中 文 信 息 学 报
·第 18 卷 第 2 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.18 No.2

文章编号:1003－0077(2004)02－0015－08

OpenE:一种基于 n-gram 共现的自动机器翻译评测方法*

孙连恒,杨 莹,姚天顺

(东北大学 计算机软件与理论研究所语言工程实验室机器翻译评测研究组,辽宁 沈阳 110004)

摘要:在机器翻译研究领域中,评测工作发挥着重要的作用,它不仅仅是简单地对各个系统输出结果进行比较,它还对关键技术的发展起到了促进作用。译文质量的评测工作长期以来一直以人工的方式进行。随着机器翻译研究发展的需要,自动的译文评测研究已经成为机器翻译研究中的一个重要课题。本文讨论了基于 n-gram 共现的自动机器翻译评测框架,介绍了 BLEU、NIST、OpenE 三种自动评价方法,并通过实验详细分析了三种方法的优缺点。其中的 OpenE 采用了本文提出了一种新的片断信息量计算方法。它有效地利用了一个局部语料库(参考译文库)和全局语料库(目标语句子库)。实验结果表明这种方法对于机器翻译评价来说是比较有效的。

关键词:人工智能;机器翻译;机器翻译评测;信息量计算;n-gram 共现

中图分类号:TP391.2 **文献标识码**:A

OpenE: an Automatic Method of MT Evaluation Based on N-gram Co-occurrence

SUN Lian-heng, YANG Ying, YAO Tian-shun

(MTE group in Language Engineering Lab Dept.

of Computer Science and Technology Northeastern University, Shenyang, Liaoning 110004, China)

Abstract: Evaluations are very helpful for the research of Machine Translation (MT). The aim of evaluations is not only to output the differences among MT systems, but also to stimulate the improvement of key technologies in this area. In the past, the evaluations of MT are performed by human. With the increasing needs of MT research, the automatization of MT evaluations becomes more and more important. This paper introduces the basic framework of automatic MT evaluation using n-gram co-occurrence statistics. Three methods (BLEU, NIST and OpenE) based on this framework are described. The advantages and disadvantages of these methods are also discussed through the analysis of several experiments. Among these methods, OpenE adopts a new method of n-gram weighting which employs a local corpus and a large global corpus. Through the experiments, this method is proved to be practical for machine translation evaluation.

Key words: artificial intelligence; machine translation; MT evaluation; information computing; n-gram co-occurrence

1 简介

评测工作在各个研究领域都发挥着重要的作用,它不仅仅是简单的对各个研究成果进行比较,它还对关键技术的发展起到了促进作用。在机器翻译的研究工作中,对译文的评测工作

* 收稿日期:2003－09－05

基金项目:国家重点基础研究资助项目(G19980305011)

作者简介:孙连恒(1978—),男,硕士研究生,主要研究方向为机器翻译及其评测 .

一直在以人工方式进行。近年来,国外的研究机构已经开始了自动机器翻译评测的研究工作,并取得了喜人的成果。在 IBM 推出它的自动评价系统 BLEU 后,NIST 在其基础上又添加了片断信息量计算方法。本文介绍的是我们在承担国家 973 汉英机器翻译评测项目的基础上,提出的一个新的片断信息量计算方法。这个方法被应用于一套自动译文评测工具——OpenE。

2 基于 n-gram 共现的评测简介

2.1 两种实际的方法

2.1.1 BLEU (BiLingual Evaluation Understudy)

这种方法由 IBM 提出[1]。虽然它不能称得上是一种完善的自动评测方法,有许多部分需要改进,但是其构思却是很巧妙。让我们先通过一个直观的例子来了解这种方法。

英语译文:Can I have a word with you?

参考答案 1:May I talk to you?

参考答案 2:Can I have a talk with you?

要计算译文的 BLEU 得分,先要算出各个 n-gram 的精确度(precision)。具体算法就是先把译文中所有出现在参考答案集合中的 n 元片断数目累加起来再除以译文的总片断数。例如,上面例子中的一元片断准确度(unigram precision)是 6/7,二元片断准确度(bigram precision)是 4/6。BLEU 还针对机器翻译的特点改进了精确度算法,有兴趣的读者可以参考相关论文[1]。

最终得分的计算是取各元文法的平均值,可以是指数平均值、算术平均值等。最后,为了防止较短句子的得分过高,得分还要乘以一个长度罚分比(brevity penalty)。

$$BP = \begin{cases} 1 & \text{if } c < r \\ \mathrm{EXP}(1 - R/C) & \text{if } c \geqslant r \end{cases}$$

其中,c 表示被测译文单词数,r 表示参考答案中单词数最接近 c 的句子的单词数。

最终的得分公式为(这里取了各元文法的指数平均值):

$$BLEU = BP \cdot \exp\left(\left(\sum_{n=1}^{N}(w_n \log p_n)\right)\right)$$

w_n 表示各阶匹配正确率的权重,p_n 就是匹配正确率。

BLEU 的方法被证明与人工评测有很高的相关性[1]。

2.1.2 NIST

NIST(National Institute of Standards and Technology)方法是在 BLEU 方法上的一种改进[2]。它并不是简单的将匹配的 n-gram 片断数目累加起来,而是求出每个 n-gram 的信息量(information),然后累加起来再除以整个译文的 n-gram 片断数目。信息量的计算公式是:

$$Info(w_1 \cdots w_n) = \log_2\left(\frac{\text{the \# of occurrences of } w_1 \cdots w_{n-1}}{\text{the \# of occurrences of } w_1 \cdots w_n}\right)$$

NIST 的网站提供了一个自动译文评测工具 mteval-kit,它用 PERL 实现了 BLEU 和 NIST 的方法。当 $n=1$ 时:

$$Info(w_1) = \log_2\left(\frac{\text{the \# of all the words}}{\text{the \# of occurrences of } w_1}\right)$$

NIST 的评分公式为:

$$Score = \sum_{n=1}^{N} \left\{ \sum_{\substack{\text{all } w_1 \cdots w_n \\ \text{that co-occur}}} Info(w_1 \cdots w_n) / \sum_{\substack{\text{all } w_1 \cdots w_n \\ \text{in sys output}}} (1) \right\} \cdot \exp\left\{ \beta \log^2 \left[\min\left(\frac{L_{sys}}{L_{ref}}, 1 \right) \right] \right\}$$

可以看出 NIST 最终得分分为两个部分，一部分是各元 n-gram 的平均信息量的总和；另一部分是长度罚分比。其中，β 是一个可变参数，它在 L_{sys} 是 L_{ref} 的 2/3 时使罚分比是 1/2。但是 NIST 的罚分比同 BLEU 相比在译文句子与参考答案句子的长度差异较小时对最终得分的影响要小一些。

经过实验比较，NIST 方法的区分度（Sensitivity）和同人工评分的关联度（Correlation）比 BLEU 都要好[2]。但是从对后面的一些实验结果的分析，我们可以看出 NIST 的方法存在一些缺陷，那就是它不能很好反映高阶 n-gram 的匹配情况。

2.2 N-gram 共现评测方法的关键部分

上面介绍了基于 n-gram 共现的两个自动译文评测方法，从中我们可以大概了解基于 n-gram 共现的机器翻译的框架。这个框架可分为三大部分：片断信息量权重计算、分数规范化和长度罚分比。其中的每一部分都有许多方案可供选择。下面将要讨论的就是在这个框架下的三个关键部分。

2.2.1 信息量权重计算

给片断分配权重的方法在信息检索中有很多，比较著名方法的是布尔权重、词频权重、$tf \times idf$、熵权重等[3]。其中，布尔权重的只包括 0 或 1，因此 BLEU 可以看成是采用了布尔权重的计分法；词频权重给那些出现频率高的片段更多地权重；$tf \times idf$ 和熵权重都基于这样一种假设：如果某个片断在某个局部语料库中的出现频率较高，但是在全局语料库中的出现频率却较低，那么这个片断对于局部语料库来说就应该具有较高的权重。在信息检索中全局语料库就是文档的集合，局部语料库就是其中的某些文档；如果将这两种方法应用于机器翻译评测中，全局语料库可以是一个目标语句子库，局部语料库是对应于一个原文句子的一组译文参考答案。后面将要介绍的 OpenE 方法中的片段信息量计算就是基于这种假设的。

2.2.2 规范化的分数

计算了每个 n-gram 片断的权重之后，我们需要解决的下一个问题是如何将权重转化为规范化的分数（normalized score）。因为只有给出了规范化的分数，我们才能判断一个翻译系统的输出效果到底如何。而通过不同的评测方法对同一个翻译系统进行打分时，只有规范化的分数才能使这些不同的评测方法具有可比性。

我们可以看出，BLEU 的得分直接使用了匹配精确度。NIST 却没有一个规范化的分数而是直接给出它的 n-gram 平均信息量。

2.2.3 长度罚分比

我们知道，通常同一个句子的不同译文的长度是比较接近的，而机器翻译的输出有时会输出过长或过短的结果。在 BLEU 和 NIST 中分数计算都是采用匹配的 n-gram 数目与被测译文总的 n-gram 数目的比值。这种方法对较短的句子有偏向，也就是越短的句子会有越高的得分，解决的方法是用罚分比将较短的句子分数罚去一些。

3 OpenE（Open Evaluation）

OpenE 是一个汉英机器翻译自动评测平台，它属于 973 汉英机器翻译评测项目的一部分，它包括对分词、词性标注、短语结构以及最终译文的全自动评测。其中译文评测方法就是基于 n-gram 共现的框架。这种方法是对 NIST 方法的一种改进。像 NIST 一样它也要计算每个 n-

gram 片断的信息量，然后累加，最后除以参考答案的平均句信息量，只是这个信息量的计算采用了另一种新的方法。

信息量的计算公式如下：

$$Info(w_1\cdots w_n) = \log_2(f_{ref}^{w_1\cdots w_n}/f_{corpus}^{w_1\cdots w_n})$$

上式中，$f_{ref}^{w_1\cdots w_n}$和$f_{corpus}^{w_1\cdots w_n}$分别是 n-gram 在参考答案和全局语料库的出现频率。这个信息量的计算方法的想法来源于 $tf\times idf$ 的启发。在信息检索的权重计算中 tf 表示某个检索项的在某篇文档中的出现频率，idf 表示反向文本项频率(inversed document frequency)。其中 $idf=\log(N/nk)$，nk 是包含检索项的文档数目，N 表示所有文档的总数。最后每个检索项的权重为 $tf\times idf$。但是，用于自动评测的权重计算方法又不完全等同于 $tf\times idf$。因为评测的全局语料库并不像信息检索的语料库那样被分成一个一个的文档。于是我们选择 $f_{ref}^{w_1\cdots w_n}$来代替 tf，用 $1/f_{corpus}^{w_1\cdots w_n}$来代替 idf，最后再取对数。取对数的目的是压缩权重的值域。这样我们便得到了上面的权重计算公式。

再一次看上面的例子：

英语译文：Can I have a word with you?

参考答案 1：May I talk to you?

参考答案 2：Can I have a talk with you?

单词 Can 在参考答案中的出现频率为 1/12，而在全局语料库中得出下频率为 1/1000(不是真实值)，因此根据上面的公式我们可以得出 Can 的信息量为 $\log_2(\frac{1}{12}/\frac{1}{1000})=6.38$。同理，我们可以计算所有的 n-gram 的信息量。

各阶的得分计算如下：

$$p_n = \min\left\{\frac{\sum_{\substack{\text{all}\,w_1\cdots w_n\\ \text{that co-occour}}} Info(w_1\cdots w_n)}{average\_ref\_info_n}, 1\right\}$$

其中，$average\_ref\_infon$ 为参考答案的 n-gram 平均句信息量计算方法如下：

$$average\_ref\_info_n = \frac{\sum_{\substack{\text{all}\,w_1\cdots w_n\\ \text{in the ref set}}} Info(w_1\cdots w_n)}{\text{number of ref sentences}}$$

OpenE 算法采用匹配的 n-gram 权重之和与参考译文每个句子 n-gram 权重之和的平均值之间的比值。这个方法对长句子有偏向，因此也要采用罚分比，与 BLEU 和 NIST 不同的是这个罚分比针对的是长句子而不是短句子。OpenE 的长度罚分比为：

$$BP = \exp\left\{\beta\log^2\left[\min\left(\frac{\overline{L}_{ref}}{L_{sys}}, 1\right)\right]\right\}$$

这个罚分比同 NIST 的罚分比很相似，只是将 L_{sys}和L_{ref}调换了位置。

最后的得分为：

$$score = BP\times\sum_{n=1}^{N}(w_n p_n)$$

其中 $w_n=1/N$，表示各阶 p_n 占总分的比重。

可以看出，OpenE 的片段权重计算方法不仅考虑了片断在当前参考答案中的出现频率，而且还加入了一个全局的大语料库信息。这样可以避免给那些没有意义但是却在答案中出现频

率很高的片断以过高的权值。

4 几种方法的比较

4.1 实验数据

下面我们将通过实验来比较这三种方法的效果。BLEU 在他们的实验中已经论证了基于 n-gram 共现的评测的可行性，因此我们这里就不再重复这些实验。我们实验主要集中在对各个自动评测系统之间的比较上面。为了比较三种自动评测系统，我们选用人工评分作为参照，看看他们哪个与人工评分更加接近。我们的实验语料是 LDC 新华社文本，共有 440 个句子。其中每个汉语句子的英语译文参考答案有 11 个。进行 n-gram 匹配时，最大的 n-gram 长度为 $N=4$，各阶的 n-gram 占总分的比例 $w_n=1/N$。OpenE 方法中用到的全局语料库为 BNC (British National Corpus)。我们先用人工方式给 LDC 中的两个机器翻译系统输出($tb1$, $tb2$)进行了评分，再用上面提到三种方法给这两个系统进行了自动评分。

这里要讨论一下全局语料库的选择。我们面临两个方案：一个是所有句子参考译文的集合，另一个是规模很大的、能够反映片断或单词在真实情况中出现频率的目标语语料库。我们认为，只要语料库的规模足够大，它就能反映片断在一般情况下出现的频度，至于选择哪个方案并没有本质的区别。但是在我们现有的实验环境中，前者的规模比较小，只有 4000 多个英文句子，而且这些句子只是根据 400 多个源语句子生成的，因此不能很好的反映片断的真实出现频度，因此我们便决定在本次实验中选择后者——一个大规模的目标语语料库。

其中人工评分包括两个部分——充分度和流畅度。下面是人工评分的标准：

• 充分度评分标准：

5 分：完全表达了原文的意思；

4 分：基本表达了原文的意义，但有个别地方(1 处)不准确；

3 分：表达主要意义，有 2 处错误；

2 分：3 处错误；

1 分：4 处错误；

0 分：4 处以上错误；

• 流畅度评分标准：

5 分：语序正确，非常流畅；

4 分：基本通顺，关键部分正确；

3 分：基本通顺，关键部分 1 处错误；

2 分：关键部分有 2 处错误；

1 分：关键部分 3－4 处错误；

0 分：关键部分 5 处以上错误；

图 1 和图 2 分别显示了对于机器翻译系统 $tb1$ 和 $tb2$，440 个译文句子自动评价得分同人工评价得分的比较。其中 NIST 得分已经被规范至[0,1]区间。每一个点表示一个句子，它的纵坐标是人工评分，横坐标是自动评分。

从图中我们可以看出，NIST 方法与人工评分比较接近，OpenE 次之，BLEU 得分有些偏小。需要说明的是，NIST 得分已经被规范化至[0,1]。具体方法是用 NIST 得出的信息量除以所有译文参考答案的平均信息量。

三种方法的得分都主要集中在斜率为 1 的直线附近，但是它们的得分普遍要比人工评分

低。其原因应该是尽管参考答案有 11 个,但还是不能将所有可能的片段包含进去。这就造成有些片断在人工方法开来是正确的但是在自动方法中却没有被计入总分。

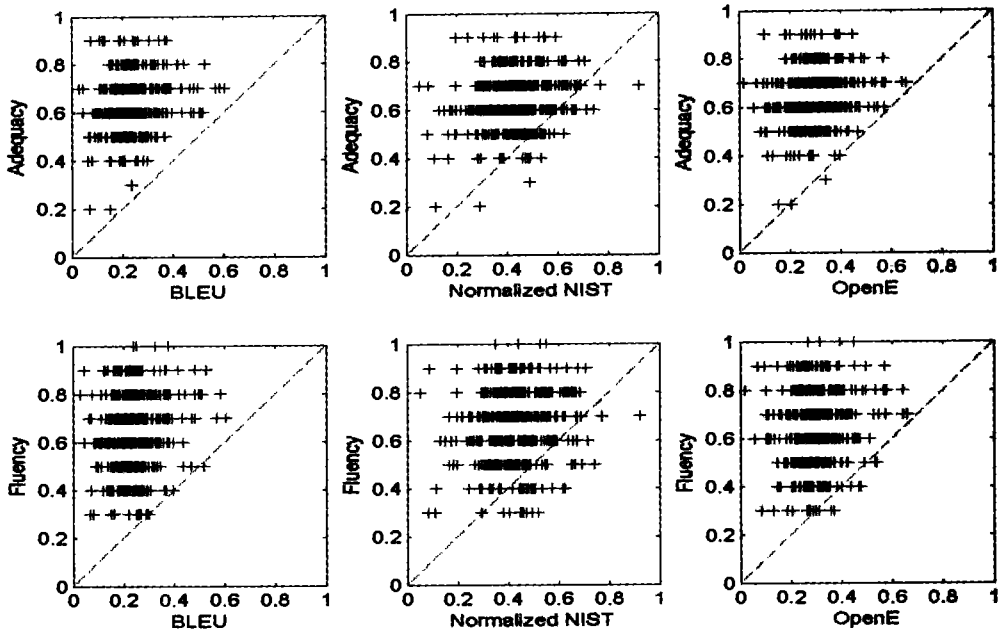

图 1　三种评分方法同人工方法的比较(对于系统 tb1 的翻译结果)

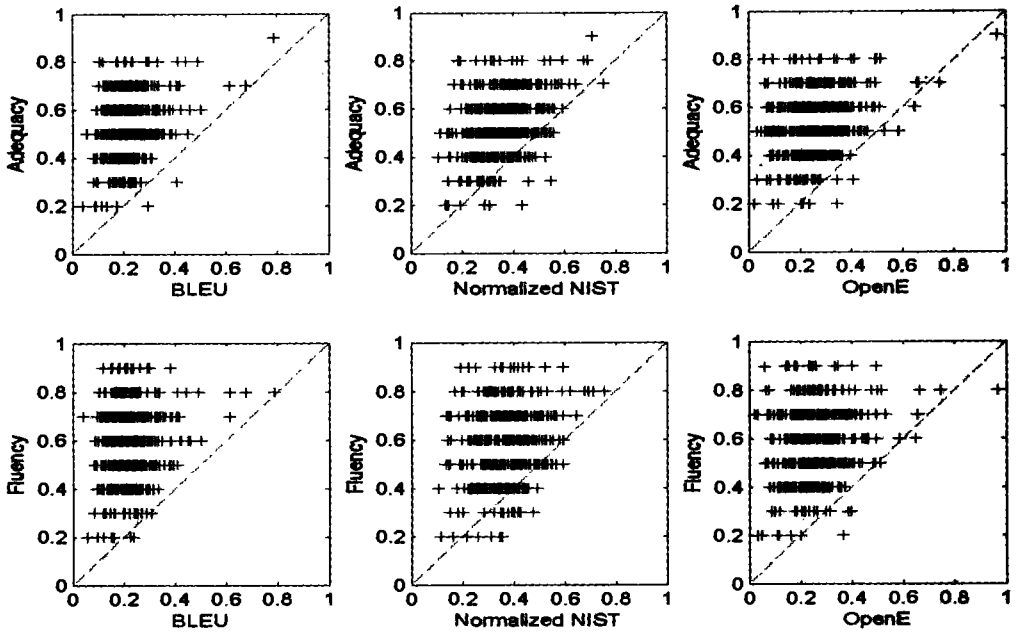

图 2　三种评分方法同人工方法的比较(对于系统 tb2 的翻译结果)

图 3 显示了三种自动评价方法对两个机器翻译系统的区分度,衡量的指标是 F-ratio。可以看出 NIST 的方法区分度最高,OpenE 次之,BLEU 最低。

4.2 三种方法的优缺点

这里分析一下各种方法的优点和缺点。

• BLEU 方法的优点是算法简单，速度快；缺点是没有对 n-gram 片断给予必要的权重，因此 BLEU 不能区分哪些片断对最终结果的贡献比较大。

• NIST 方法的优点是给每个 n-gram 片断分配了一个自动统计权重，这样 NIST 的得分可以比 BLEU 方法更能反映被测译文句子和参考答案句子之间的语义匹配程度。但是 NIST 存在一定的缺陷，详细的分析请见下面的实验数据。

对系统 *tb*1 的输出，我们统计了三种方法 $N=1$ 至 5 的得分。

从图 4 可以看出，NIST 得分随着 N 的增大(特别是在 $N>2$ 时)变化较小，而另外两种方法的得分却随着 N 的增大急剧减小。我们认为这种现象是由于不同的 n-gram 信息量计算方法造成的。

于是我们按照不同的信息量计算方法分别统计了一下 LDC 新华社文本所有参考答案的 n-gram 平均信息量，具体方法是按照各自的信息量公式求出所有 n-gram 的信息量，累加之后再除以 n-gram 的总数。结果如表 1。

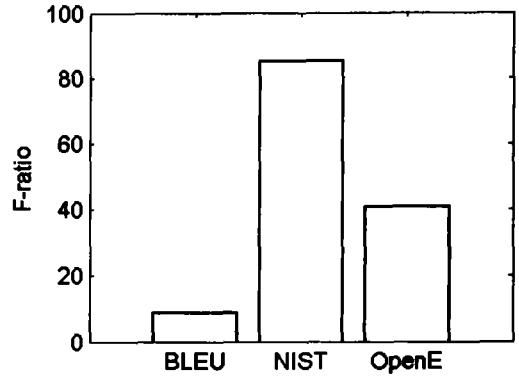

图 3 三种自动评价方法的区分度比较

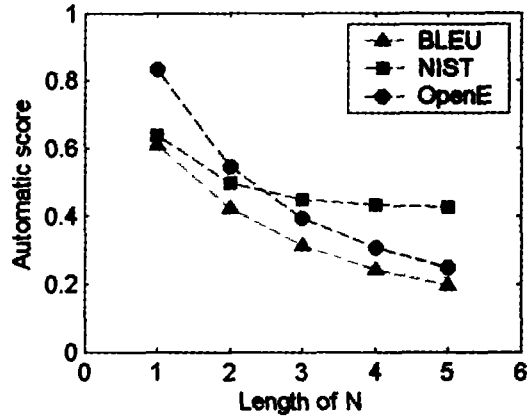

图 4 系统 tb1 在 $N=1$ 至 5 时三种方法的得分

表 1 新华文本所有译文参考答案的平均 n-gram 信息量

	1－gram	2－gram	3－gram	4－gram
BLEU	1	1	1	1
NIST	9.31	4.71	1.85	0.67
OpenE	8.46	16.64	17.78	18.14

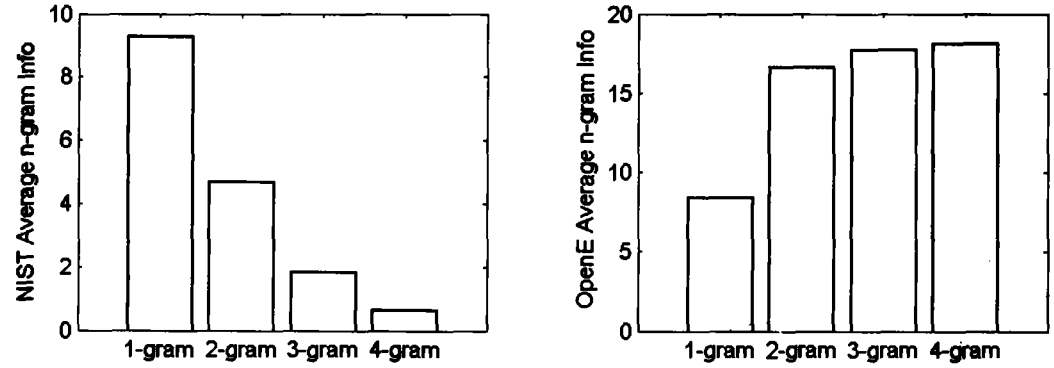

图 5 NIST 和 OpenE 的平均 n-gram 信息量分布比较图

从图 5 可以看出 NIST 的各阶 n-gram 平均信息量分布很不均匀，它的信息量主要集中在 1－gram 上，也就是说 NIST 的最终得分中 1－gram 要占很大比重，高阶 n-gram 片断所起的作用非常小。这导致了 NIST 的得分受最大的 n-gram 长度 N 的影响较小(如图 4 所示)，从而不能反映高阶 n-gram 的匹配情况。其它两种方法的各阶 n-gram 的信息量分布比较均匀，因此最后得分都能很好反映各阶 n-gram 的匹配情况。

• OpenE 方法弥补了上述两种方法的不足，既加入了统计权重，又保证了各阶 n-gram 权重的均匀分配。但是 OpenE 方法的最后得分要比人工评分偏低。我们认为原因有两个，一是人工的评分同自动评分方法的方法和标准存在一定的差异；二是高阶 n-gram 匹配正确率很低造成了最后得分的偏低。从实验数据上我们看出，OpenE 方法的效果并没有取得显著的提高，但是它所采用的权重计算方法却能挖掘更多的语义信息(通过平均分配各阶 n-gram 权重)用于译文自动评判。

5 未来的工作

从前面的结果来看，自动评测方法已经取得了一定的成功，它在一定程度上可以与人工评分相关联。但是自动得分还是没有同人工得分完全一致，主要表现在自动评分总是比人工评分偏低。我们认为，在基于 n-gram 共现的自动评测框架中，未来的工作主要集中在两个方面，一是寻求更好的片断信息量的计算方法，使片断信息量能够充分反映它的语义重要性；二是建立大规模参考答案库，使得所有的可能译文片断都可以在答案库中被尽可能地找到。

对机器翻译的人工评测可以分为许多方面，例如准确度、可信度、忠实度等[4~6]。但是使这些评测的自动化却是一个很大的问题，原因在于机器不可能像人类那样利用自己的语法和其他方面的知识去理解译文然后给出评价，它只能以其独有的方式对译文进行评价。现今用于译文质量自动评测的方法都是根据片断的重要程度来计算得分的。由此看出，机器评判和人工评判是分别从两个角度去看待译文质量的。如何去将这两个不同角度的结果统一起来，是自动评测的急需解决的关键问题。

参 考 文 献：

[1] Kishore Papineni, et al. BLEU: a method for automatic evaluation of machine translation[R]. Technical Report RC22176 (W0109－022), IBM Research Division, Thomas J. Watson Research Center, 2001.

[2] Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics [R]. NIST Research Report, 2002.

[3] Kjersti Aas, Line Eikvil. Text Categorisation: A Survey[M]. Raport NR 941. Norwegian Computing Center, 1999.

[4] E.H. Hovy. Toward finely differentiated evaluation metrics for machine translation[A]. In: Proceedings of the Eagles Workshop on Standards and Evaluation, Pisa, Italy, 1999.

[5] EAGLES. Evaluation of Natural Language Processing Systems FINAL REPORT[R]. EAGLES DOCUMENT EAG-II-EWG-PR.1, 1999.

[6] J.S. White, T. O'Connell. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches[A]. In: Proceedings of the First Conference of the Association for Machine Translation in the Americas, 193－205, Columbia, Maryland, 1994.