汉语分词在机器翻译评价中的影响*

徐冰 姚建民 杨沐昀 赵铁军

哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001

E-mail: {xb;james;ymy;tjzhao}@mtlab.hit.edu.cn

摘要：机器翻译评价对机器翻译系统的发展有重要的推动作用。本文针对目前流行的 IBM 提出的机器翻译自动评价方法，探讨了分词信息对于外汉机器翻译评价的影响。实验结果证明在评测汉语译文中用汉语分词方法将提高自动评测的准确度。

关键词：机器翻译评价，n 元关系，分词

Effect of Chinese Word Segmentation in Machine Translation Evaluation

Xu Bing, Yao Jianmin, Yang Muyun, Zhao Tiejun

School of Computer Science & Technology, Harbin Institute of Technology, Harbin, 150001

E-mail: {xb;james;ymy;tjzhao}@mtlab.hit.edu.cn

Abstract: The evaluation of machine translation (MT) plays an important role in development of MT system. This paper discusses the effect of word segmentation in evaluation of foreign-Chinese MT, especially in the popular automatic method proposed by IBM.. The experiment shows that application of word segmentation in evaluating Chinese translation will improve the precision.

Keywords: machine translation evaluation, N-gram relation, word segmentation

1. 引言

机器翻译评测是推动机器翻译系统发展的重要手段。人工的机器翻译评测虽然正确率高，但是花费的代价也很高，所以一种有效的机器翻译自动评测方法可以节省很多人力的投入。近几年在机器翻译领域中提出了许多新的机器翻译的自动评价方法，每一种方法都要回答这样两个问题[1]：（1）如何评价一个机器翻译系统好？（2）如果有两个机器翻译系统你怎样评价出哪个更好？

评价一个机器翻译系统的好坏包括多项内容，从用户的角度来说，他们更关心的是机器译文质量的好坏，同时译文质量也是评价机器翻译系统的最关键的指标[2]。但是，译文质量的好坏很难准确地评价，量化标准对人来说仍是一项十分棘手的任务，所以一般都采用机器译文的打分和人工译文的打分相比较的方法，并且认为越接近于人工译文打分的机器译文越

*本文研究受到国家 863 计划资助(项目号 2002AA117010-09)。

好[3]。

回顾近几年的机器翻译的自动评测方法，大部分研究集中在机器译文和标准译文之间的相似度计算方面。例如， Jones 提出应用语言学上的相关信息作为研究译文质量的方法，比如句法树、n 元关系、语义同现等[4]，Brew 使用词频、词性标注及其他文本特征决定译文质量[5]。另外还有一些方法是通过对机器译文和人工译文的比较来评价译文质量的好坏，比如 Yokoyama 提出基于评测方法的双向机器翻译，他对输出日文和原日文在单词识别上，修饰成分的正确性上进行比较[6]。Akiba 通过计算机器译文和人工译文对应词的距离来评价机器译文质量[7]，这种计算距离的方法还有 Alshawi 也采用过[8]。近年来 n 元关系的方法被许多研究者普遍关注，如 Langkilde[9]和 Kishore Papineni[3]等。

本文重点研究了机器译文和人工译文的 n 元关系匹配的方法，采用了 IBM 提出的 BLEU 算法的基本原理，自动评价了译文为汉语的 500 个句子的质量。研究中重点考虑到汉语是按字书写的语言，词才是汉语表意的基本单位，直接以字为单位评价可能出现评测上的偏差，所以又以词为单位对汉语译文进行了评测。实验结果表明分词后的汉语句子在自动评价中得分更接近人工打分。

2. BLEU 基本原理

为了说明机器译文的质量，IBM 提出的 BLEU 方法采用了译文是英文的机器译文和人工译文的 n 元关系匹配，相匹配的词的个数越多，机器译文的得分就越高。虽然 BLEU 方法是针对机器译文是英文的句子，但是对于汉语理论上同样适用，只不过其中的 n 元关系存在按字计算还是按词计算的问题。

下面我们以原文是英文的汉语译文为例来具体说明字和词的一元关系及二元关系的匹配过程。

原文：Accuracy is most important in translation.

机器译文：准确性是在译文里最重要的。

人工译文：正确性在翻译中是最重要的。

如果我们以字为单位对比机器译文和人工译文的句子，那么采用一元关系的评测方法准确度应该是 9/12=0.75，其中“确”、“性”、“是”、“在”、“译”、“最”、“重”、“要”、“的”这 9 个字在人工译文里出现，机器译文共有 12 个字，所以得到该句机器译文的得分。

采用二元关系的评测方法得分应该是 4/11≈0.36，因为在机器译文中出现的只有“确性”、“最重”、“重要”、“要的”与人工译文中的词能匹配上。

如果机器译文和人工译文的句子先分词，那么分词后的句子就变为：

机器译文：准确性 是 在 译文 里 最 重要 的 。

人工译文：正确性 在 翻译 中 是 最 重要 的 。

则分词后按一元关系评测得分为 5/8≈0.63，与人工译文能匹配上的有“是”、“在”、“最”、“重要”、“的”。按二元关系评测得分为 2/7≈0.29，与人工译文匹配上的有“最 重要”和“重要 的”。

在 IBM 提出的 BLEU 算法中，句子得分采用了下述公式进行处理。

首先该算法考虑到了机器译文相对于人工译文的长度问题。一种情况是机器译文只包含原文中一两个词的译文，故这种机器译文比人工译文要短很多，那么在打分时要乘以一个惩罚因子；另一种情况是机器译文包含原文中的某个词的多个译文，那么机器译文就要比人工译文长，在匹配的过程中只对先匹配上的词予以承认。下面的公式计算了惩罚因子，c 代表机器译文，r 代表人工译文。

得到惩罚因子后，在计算每一句的机器译文分值时，就应用下面的公式：

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \le r \end{cases}$$

其中，P_n代表各元关系的精确度，Wn 表示如果 n 取到三元关系时，记为 1/3。

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

3. 实验及分析

3.1 实验内容

我们使用了 9000 句双语语料库中的 500 句作为测试集进行测试。实验中所指的人工译文是语料库中的标准汉语译文，机器译文是由某个英汉翻译系统得出的汉语译文。

人工给机器译文评分的过程是由 6 名专家独立的给每句机器译文打分，每句的平均分作为人工评测机器译文质量的结果。打分的标准采用六等级制，详见表 1。

然后我们采用匹配方法对机器译文进行自动评测。第一步是将机器译文对照人工译文直接应用字的一元关系和二元关系，计算出准确度；第二步是将机器译文和人工译文进行分词，然后采用词的一元关系和二元关系计算出准确度。

表 1 人工评测标准

分 数	评 价 指 标
1	译文准确、流畅地传达了原文的信息，除个别错别字外，无需修改；
0.8	译文传达了原文的信息。不用参照原文，就能明白译文的意思，但是译文在语法、择词选择、汉语表达习惯等方面多少有些问题，需要修改。不过这种修改无需参照原文也能有把握地进行，且修改也较容易；
0.6	译文大致表达了原文的意思，局部与原文有出入，一般情况下需要参照原文才能改正。有些情况即使无需参照原文也能猜测到原文的意思，但译文的不妥明显是由于翻译程序的缺陷造成的；
0.4	译文有一部分符合原文的一部分意思，全句没有译对，不过原文全句的词都孤立地译出来了，对人工后编辑有点用处；
0.2	看了译文不知所云或者意思完全不对。不过总有一些局部或词语是译对了的；
0	完全没有译出来。

3.2 实验结果及分析

根据上面的人工打分标准人工评测后我们得到表 2。

表 2 译文质量的人工评测结果

分值	句子数	百分比
[0,0.2]	46	9.2%
(0.2,0.4]	127	25.4%
(0.4,0.6]	190	38%
(0.6,0.8]	93	18.6%
(0.8,1]	44	8.8%
译文质量得分	58.4	

在表 2 中，随着分值的增高译文质量也增高，由每个句子的打分结果可计算出系统的总的译文质量得分，若按百分制计算得分应为 58.4。

按照字和词的一元关系和二元关系自动评测机器译文得到表 3。在表 3 中我们列出了不同方法采用百分制得到的评测分,图 1 是对应表 3 的分值图。

表 3 译文质量的评测结果

方法	分值（百分制）
人工打分	58.4
字一元关系	64.7
词一元关系	59.0
字 BLEU	41.3
字二元关系	40.7
词二元关系	21.4
词 BLEU	25.2

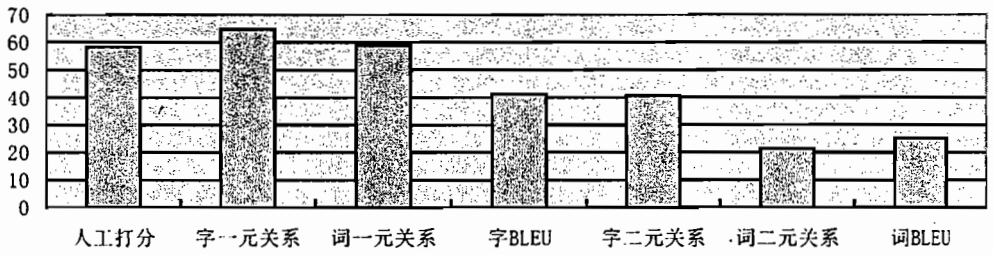

图 1 人工评测和自动评测分值图

分析表 3 可以得出如下结论：

（1）按字和按词的一元关系得分结果较接近人工打分的分值，尤其是按词的一元关系打分更接近人工打分结果，这说明给句子分词后再采用一元关系处理的方法提高了自动评测的准确性，外汉机器翻译评测一般应在“词”的层面上构建模型。

（2）按字的二元关系与词的一元关系得分结果相差较大，这是由于字二元关系将词与词的边界关系也记入了得分，而汉语本身的语序灵活性比较大，所以简单的利用字的二元关系计算汉语句子的相似度会造成计算结果偏低的现象，分词后采用一元关系评测更合理。

（3）分别按字和按词使用 BLEU 算法得到的分值与人工打分相差较大，一部分是由于

判罚因子的关系，在汉语译文中由于语序的灵活性不能通过比较机器译文和人工译文长度来进行判罚；另外是由于二元关系的分值较低导致最终分值较低。

为了说明实验结果的可靠性，下面列出了 200 句人工打分和自动打分的分值比较曲线。

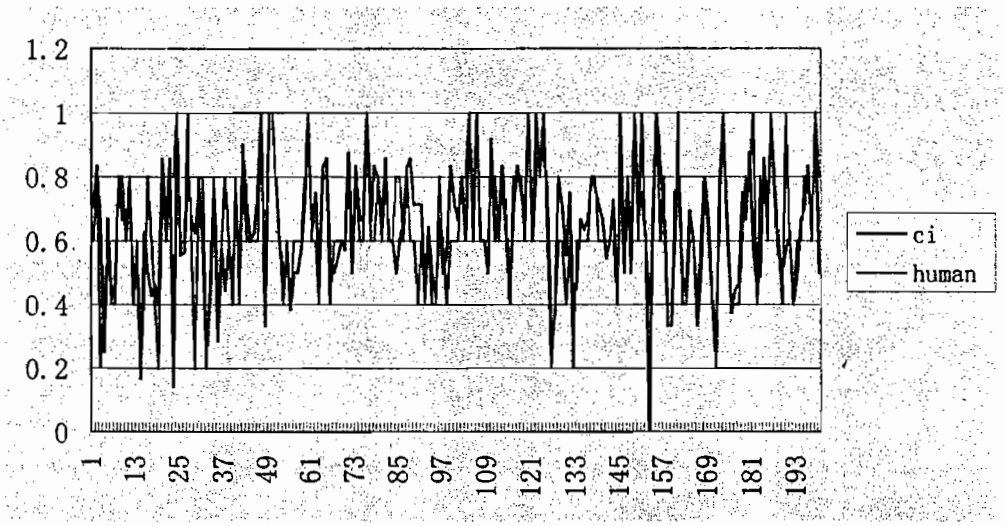

图 2 人工打分与词一元关系分值的比较

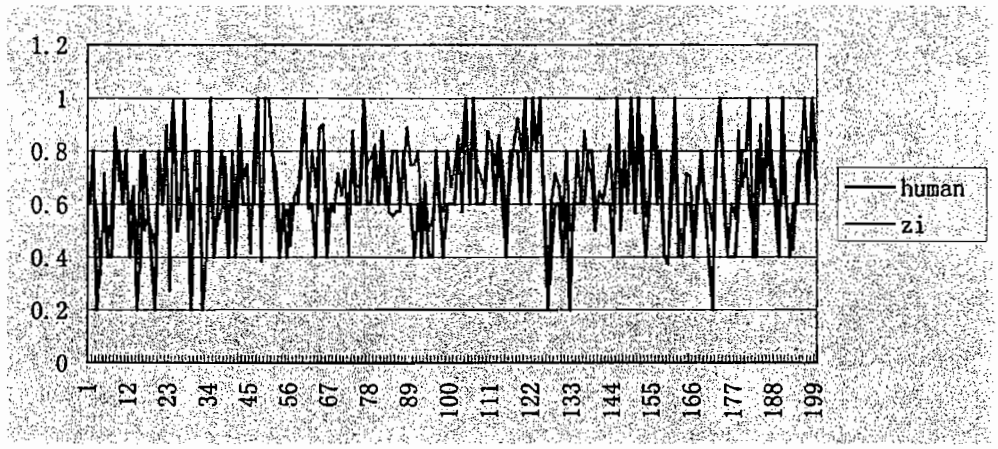

图 3 人工打分与字一元关系分值的比较

从图 2 和图 3 可以看到大部分句子的人工打分和机器打分基本上在同一个分值段内变化，偏差不是很大。在图 2 中人工打分和词一元关系分值的平均偏差为 0.08,图 3 中人工打分与字一元关系分值的平均偏差为 0.12。

进一步观察我们还发现，实验中对于机器译文人工打分为 1 的句子，按字一元关系的得分往往比按词一元关系的得分要高。初步分析其中的原因，我们发现这是由于在机器译文中出现和人工译文同义的词时，采用分词的方法就认为该词没有和人工译文中的词匹配上，而事实上这两个词意思相同。如何处理这种语言表层实现不同而语义相同的问题，实际上是目

前这种简单的基于 n 元关系的自动评测技术面临的一大挑战，是进一步研究中需解决的关键问题。

4. 结论

汉语是按字书写的语言，但是词才是汉语表意的基本单位。本文研究表明，如果忽略汉语这一特点，直接以字为单位采用类似基于 n 元关系的自动评测方法，所获得的结果会出现有一定的偏差。实验表明，外汉机器翻译评测应该先对汉语进行分词再应用 n 元关系匹配的评测方式，并且采用一元关系评测准确度更为合理。

参 考 文 献

[1] Niamh Bohan, Elisabeth Breidt, Martin Volk, Evaluating Translation Quality as Input to Product Development, 2nd International Conference on Language Resources and Evaluation, Athens, 2000.

[2] 赵铁军等：《机器翻译原理》，哈尔滨工业大学出版社，2000。

[3] Kishore Papineni, Salim Roukos, BLEU: a Method for Automatic Evaluation of Machine Translation,2001.

[4] Douglas A.Jones, Gregory M.Rusk, Toward a Scoring Function for quality-driven Machine Translation, Proceedings of COLING-2000.,2000.

[5] Brew C, Thompson H.S, Automatic Evaluation of Computer Generated Text: A Progress Report on the TextEval Project, Proceedings of the Human Language Technology workshop,108-113,1994.

[6] Shoichi Yokoyama, Hideki Kashioka, etc. An Automatic Evaluation Method for Machine Translation using Two-way MT, 8th MT Summit conference, Santiago de Compostel,2001.

[7] Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita, Using Multiple Edit Distance to Automatically Rank Machine Translation Output, Mt summit conference, Santiago de compostela,2001.

[8] Alshawi, H., S. Bangalore, and S. Douglas. . Automatic acquisition of hierarchical transduction models for machine translation. In Proceedings of the 36 th Annual Meeting of the Association for Computational Linguistics, Montreal Canada, Vol.I: 41-47,1998.

[9] Langkilde, I., and K. Knight. Generation that exploits corpus-based statistical knowledge. In Proceedings of the 36 th Annual Meeting of the Association for Computational Linguistics, and 17 th International Conference on Computational Linguistics, Montreal, Canada. 704-710,1998.