTC4 用于健壮性分析。

数据集的相关统计见表 2。在评价前所有句子都进行分词和小写处理, 并且去除译文中的未登录词。两个数据集上的每一个系统译文都有至少两个人工评价者给出的评分: 流利度(Fluency)和适当性(Adequacy), 分值在 1 到 5 之间。

表 2　MTC2 和 MTC4 数据信息

Table 2　MTC2 and MTC4 dataset

信息	MTC2	MTC4
系统数	3	6
文档数	100	100
句子数	878	919
参考译文数量	4	4
源语言	Chinese	Chinese
目标语言	English	English
题材	Newswire	Newswire

本文根据文献[13]的方法对人工评分进行归一化处理。由于评测数据没有直接给出文档的评分, 本文采用文献[9]方法获得文档的近似得分①: 一篇文档的评分是该文档中所有句子上的评分的均值。因此在实验中, 每一个句子和每一个文档都有一个 Adequacy 评分和一个 Fluency 评分。本文使用 Spearman 等级相关系数 ρ 测量自动评价方法的结果与人工评分的相关性。

3.2　词汇类型统计

在进行词汇衔接评价的实验前, 需要先对数据集中的词进行过滤。在本实验中, 停用词为常用停用词表中的词和标点符号以及含有非 26 个英语字母字符的单词, 文档经过停用词过滤后剩余的即为实义词, 这些实义词会应用在词汇衔接的分析中。数据统计如表 3 所示, MT 表示机器译文, HT 表示人工译文。

表 3　MTC2 评价数据中的词汇统计

Table 3　Statistics of words with different POS on MTC2

词汇类型	MTC2	
	MT	HT
所有词	29038	27394
实义词	16224	13660
名词	12104	9959
动词	6804	5555
形容词	4133	3651
副词	1307	1028
其他	756	970

① 随着评测单位的扩大, 人工直接给出文档的得分也成为一项艰巨的任务; 此外, 人的主观性和差异性都会带来误差。

表 3 中的数据是每个系统统计结果的平均数, 词性的统计使用 WordNet, 并取一个词的所有词性。从表 3 可以看出, 大约有一半左右的词是实义词; 而实义词中, 多数词具有名词词性, 其次为动词、形容词和副词, 而在 4 个词性之外的词只是少部分。

3.3 词汇衔接评价独立使用的结果

本文提出的基于 PageRank 的词汇衔接方法是建立在词图的基础上, 而词图的构建与窗口大小有关, 因此本节实验用来观察不同窗口大小下评价性能的变化。实验中取词干作为图节点, 即构建词图时每个词经过 Porter Stemmer 处理。

在 MTC2 上的实验结果如图 3 所示。从图 3 可以看出, 随着窗口的不断增大, WLC 的性能不断下降; 与 WLC 相比, 虽然 PWLC 也总体上呈现下降趋势, 但是当窗口增大时, PWLC 性能逐渐高于 WLC。该图中的下降趋势的一个可能原因是: 随着窗口的增大, 词图中的边也随之不断增多, 加入的这些边连接的都是距离较长的一些词汇, 但是这些距离信息却不能反映在词图中, 因此词图中词的区分度会变小。因为 PWLC 在词图的节点中加入了词汇的权重信息, 在运行 PageRank 时含有较高的权重的节点受到窗口的影响相对较小, 所以其性能的下降也相对平缓。

从图 3 可以发现, 用 PageRank 训练词汇权重后, 词汇衔接方法的评价性能有显著的提高。同时, 虽然窗口增大使得 WLC 和 PWLC 性能有所下降, 但是依然高于 LC, 尤其在 Fluency 上性能的差异更加显著。而且与 Adequacy 相比, Fluency 上的相关度受到窗口变化的影响较小, 也可以说明基于权重的词汇衔接在评价文档连贯性方面具有比较大的优势。

3.4 词汇衔接评价与其他评价联合使用的结果

与目前主流的方法不同, 词汇衔接方法直接对文档进行评价。因此, 本文将 LC, WLC 和 PWLC 三种方法与其他方法进行融合, 以提高它们在文档评价上的性能。WLC 和 PWLC 的窗口大小都取 10, 此时两者的性能有一定的差距, 而且差距不是最大, 这样便于比较两者在融合方面的差异, 而且具有较小的过拟合风险。

实验选取 BLEU、TER 和 METEOR 三种方法, 这三种方法都是目前比较主流且具有代表性的方法。同时这三种方法也分属于 3 个类别, 它们与词汇衔接评价的融合采用式(6)的形式。

文献[9]在实验中优化了这三种方法与 LC 融合时的参数, 优化目标是最大化与人工 Adequacy 评分的 Pearson 相关系数, 优化后的结果如表 4 所示。此优化结果被直接应用在本文实验中。

表 5 给出不同的评价方法以及它们与基于词汇衔接的 3 个方法融合后与人工评价的 Spearman 相关系数。从表中可以看出, 三种基于词汇衔接的方法都有效地提高了三种主流的评价方法在文档级的评价效果。尤其在 BLEU 和 TER 上表现的最为显著。LC 使 BLEU 和 TER 在 Adequacy 上取得最好的评分, 本文提出的 WLC 和 PWLC 在 Fluency 上表现得更好。而且对比 WLC 和 PWLC 发现, 这两种方法在 MTC2 数据上没有显著的差异。

3.5 词汇衔接评价方法的健壮性

为了测试词汇评价方法的健壮性, 本文用 3 种评价衔接性的方法以相同的设置在 MTC4 上做了相应的实验。表 6 是 3 种词汇衔接评价在 MTC4 上与人工的 Spearman 相关系数。

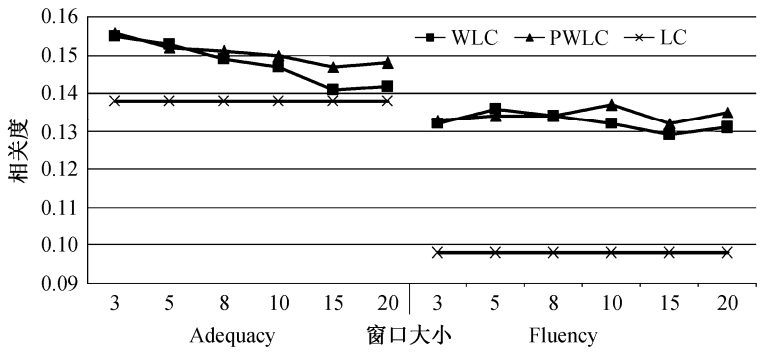

图 3 LC与不同窗口大小下的WLC 和 PWLC 在文档评价上的 Spearman 相关系数

Fig. 3 Spearman coefficients of LC, WLC and PWLC under different window size

表 4　将 LC 加入到其他方法后的优化参数

Table 4　Advanced tuning parameters of combining traditional metric with LC

评价方法	α
BLEU	0.29
TER	0.38
METEOR	0.18

表 5　不同方法融合词汇衔接后在 MTC2 上的 Spearman 相关系数

Table 5　Spearman coefficients of different metrics on MTC2

方法	MTC2	
	Adequacy	Fluency
LC	0.1391	0.097
WLC	0.1465	0.1329
PWLC	0.1504	0.1352
BLEU	0.0906	0.1087
BLEU+LC	0.1487	0.1513
BLEU+WLC	0.1397	0.1554
BLEU+PWLC	0.1373	0.1530
TER	–0.1738	–0.1886
TER+LC	–0.2160	–0.2143
TER+WLC	–0.2081	–0.2237
TER+PWLC	–0.2065	–0.2223
METEOR	0.3349	0.2753
METEOR+LC	0.3509	0.2914
METEOR+WLC	0.3525	0.2972
METEOR+PWLC	0.3531	0.2964

将表 6 与表 5 进行对比，发现在两个数据集上，本文提出的两个方法 WLC 和 PWLC 在评价中优于 LC，且 PWLC 在多数情况下表现得更好。虽然在与 METEOR 的融合时，在 MTC4 上的效果并不显著，却也可以看出词汇衔接在多数情况为融合评价带来了帮助，而且优化的参数是根据 MTC2 上的实验进行选取的，因此总体上表明，词汇衔接方法在两个数据集上具有较好的一致性。

3.6　词汇衔接评价与主流评价方法的相关度

考虑到词汇衔接在不同方法上融合效果的不一致性，本文做了一组实验来计算词汇衔接与各种不同的评价方法间的相关度，如表 7 所示。可以看出，在 MTC2 和 MTC4 上，BLEU 和 TER 与词汇衔接的相关度最低，这两种方法与词汇衔接的结合产生的效果也最好；相反，METEOR 与词汇衔接的相关度最高，而相应地，实验中词汇衔接也在这两个方法上的表现最不突出。本文认为这种较高的相关度是词汇衔接在 METEOR 效果不显著的一个原因。

表 6　不同方法融合词汇衔接后在 MTC4 数据集上的 Spearman 相关系数

Table 6　Spearman coefficients of different metrics on MTC4

方法	MTC4	
	Adequacy	Fluency
LC	0.3160	0.2502
WLC	0.3417	0.2815
PWLC	0.3576	0.2927
BLEU	0.6055	0.5093
BLEU+LC	0.6250	0.5266
BLEU+WLC	0.6284	0.5330
BLEU+PWLC	0.6311	0.5349
TER	–0.5173	–0.4572
TER+LC	–0.5704	–0.4956
TER+WLC	–0.5736	–0.5059
TER+PWLC	–0.5765	–0.5067
METEOR	0.6981	0.5543
METEOR+LC	0.6956	0.5537
METEOR+WLC	0.6958	0.5570
METEOR+PWLC	0.6969	0.5579

表 7　词汇衔接与不同评价方法间的文档评分的相关度

Table 7　Correlation between lexical cohesion–based metric and other traditional metrics

方法		BLEU	TER	METEOR
MTC2	LC	–0.1324	–0.0161	0.0869
	WLC	–0.1038	–0.0486	0.0676
	PWLC	–0.0973	–0.0649	0.0703
MTC4	LC	0.2529	–0.1380	0.3866
	WLC	0.2886	–0.1786	0.4378
	PWLC	0.2973	–0.1911	0.4491

4　总结

本文在词汇衔接评价 LC 的基础上提出了 WLC 和 PWLC 两个评价方法。它们采用基于词图的 PageRank 得到词汇的权重，且 PWLC 在计算权重时又考虑了词汇的词性带来的影响。最后，本文使用了一种简单的方法将词汇衔接评价加入到已有评价方法中。

在词汇衔接性的评价上，本文提出的基于 PageRank 的方法 WLC 以及基于词性权重的 PWLC 方法优于已有的 LC 方法。同时将这 3 种方法融合进 BLEU，TER 和 METEOR 之后，发现它们能有效地提高 BLUE 和 TER 在文档评价上的效果，但是并

没有对 METEOR 产生明显的帮助。通过分析各种方法之间的相关性发现，LC, WLC 和 PWLC 与 METEOR 和 PBE 的相关度较高，这可能是融合效果较小的一个原因。

参考文献

[1] Visser E M, Fuji M. Using sentence connectors for evaluating MT output // Proceedings of the 16th Conference on Computational Linguistics. Stroudsburg, 1996: 1066–1069

[2] De Beaugrande R, Dressler W. Introduction to text linguistics. London and New York: Longman, 1981

[3] Margaret K, Andrei P B, Eduard H. FEMTI: creating and using a framework for mt evaluation // Proceedings of MT Summit IX. New Orleans, 2003: 224–231

[4] Halliday M A K, Hasan R. Cohesion in English. London: Longman Pub Group, 1976

[5] Gim´enez J. IQMT v 2.1. Technical manual[EB/OL]. (2007)[2013–05–20]. http://www.lsi.upc.edu/nlp/IQMT/IQMT.v2.1.pdf

[6] Gim´enez J, M`arquez L, Comelles E, et al. Document-level automatic MT evaluation based on discourse representations // Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Stroudsburg, 2010: 333–338

[7] Kamp H, Reyle U. From discourse to logic: introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory. Studies in Linguistics and Philosophy Series, 1993: 81–91

[8] Burstein J. The E-rater scoring engine: automated essay scoring with natural language processing. Automated Essay Scoring: A Cross-disciplinary Perspective, 2003: 113–121

[9] Wong B, Kit C. Extending machine translation evaluation metrics with lexical cohesion to document level // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea, 2012: 1060–1068

[10] Litvak M, Last M. Graph-based keyword extraction for single-document summarization // Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization. Manchester, 2008: 17–24

[11] Schenker A, Bunke H, Last M, et al. Graph-theoretic techniques for web content mining. Machine Perception and Artificial Intelligence. World Scientific Publishing Co, Inc, River Edge, NJ, 2005

[12] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the web[EB/OL]. (1999) [2013–05–20]. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

[13] Blatz J, Fitzgerald E, Foster G, et al. Confidence estimation for machine translation[EB/OL]. (2004) [2013–05–20]. http://web.eecs.umich.edu/~kulesza/pubs/confest_report04.pdf