基于 N-gram 的无参考译文机器翻译自动评测方法[1]

徐金安　蒋俊杰
北京交通大学计算机与信息技术学院　北京　100044
E-mail: {jaxu, 10120469}@bjtu.edu.cn

摘 要：本文提出了一种无需参考译文的机器翻译自动评测方法，基本思想是利用源语言和目标语言的语言模型分别计算源语言句子和系统译文中所有 n-gram 的平均概率，再利用 n-gram 平均概率对系统译文进行打分。实验结果表明，本文方法与 BLEU、NIST 等自动评测方法的评测结果保持了很好的一致性。本文方法的主要贡献体现在普通用户可以在无参考译文的情况下，得到机器翻译系统译文的可信度，进而增强机器翻译自动评测的实用性。

关键字：机器翻译评测；自动评测；无参考译文

A N-gram based Automatic MT Evaluation Method without Reference Translations

Jinan Xu, Junjie Jiang
School of Computer and Information Technology,
Beijing Jiaotong University, Beijing 100044, China
E-mail: {jaxu, 10120469}@bjtu.edu.cn

Abstract: *This paper presents a novel automatic MT evaluation method without human reference translations. The basic idea of our proposed approach consists of three parts. Firstly, calculate average n-grams probability of source sentence with source language models, and similarly, calculate average n-grams probability of machine-translated sentence with target language models, finally, use the relative error of two average n-grams probabilities to mark machine-translated sentence. The experimental results show that our method can achieve high correlations with a few automatic MT evaluation metrics, such as BLEU, NIST, etc. The main contribution of this paper is that users can get MT evaluation reliability in the absence of human reference translations, which greatly improving the utility of MT evaluation metrics.*

Keywords: *machine translation evaluation; automatic evaluation; without reference translations*

1 引言

由于人工评测具有周期长、成本高的缺点，人们提出了多种机器翻译自动评测方法。大多数自动评测方法通过计算系统译文与参考译文的相似度进行评分。按照所使用的资源和计算方法进行分类，机器翻译自动评测方法通常可分为三大类。第一类采用基于 n 元匹配的方式，如 BLEU[Papineni et al., 2002]、NIST[NIST, 2002]等，此类评测方法通过计算系统译文与参考译文的共现 n-gram 对系统译文质量进行评价。第二类采用基于编辑距离的方式，如 WER[Sonja et al., 2000]，PER[Leusch et al., 2003]，TER[Snover et al., 2006]等。第三类采用基于语言知识的方式，如 METEOR[Banerjee et al., 2005]，PosBLEU[Popović and Ney, 2007]等，这类评测方法通常引入了句法结构，其效果依赖于句法分析工具的有效性和健壮性。

上述机器翻译自动评测方法采用不同的数学模型对系统译文质量进行评价，取得了比较好的效果。对研究人员而言，自动评测方法通过计算翻译系统改进前后译文的评价结果，可

[1]本文研究得到以下项目资助：中央高校基本科研业务费专项资金(2009JBM027)(K11JB00210)；北京市重点学科共建项目（计算机应用技术）；中科院计算技术研究所智能信息处理重点实验室开放课题(IIP2010-4).

以帮助改良翻译系统，进而促进机器翻译技术的研究和发展。但是，对于广大的机器翻译用户而言，获取参考译文并对翻译结果进行评价，通常是一件不合逻辑又非常困难的事情。另一方面，随着机器翻译技术的飞速发展，机器翻译用户的数量在不断增加。一个非常重要的问题是：如何在没有参考译文的情况下，对译文质量进行打分，为机器翻译用户提供值得信赖的译文质量评价服务。

为了解决上述问题，本文提出了一种无参考译文的机器翻译自动评测方法。基本思想是利用源语言和目标语言的语言模型分别计算源语言句子和系统译文中所有 n-gram 的平均概率，再计算 n-gram 平均概率的相对误差作为评分，从而对系统译文质量进行评价。实验结果显示，本文提案方法评测结果与 BLEU、NIST 等自动评测方法的评测结果保持了很好的一致性。

本文第二部分回顾了一些无参考译文条件下的自动评测方法研究工作，第三部分论述本文提案方法的基本原理，第四部分论述不同语言间(法英、德英)进行实验验证的结果和分析，第五部分是结论和下一步工作。

2 相关工作

普通机器翻译用户通常会在不同系统译文间挑选评价更高的译文使用，因而迫切需要一种无参考译文条件下的自动评测方法对译文质量进行可信的评价。近年来，针对无参考译文的自动评测方法，研究人员作了大量的相关工作。

Michael Gamon 等人提出了一种采用语言模型和支持向量机技术的自动评测方法，可以找出译文中翻译质量较差的句子[Gamon et al., 2005]。Andrew Mutton 等人使用机器学习技术，对译文的流利度进行了有效的评价[Muttor et al., 2007]。Joshua S.Albrecht 等人在自动评测中引入回归学习方法，取得了一定的成效[S.Albrecht et al., 2007]。Reinhard Rapp 认为针对句子级评测 BLEU 等方法的效果不佳，提出计算 Back-translation 句子与源语言句子的正交相似度对系统译文进行评价[Rapp, 2009]。

虽然研究人员作了许多尝试，但目前仍无一种有效的无参考译文条件下的自动评测方法，无参考译文的自动评测方法研究依然是一个极具挑战性和实用性的课题。因此，本文提出了一种基于 N-gram 的无参考译文机器翻译自动评测方法。首先，利用源语言和目标语言的语言模型分别计算源语言句子和系统译文中所有 n-gram 的平均概率，然后计算 n-gram 平均概率的相对误差作为评分，从而对系统译文质量进行评价。

3 本文提案方法

图 1 是依据本文提案方法基本思想设计的流程框图：

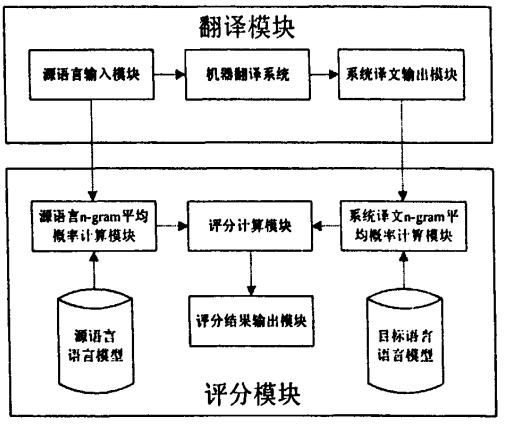

图 1. 提案方法流程

如图 1 所示，本文提案方法流程包括翻译模块和评分模块。在翻译模块中，首先输入源语言句子，经过机器翻译系统翻译后得到系统译文并输出。在评分模块中，分别利用源语言语言模型和目标语言语言模型计算源语言和系统译文的 n-gram 平均概率，输入至评分计算模块，计算评分并输出。

3.1 基本原理

大多数机器翻译自动评测方法通过计算系统译文与参考译文的相似度对系统译文质量进行评价。在参考译文不存在的情况下，本文提案方法通过计算系统译文与源语言句子的相似度进行评价。具体来讲，计算源语言句子和系统译文的 n-gram 平均概率，利用 n-gram 平均概率的相对误差对系统译文进行打分。

系统译文是对源语言句子用另一种语言或方式的诠释，因此可以将源语言句子视为一种“特殊”的参考译文。理论上，源语言句子及其正确翻译结果之间，彼此的 n-gram 具有严格的映射关系。译文中翻译质量较好的 n-gram，与源语言句子中对应的 n-gram，传递了相近的信息。从信息论的角度而言，两者具有相近的信息量，因而具有相近的概率值。因此，可以用系统译文与源语言句子中对应 n-gram 概率值的相似度来表示系统译文与源语言句子的相似度。然而，找出对应的 n-gram 是不易的。因此，本文使用句子中所有 n-gram 的平均概率来代替对应 n-gram 的平均概率。

为计算句子的 n-gram 平均概率，首先需要计算句子概率。假定句子 $s = w_1 w_2 \dots w_l$，其中 w_i 为一个字、词或短语，以 n-gram model 为例，使用公式 1 计算句子 s 的概率：

$$p(s) = \prod_{i=1}^{l+1} p(w_i \mid w_{i-n+1}^{i-1}) \tag{1}$$

公式 1 中，$p(s)$ 表示句子 s 的概率，$p(w_i \mid w_{i-n+1}^{i-1})$ 表示 $w_{i-n+1} w_{i-n+2} \cdots w_{i-1}$ 后出现 w_i 的概率。在计算 $p(s)$ 之后，根据公式(2)可得句子的 n-gram 平均概率：

$$\overline{p(s)} = \sqrt[Count(n-gram)]{p(s)} \tag{2}$$

在公式 2 中，$\overline{p(s)}$ 表示句子的 n-gram 平均概率，$p(s)$ 表示句子 s 的概率，*Count(n-gram)* 表示句子 s 中所包含 n-gram 的个数，这里的 n 为公式 1 中计算句子概率所使用的 n 值。

根据本文提案方法，计算源语言句子和系统译文的 n-gram 平均概率后，通过比较 n-gram 平均概率对系统译文质量进行评价。本文使用公式 3，计算系统译文与源语言句子的 n-gram 平均概率的相对误差作为系统译文的评分：

$$score = \left| \frac{\overline{p_t} - \overline{p_s}}{\overline{p_s}} \right| \tag{3}$$

在公式 3 中，*score* 表示系统译文的评分，$\overline{p_t}$ 表示系统译文的 n-gram 平均概率，$\overline{p_s}$ 表示源语言句子的 n-gram 平均概率。该评分越小，表明系统译文翻译质量评价越高。

3.2 句子长度惩罚

一个源语言句子通常对应着多个长度不等的译文句子。若某个“取巧”的系统译文仅保留了那些翻译评价高的 n-gram，其评分将会变得很高。考虑以下情况：

源语言句子：党指挥枪是党的行动指南。

候选译文 1: It is a guide

候选译文 2: It is a guide to action that ensures that the military will forever heed Party commands.

根据本文之前所述方法进行评分，候选译文 1 的评分将会高于候选译文 2。这明显是错误的。事实上，某些长度过长或过短的系统译文会翻译那些比较确定的 n-gram，从而获得较高的评分。因此，需要考虑句子长度带来的影响。

大多数情况下，一种语言的长句翻译为另一种语言后也是长句，短句翻译为另一种语言后也是短句。借鉴句对齐技术中的句子长度比模型[Church, 1993]，本文给定一个区间 [*minLenRatio*, *maxLenRatio*]，如果句子长度比 $\frac{|t|}{|s|} \notin$ [*minLenRatio*, *maxLenRatio*]（其中$|t|$为系统译文句子长度，$|s|$为源语言句子长度），则对该系统译文进行惩罚。具体地，通过修正系统译文的 n-gram 平均概率，加大系统译文与源语言句子的 n-gram 平均概率之间的相对误差。

句子长度比模型假定 $\frac{|t|}{|s|}$ 服从正态分布，其期望 μ 和方差 σ^2 可从双语平行语料库中统计得到。本文中令[*minLenRatio*, *maxLenRatio*]=[$\mu - \sigma$, $\mu + \sigma$]。根据公式 4 计算系统译文的长度惩罚因子 *BP*：

$$BP = \begin{cases} e^{(1-\frac{|s|\cdot\mu}{|t|})} & |t| < |s| \cdot minLenRatio \\ 1 & |s| \cdot minLenRatio \le |t| \le |s| \cdot maxLenRatio \\ e^{(1-\frac{|t|}{|s|\cdot\mu})} & |t| > |s| \cdot maxLenRatio \end{cases} \tag{4}$$

在公式 4 中，*BP* 表示系统译文长度惩罚因子，$BP \in (0,1]$，$|t|$表示系统译文 t 的 n-gram 个数，$|s|$表示源语言句子 s 的 n-gram 个数。根据公式 5，使用长度惩罚因子 *BP* 对系统译文的 n-gram 平均概率作惩罚。

$$\overline{p_t}' = \begin{cases} \overline{p_t} / BP & \overline{p_t} \geq \overline{p_s} \\ \overline{p_t} \cdot BP & \overline{p_t} < \overline{p_s} \end{cases} \tag{5}$$

在公式 5 中，$\overline{p(t)}'$ 表示惩罚后的系统译文 n-gram 平均概率，$\overline{p(t)}$ 表示根据公式 2 计算的惩罚前的系统译文 n-gram 平均概率，$\overline{p(s)}$ 表示源语言句子的 n-gram 平均概率。根据公式 6 计算得到系统译文的最终评分。

$$score = \left| \frac{\overline{p_t}' - \overline{p_s}}{\overline{p_s}} \right| \tag{6}$$

3.3 篇章级得分

本文提案方法的基本计算单元是句子。但更多情况下需要评价的是整个篇章或文本的翻译质量。针对包含多个句子的篇章，本文提案方法首先计算各句子的 n-gram 平均概率，再根据公式 7，计算源语言与系统译文中所有句子的 n-gram 平均概率的和，使用 n-gram 平均概率和的相对误差，作为整个篇章的评分。

$$score(text) = \left| \frac{\sum_{i=1}^{n} \overline{p_t}' - \sum_{i=1}^{n} \overline{p_s}}{\sum_{i=1}^{n} \overline{p_s}} \right| \tag{7}$$

公式 7 中，$score(text)$ 表示整个系统译文篇章的评分，$\overline{p_t}'$ 表示系统译文的 n-gram 平均概率，$\overline{p_s}$ 表示源语言句子的 n-gram 平均概率，n 为篇章中句子的数量。

4 实验结果与分析

4.1 实验数据

为了验证本文提案方法的有效性，本文进行了法英翻译和德英翻译的实验验证，所使用数据来源于欧洲议会平行语料库[2]。

本文实验随机从平行语料库中抽取 2000 句对作为开放测试集(T3)，剩余部分作训练集，用于建立语言模型。并随机从训练集中抽取 2 组各 2000 句对作为封闭测试集(T1, T2)。具体实验数据如表 1 所示。

表 1. 实验数据

		法英	德英
训练集		1,825,077 句对	1,739,154 句对
测试集	T1(封闭)	2,000 句对	2,000 句对
	T2(封闭)	2,000 句对	2,000 句对
	T3(开放)	2,000 句对	2,000 句对

4.2 句子长度比模型

句子长度可以定义为字节或单词的个数。本文实验中使用单词个数表示句长。由表 1 所示的双语平行语料库计算得到句子长度比的期望 μ 和方差 σ^2，如表 2 所示。

[2] http://www.statmt.org/europarl/

表 2. 句子长度比统计数据

	法英	德英
期望 μ	0.9239	1.0731
方差 σ^2	0.0442	0.0533

从法英平行语料库中计算得 μ=0.9239, σ^2=0.0442, [*minLenRatio*, *maxLenRatio*]=[0.7137, 1.1341]；从德英平行语料库中计算得 μ=1.0731，σ^2=0.0533，[*minLenRatio*, *maxLenRatio*]=[0.8589, 1.3045]。

4.3 实验过程

本文实验使用 bing 在线翻译系统[3]与 google 在线翻译系统[4]对测试集进行翻译，得到两个不同系统（分别命名为系统 S1，系统 S2）的系统译文。实验设计了两组对比方法对系统译文质量进行评价：方法 1 使用 trigram model 计算句子概率，方法 2 使用插值 class-based trigram model 计算句子概率。在计算出句子概率的基础上，两组方法均根据公式 2 计算源语言句子和系统译文的 n-gram(n=3)平均概率，使用公式 4，5 对系统译文的 n-gram(n=3)平均概率进行惩罚，最后根据公式 7 计算系统译文的篇章级评分。

4.4 实验结果

本文实验使用方法 1 和方法 2 对两个系统在不同测试集上的系统译文进行评价，并与其它自动评测方法的评测结果进行对比。

法英翻译评测实验结果如表 2 所示。其中，T1+S1 表示使用系统 S1 翻译测试集 T1 得到的系统译文，类似地，T1+S2 表示使用系统 S2 翻译测试集 T1 得到的系统译文。表格中的粗体表示评价更高的系统译文，注意本文自动评测方法是评分越低，译文质量越好。

表 3. 法英翻译评测实验结果

	T1+S1	T1+S2	T2+S1	T2+S2	T3+S1	T3+S2
Method1	**0.3882**	0.4254	**0.3956**	0.4052	**0.2721**	0.3223
Method2	**0.3150**	0.3688	**0.3239**	0.3412	**0.2134**	0.2860
BLEU	**0.3506**	0.3173	**0.3408**	0.3176	**0.3064**	0.2789
NIST	**8.2050**	8.0268	**8.0536**	7.9966	**7.6305**	7.4436
Meteor	**0.6639**	0.6468	**0.6584**	0.6459	**0.6272**	0.6105
TER	**0.4852**	0.4888	**0.4949**	0.4900	**0.5309**	0.5322

德英翻译评测实验结果如表 3 所示。

表 4. 德英翻译评测实验结果

	T1+S1	T1+S2	T2+S1	T2+S2	T3+S1	T3+S2
Method1	0.1069	**0.0398**	0.1210	**0.0888**	**0.1177**	0.1844
Method2	**0.1983**	0.2595	**0.1631**	0.1768	**0.2568**	0.2996
BLEU	**0.2985**	0.2953	**0.3004**	0.2948	**0.2733**	0.2667
NIST	7.5710	**7.6270**	7.6151	**7.6570**	7.1820	**7.1861**
Meteor	0.6313	**0.6324**	**0.6348**	0.6343	0.5959	**0.5980**
TER	0.5461	**0.5342**	0.5452	**0.5330**	0.5814	**0.5694**

4.5 讨论

如表 3 所示，针对法英翻译评测，本文提案方法评测结果与 BLEU、NIST 等自动评测方法评测结果保持了很好的一致性。在表 4 中，两种方法的评价结果不太一致，其中方法 2 的评测结果与 BLEU 保持了较好的一致性。同时，方法 2 的评测结果较方法 1 也更稳定。主要原因是插值 class-based n-gram model 相较 n-gram model 引入了类信息，在语言模型训

[3] http://www.microsofttranslator.com/
[4] http://translate.google.cn/

练集规模不足时计算的句子概率更准确。

表 4 中，BLEU 与 NIST 给出了相反的评价结果。其原因在于，与 BLEU 方法相比，NIST 除了调整同现单元的记分方法以外，还修改了长度惩罚因子。根据 NIST 研究人员的实验结果，NIST 在稳定性和可靠性方面都优于 BLEU，尤其是对于译文的忠实度评测，NIST 方法更接近于人工评测。

综上所述，本文提案方法在一定条件下能对系统译文做出正确的评价。另外，用于建立语言模型的语料库规模将直接影响句子概率的准确度，从而影响最终的评价结果。但是，针对特定领域，构建大规模单语语料库要比构建双语平行语料库容易得多。因此，可以通过使用大规模单语语料库构建语言模型，来提高本文提案方法评测结果的可靠性。

5 结论与下一步工作

本文提出了一种基于句子 n-gram 平均概率的无参考译文的机器翻译自动评测方法。提案方法考虑了句子长度带来的影响，同时提出了句子级与篇章级的不同评分计算方法。实验结果表明，本文提案方法对不同系统译文作出了有效评价，并与 BLEU、NIST 等自动评测方法结果保持很好的一致性。同时也注意到：

- 本文提案方法的有效性需要大规模语言模型支撑。小规模语言模型易导致数据稀疏问题，使得句子概率计算不精确，影响最终评价结果。
- 在语言模型规模较小时，插值 class-based n-gram model 可适当缓解数据稀疏问题。当语言模型训练文本足够大时，类(class)的标注通常会花费大量人力物力，另一方面此时大规模的 n-gram model 已具有较好的表现，完全可以用 n-gram model 代替。
- 在获悉篇章级的译文可信度时，可结合句子级评分，筛选出翻译较差的句子。
- 本文提案方法在语系相近的双语间评测效果较好。主要是因为相近的语言间 n-gram 的对应关系比较单一。

本文提出的自动评测方法的优势在于摆脱了参考译文的束缚，且算法简单快速。其主要贡献有两点。首先，由于评测结果不依赖于参考译文，可以为研究人员提供大规模测试集的评测结果，有效减小研究人员的工作量。其次，对普通的机器翻译用户而言，提案方法可以协助用户在无参考译文的情况下获悉翻译结果的可信度。因此，提案方法具有非常广阔的应用前景。

下一步，我们将在本文提案方法的基础上，使用大规模真实文本的语言模型进行实验验证，并研究导入句法结构和语义信息，进一步提高提案方法的可靠性；另外，本文只在语系相近的语言对（法英，德英）之间进行实验，未来将进一步考察验证不同语种之间（如中英）提案方法的有效性和稳定性。

参考文献

Banerjee Satanjeev and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *Proceedings of ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65-72.

Church, Kenneth Ward. 1993. Char_align: A Program for Aligning Parallel Texts at the Character Level. In: *Proceedings of ACL-93*, Columbus OH.

Coughlin Deborah. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. In: *Proceedings of MT Summit IX*, pages 63-70.

Gamon Michael, Anthony Aue, and Martine Smets. 2005. Sentence-level MT evaluation without reference translations: Beyond language modeling. In: *European Association for Machine Translation (EAMT) 2005*,

May.

Hiroshi Echizen-ya and Kenji Araki. 2007. Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum. In: *Proceedings of MT Summit XII*, pages 151-158.

Hiroshi Echizen-ya and Kenji Araki. 2010 Automatic Evaluation Method for Machine Translation using Noun-Phrase Chunking. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 108-117.

Leusch Gregor, Nicola Ueffing and Hermann Ney. 2003. A Novel String-to-String Distance Measure with Applications to Machine Translation. In: *Proceedings of MT Summit IX*, New Orleans, USA, pages 240-247.

Lin Chin-Yew and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In: *Proceedings of ACL 2004*, pages 606-613.

Mutton Andrew, Mark Dras, Stephen Wan and Robert Dale. 2007. GLEU: Automatic Evaluation of Sentence-Level Fluency. In: *Proceedings of ACL 2007*, pages 344-351.

NIST. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf.

Papineni Kishore, Salim Roukos, Todd Ward and WeiJing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311-318.

Popović Maja and Hermann Ney. 2007. Word Error Rates: Decomposition over POS classed and Applications for Error Analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 48-55.

Rapp Reinhard. 2009. The Back-translation Score: Automatic MT Evaluation at the Sentence Level without Reference Translations. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 133-136.

S.Albrecht Joshua and Rebecca Hwa. 2007. Regression for Sentence-Level MT Evaluation with Pseudo References. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 296-303.

Snover Matthew and Dorr Bonnie, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In: *Proceedings of the 7th Conference of AMTA*, pages 223-231.

Sonja Nießen, Franz Josef Och, Gregor Leusch and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

赵红梅，刘群．机器翻译及其评测技术简介．术语标准化与信息技术，2010(1).

宗成庆．统计自然语言处理[M]．北京：清华大学出版社，2008(5).