·人工智能及识别技术· 文章编号: 1000-3428(2017)12-0179-05 文献标志码: A 中图分类号: TP18

融合主题的汉越冶金领域统计机器翻译方法

周 珂,余正涛,高盛祥

(昆明理工大学 信息工程与自动化学院,昆明 650500)

摘 要: 为有效利用领域知识提高汉越冶金领域的机器翻译效果,利用术语与主题的分布关系约束术语的选择,提出融合主题的统计机器翻译方法。建立术语翻译模型,抽取短语概率表时将术语对源语言文档的主题信息融入到翻译概率表中,利用主题分布描述术语与主题之间的关系,同时使用冶金领域语料训练翻译模型,将 2 个模型融合到基于短语的汉越翻译系统解码过程中,指导选择出最符合领域特性的术语及译文。融合模型前后的对比实验结果表明,与基于短语的机器翻译方法相比,该方法可有效提高汉越冶金领域的翻译性能。

关键词: 冶金领域; 统计机器翻译; 主题模型; 术语; 联合翻译模型

中文引用格式: 周 珂,余正涛,高盛祥. 融合主题的汉越冶金领域统计机器翻译方法[J]. 计算机工程,2017,43(12):179-183.

英文引用格式: ZHOU Ke,YU Zhengtao,GAO Shengxiang. Statistical Machine Translation Method Integrating Topic for Chinese-Vietnamese Metallurgy Field[J]. Computer Engineering,2017,43(12):179-183.

Statistical Machine Translation Method Integrating Topic for Chinese-Vietnamese Metallurgy Field

ZHOU Ke,YU Zhengtao,GAO Shengxiang

(School of Information Engineering and Automation,Kunming University of Science and Technology,Kunming 650500,China)

【Abstract】 In order to effectively use domain knowledge to improve the effect of Chinese-Vietnamese metallurgy field translation,this paper proposes a method of statistical machine translation,which uses the terms and topics' distribution relation to constrain the choice of terms. It constructs term translation model in the process of phrase probability table extraction,the topic information of the source language document for the term pair is incorporated into the translation probability table,and the relationship between the term and the topic is described by topic distribution. Based on the translation model of metallurgy field corpus training,the two models are integrated into the decoding process of Chinese-Vietnamese phrase-based translation system to guide the selection of terms and translations that best match the domain characteristics. The contrastive experiments are carried out before and after the fusion model. The results show that,compared with the phrase-based machine translation method,the proposed method can increase translation performance in the Chinese-Vietnamese metallurgy field effectively.

【Key words】 metallurgy field; statistical machine translation; topic model; term; united translation model

DOI: 10.3969/j.issn.1000-3428.2017.12.033

0 概述

越南与我国在冶金领域交流密切,随着冶金资料数量与日剧增[1],人工翻译虽然译文质量高,但是速度较慢,难于满足文本日益增长的需要,而机器翻译是解决这个问题的有效手段。

在机器翻译研究方面,统计机器翻译先后经历了基于词、短语、句法的翻译方法[2-4],已成为当前机器翻译主流的翻译方法。在汉英特定领域的机器翻译方面,专利领域、文献领域等都有较有效的翻译方法。为改善专利文献的机器翻译效果,文献[5]提出了一种基于模板的机器翻译方法,通过分析汉英双语对齐的专利文献语料,人工书写了 600 余条模板,将其与已有的一个基于规则的机器翻译系统相融合。文献[6]针对专利文本翻译中的复杂语句,提出了一种基于混合策略的方法,融合语义分析技术和基于规则的翻译技术,重点实现句子中心动词识别和句子中有嵌套结构存在的名称短语的分析。目前国际上针对越南语翻译集中在以印欧语言为源语的研究,如英越、法越、Google

基金项目: 国家自然科学基金(61672271,61761026,61732005);昆明理工大学校人才培养项目(省级)(KKSY201703005)。

作者简介: 周 珂(1991—),男,硕士研究生,主研方向为机器翻译; 余正涛,教授、博士生导师; 高盛祥(通信作者),博士。

收稿日期: 2016-11-14 修回日期: 2016-12-17 E-mail: gaoshengxiang.yn@foxmail.com

在线翻译等机译系统，在汉越的领域机器翻译方面，研究工作还处于起步阶段。

冶金领域中存在大量的冶金术语，这些领域术语[7]在汉语、越南语中都存在固定搭配和习惯用语特点。由于汉越双语语料稀缺，而且语料领域混杂，因此利用机器翻译方法进行冶金领域的文本翻译时，术语翻译往往不准确。考虑到术语的翻译经常会与原文本中的领域信息相关，本文提出融合领域知识的统计机器翻译方法。

1 研究背景

表 1 展示了 2 个冶金领域术语在现有机器翻译系统上与人工翻译的对比。其中“发汗”在冶金领域中是坯加热处理时液相渗出的一种现象，应该翻译为“Đổ mồ hôi”，而不是健康领域中“出汗”(nước sinh hoạt) 的意思；“起泡”在冶金领域中是由于气体剧烈排出，在烧结件表面形成鼓泡的一种现象，而不应翻译为“冒泡”(bọt nước)。

表 1 冶金领域汉语文本的越南语机器翻译与人工翻译对比

汉语文本	机器翻译	人工翻译
金属发汗材料是一种特殊的散热材料	nước sinh hoạt vật liệu kim loại là một loại vật liệu đặc biệt làm mát	Đổ mồ hôi vật liệu kim loại là một loại vật liệu đặc biệt làm mát
锌合金压铸件起泡的原因	Kẽm hợp kim Die Casting Nguyên nhân bọt nước	Kẽm hợp kim Die Casting Nguyên nhân phồng rộp

现有的基于短语的翻译系统[8-9]在处理这些语言现象时未考虑领域信息，导致对这些领域的专业术语翻译不准确。针对此问题，本文建立术语翻译模型和联合翻译模型。术语翻译模型是把术语的翻译概率与它所在文档的主题分布绑定到一起，从而在术语翻译时指导选择符合主题的译文；联合模型是将通用领域语料训练的翻译模型和冶金领域语料训练的模型线性融合到一起，因为通用领域语料比较丰富，训练的翻译模型可以用来做统计机器翻译的基础模型，冶金领域语料比较稀缺，训练的翻译模型不能直接用来做通用的文本翻译，但是可以将它作为冶金领域的领域知识，而本文就是利用已有小规模的特定领域语料库，重新调整翻译模型中的知识分布，达到使通用机器翻译具有领域性的目的。

本文通过冶金领域和通用领域的平行语料来训练联合模型，利用已有的冶金术语库来训练术语翻译模型，并将 2 种模型融合到 Moses 统计机器翻译系统中。

2 融合主题信息的翻译模型

2.1 主题模型

主题模型是对文字中隐含主题的一种建模方法。LDA 是现在用的最普遍的主题模型[10]，因此，本文用它来挖掘主题信息。

LDA 是一种非监督学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息。它采用了词袋的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为易于建模的数字信息。但是词袋方法未考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。假设一个文本中有 T 个主题，第 i 个词汇 ω_i 和主题的关系表示为式(1)。

$$p(\omega_i) = \sum_{j=1}^{T} \mathrm{P}(\omega_i | z_i = j)\, \mathrm{P}(z_i = j) \tag{1}$$

其中：z_i 是潜在的随机变量，表示第 i 个词汇标记 ω_i 取自这个主题；$\mathrm{P}(\omega_i | z_i = j)$ 是词汇 ω_i 在主题 j 的概率；$\mathrm{P}(z_i = j)$ 是 j 属于当前文本主题的概率。

假设由以 w 个词汇组成 T 个主题并形成 D 个文本，令 $\phi^{d}_{z=j} = p(z = j)$ 表示主题 j 词汇上的多项分布，其中 w 是 W 个唯一的词汇。令 $\psi^{d}_{z=j} = \mathrm{P}(z = j)$ 为对于文本 d，T 个主题上的多项分布，因此文本 d 中词汇 w 的概率为：

$$p(\omega | d) = \sum_{j=1}^{T} \phi^{z=j}_{\omega} \cdot \psi^{d}_{z=j} \tag{2}$$

LDA 模型的先验概率在 $\psi(d)$ 上作的 $Dirichlet(\alpha)$ 假设，在 $\psi(d)$ 上作对称的 $Dirichlet(\alpha)$ 的先验概率假设，并且 $\phi(z)$ 也作了对称的 $dirichlet(x)$ 先验概率的假设，如式(3)所示。

$$\begin{aligned} &\omega_i | z_i, \phi^{(z_i)} \sim Discrete(\phi^{(z_i)}), \phi^{(z_i)} \sim Dirichlet(x) \\ &z_i | \psi^{(d_i)} \sim Discrete(\phi^{(d_i)}), \psi^{(d_i)} \sim Dirichlet(\alpha) \end{aligned} \tag{3}$$

其中：x 表示抽取词汇标记之前根据主题抽样获得词频数；α 表示使用任何文档文字前主题被抽样的频数。

该模型有 2 个参数需要推断“文档-主题”分布和“主题-单词”分布。通过学习这 2 个参数，可以知道表达文档主旨的主题，以及每篇文档所涵盖的主题比例等。推断方法主要有变分-EM 算法，还有现在常用的 Gibbs 抽样法。LDA 模型自从诞生之后有了较多扩展，特别是在社会网络和社会媒体研究领域最为常见，现在已经成为了主题建模中的一个标准。

2.2 翻译模型

在冶金领域的翻译中，术语的翻译可能会产生歧义，术语的翻译应该与它所在文档的主题分布相

关。消除歧义最直接的方法就是计算术语在已给领域信息的条件翻译概率，本文使用一个文档的主题分布来表示文档的领域信息。

本文借助已有的汉越双语术语库，来计算源语言术语到目标语言术语的翻译概率，其主题分布为源语言术语在它所在文档中的分布。为每一术语对设一个 K 维向量，第 k 元 $p(\hat{e}|\hat{f},z=k)$ 表示主题为 k 的条件下源语言短语 $\hat{f}$ 到目标语言 $\hat{e}$ 的翻译概率。

对于每一个双语术语对 $<\hat{f},\hat{e}>$，可以通过 LDA 模型推断出源语言术语 $\hat{f}$ 在其文档 D 中的主题分布 $p(z|D)$。由以上办法，可以收集到一个实例集合 $\Gamma=\{(\hat{f},\hat{e},p(z|D))\}$，其中的双语术语对都有不同的文档主题分布。使用这些实例，可以由下式计算出 $p(\hat{e}|\hat{f},z=k)$ 概率:

$$p(\hat{e}|\hat{f},z=k)=\frac{\sum_{i\in\Gamma,i.\hat{e}=\hat{e},i.\hat{f}=\hat{f}}p(z|D)}{\sum_{i\in\Gamma,i.\hat{f}=\hat{f}}p(z|D)} \tag{4}$$

通过表 2 举例说明在不同主题条件下同一术语的翻译概率，可以清楚地看到源语言术语“发汗”在不同主题下翻译结果是不一样的，比如在“生物学”领域，它会被翻译成“nước sinh hoạt”，但是在冶金领域，它往往会被翻译成“Đổ mồ hôi”。

表 2　带有主题分布的术语对翻译概率举例

源语言	目标语言	$p(\hat{e}\|\hat{f},z=k)$	主题
发汗	nước sinh hoạt	0.009 7	1
	Đổ mồ hôi	0.850 0	2
	nước sinh hoạt	0.120 0	3
起泡	bọt nước	0.060 0	1
	phập phòng	0.340 0	2
	phồng rộp	0.730 0	3

本文将每个提取的双语术语与其相应的主题的条件翻译概率通过式(6)联系在一起，当翻译文档 D' 中的句子时，首先通过 LDA 模型得到它的主题分布，若待翻译句子中含有在文档 D' 中的术语，术语翻译模型表示为:

$$Term(\hat{e}|\hat{f})=\sum_{k=1}^{K}p(\hat{e}|\hat{f},z=k)\times p(z=k|D') \tag{5}$$

在翻译一个术语时，本文通过计算在文档 D' 下术语 $\hat{f}$ 到 $\hat{e}$ 的条件概率，来帮助解码器选择符合领域的术语翻译。

3　翻译模型的融合

为充分利用特定领域翻译知识，本文在将融入主题信息的翻译模型融合到机器翻译系统的同时，也将冶金领域语料训练的翻译模型线性融合到翻译系统中，进一步提升基准翻译模型的冶金领域适应能力。本文利用对数线性模型[11]将这 2 个模型作为 2 个特征融合到汉越基于短语的翻译模型中[12]，这样不会影响集成本身其他特征的计算。对于解码过程，本文设计思路框架如图 1 所示，图中显示了融合了冶金领域语料训练的翻译模型和融合主题信息的翻译模型的统计机器翻译的整个过程。

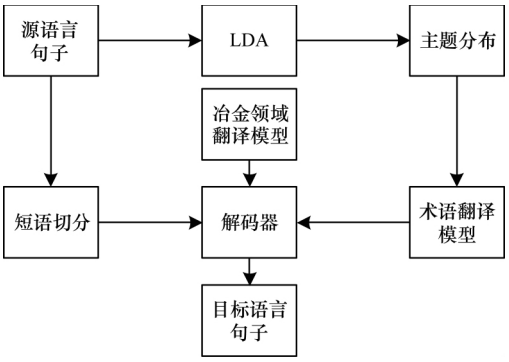

图 1　融合翻译模型的汉越机器翻译系统流程

基于短语的翻译模型对数线性表示可以用式(6)表示。

$$\begin{aligned}P(e|f)=\exp[&\lambda_{\phi}\mathrm{lb}\,\phi(f|e)+\lambda_{d}\mathrm{lb}\,d(a,b)\\&+\lambda_{LM}\mathrm{lb}\,p_{lm}(e)]\end{aligned} \tag{6}$$

其中: $\mathrm{lb}\,\phi(f|e)$ 是翻译模型; $\mathrm{lb}\,d(a,b)$ 是调序模型; $\mathrm{lb}\,p_{lm}(e)$ 是语言模型。这些特征函数不同的权重，影响着翻译的质量。

将上文提到的 2 个模型作为特征函数融合到这个短语翻译系统中，融合后的对数线性表示为式(7)。

$$\begin{aligned}P(e|f)=\exp[&\lambda_{\phi}\mathrm{lb}\,\phi(e|f)+\lambda_{d}\mathrm{lb}\,d(a,b)\\&+\lambda_{LM}\mathrm{lb}\,p_{lm}(e)+\lambda_{\mathrm{metallurgy}}\mathrm{lb}(\mathrm{metallurgy})\\&+\lambda_{\mathrm{term}}\mathrm{lb}(term)]\\\lambda_{\phi}+\lambda_{\mathrm{metallurgy}}=1\end{aligned} \tag{7}$$

其中: $\lambda_{\mathrm{metallurgy}}$ 和 λ_{term} 分别是冶金领域语料训练的翻译模型和融合主题信息的术语翻译模型的权重[13]; λ_{ϕ} 为通用领域翻译模型的权重，本文令通用领域翻译模型与冶金领域翻译模型的权重之和为 1，然后通过最小错误率训练方法来调优[14]这些参数，使得翻译结果取得最优(根据翻译评测目标评价该结果译文)。在解码阶段，搜索选择概率最高的译文为最佳候选译文。

4　实验与结果分析

4.1　实验数据

为进行汉越冶金领域机器翻译方法的评测，冶金领域语料采用收集整理的汉越冶金领域平行语料，规模为 5×10^3，而通用领域训练语料是从 5×10^4

汉越双语语料库中抽取的 3×10^4 双语句对，通用领域语料库覆盖多个领域，包含较为丰富的翻译知识。汉越双语术语库是在汉越双语语料库中抽取的 5×10^3 术语对。实验中用到的双语语料的详细统计信息如表 3 所示，对其采用相同的分词工具进行预处理。

表 3 汉越双语语料统计信息 10^3

领域	句对数		词数	
	汉语	越南语	汉语	越南语
通用领域	30	30	162	175
冶金领域	5	5	25	34

4.2 实验设置

实验采用开源机器翻译软件 Moses，搭建基于短语模型的机器翻译系统。该工具融合Giza ++工具实现词对齐，并利用最小错误率训练方法优化翻译系统中的模型参数。针对通用领域平行句对和特定领域相关性评价，本文采用 LDA 主题模型计算文档的主题分布，同时利用Giza ++工具获取通用领域平行句对的词对齐信息，并基于词对齐信息获取词汇翻译表，最终用于估计带有主题分布的短语概率表。此外，本文利用 Moses 提供的语言模型训练脚本训练4-gram语言模型，用于机器翻译系统参数调节和解码过程。本文采用 Mose[15] 内嵌的短语翻译引擎构建冶金领域汉越翻译系统，并以大小写不敏感的BLEU-4作为评价指标[16]。为了验证模型融合方法的有效性，本文设置以下 4 个机器翻译系统进行实验。

Baseline1：利用冶金领域约 5×10^3 句对训练翻译系统。

Baseline2：从通用领域平行语料库中抽取3×10^4平行句对训练翻译系统。

Baseline2 + ME：将冶金领域翻译模型融合到机器翻译系统。

Baseline2 + ME + TM：将冶金领域翻译模型和术语翻译模型融合到机器翻译系统中。

4.3 结果分析

4.3.1 主题数量

因为将主题信息融入到了术语翻译模型中，所以必须探索主题数量对术语翻译模型的影响。本文设置主题数量 k 取值范围为 10 ~ 25，得到的实验结果如表 4 所示。可以看出，当 k 取 10 ~ 15 时，模型的 BLEU-4 值是随主题数量递增而递增的，而当 k 取15 ~ 25之间时，BLEU-4 值随主题数量递增而递减，这表明本文的术语翻译模型在主题数量取 15 时效果最好。

表 4 不同主题数量下翻译模型的 BLEU-4 值变化 %

主题数量	BLEU-4 值
10	24.55
15	24.83
20	24.75
25	24.62

针对此类情况，本文假设一种最有可能发生的情况，比如当发现训练数据中术语有 10 个不同主题时，设置 $k=10$；有 15 个不同主题时，$k=15$；有 16 个不同主题时，$k=20$；有 17 个不同主题时，$k=25$。此时，数据稀疏问题就变得很严重了，当 $k=25$ 时，术语不会出现在 32% 的主题中，而当 $k=20$ 时，术语不会出现在 20% 的主题中。所以，$k=15$ 是在有效的主题信息和数据稀疏问题之间最好的平衡。因此，后续实验将主题数目都设置为 $k=15$。

4.3.2 性能比较

表 5 给出了不同的系统翻译性能。其中，Baseline1 和 Baseline2 区别在于训练翻译模型的平行语料来自不同的领域。Baseline1 系统的训练语料属于冶金领域，因此具备较高的翻译性能。

表 5 系统翻译性能 %

系统	训练语料	BLEU-4 值
Baseline1	冶金领域(5×10^3)	22.67
Baseline2	通用领域(3×10^4)	20.41
Baseline2 + UN	通用领域 + 冶金领域	25.81
Baseline2 + UN + TM	通用领域 + 冶金领域	26.03

对比 Baseline1 和 Baseline2 系统可以发现：当训练语料和目标翻译任务待测文本所属领域同是冶金领域时，往往能够取得较好的性能。原因在于，来自特定领域的待测文本往往含有此领域特有的表述方式及专有的术语，当其他领域翻译知识无法匹配时，导致其翻译系统译文质量下降。然而，冶金领域平行语料一般难以获得，而通用领域双语资源由于混杂各个领域的文本被广泛使用。Baseline2 + ME、Baseline2 + ME + TM 分别是融合相应的模型后的翻译系统，与 Baseline2 相比，分别提升了 5.40 和5.62 个BLEU 值点。可以证明，在为术语翻译过程中引入主题信息后，在术语的翻译过程中可以有效地利用主题信息来选择符合领域特性的译文，而且使用冶金领域双语资源能够有效地调整通用领域翻译模型知识分布，进而在翻译过程中，最大程度利用领域翻译知识，提升冶金领域翻译任务的译文质量。

图 2 给出了冶金领域翻译模型在不同权重下的性能变化。可以看出，当冶金领域翻译模型取得合理权重时，其系统翻译性能可以得到进一步提升，此

外，分析发现，加大冶金领域翻译模型权重未必会提升测试集的翻译质量，这是因为特定领域翻译模型往往缺乏充足的训练语料，翻译模型训练过程会遇到数据稀疏问题，由此导致冶金领域翻译知识不准确，最终降低系统翻译性能。

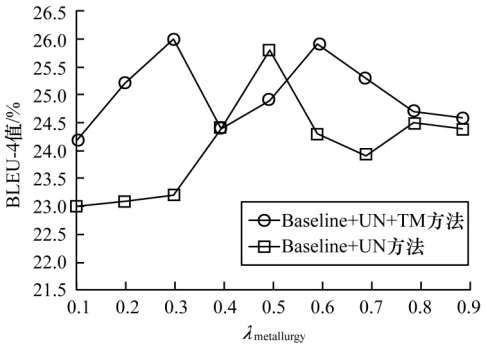

图 2　不同权重下的系统翻译性能

5　结束语

本文提出融合主题的汉越统计机器翻译方法，构建术语翻译模型和联合翻译模型。术语翻译模型是利用术语与主题的分布关系来约束术语的选择，而联合模型可以有效地调整基准翻译模型知识分布，进而在翻译过程中最大程度利用领域翻译知识。本文将这 2 个模型融合到基准翻译系统中，实验结果表明，相比改进前的翻译模型，利用该模型进行冶金领域文本翻译时在 BLEU-4 和准确度方面都有较大提升。下一步将扩大语料库规模，并对冶金领域的特征做深入分析。

参考文献

[1] 张　军. 我国西南地区在“一带一路”开放战略中的优势及定位[J]. 经济纵横，2014(11)：93-96.

[2] 刘　群. 统计机器翻译综述[J]. 中文信息学报，2003，17(4)：1-12.

[3] KOEHN P，OCH F J，MARCU D. Statistical Phrase-based Translation [C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Michigan，USA：Association for Computational Linguistics，2003：127-133.

[4] YAMADA K. A Syntax-based Translation Model [C]//Proceedings of Meeting of the Association for Computational Linguistics. Michigan，USA：Association for Computational Linguistics，2002：1-5.

[5] 张冬梅，刘小蝶，晋耀红. 基于模板的汉英专利机器翻译研究[J]. 计算机应用研究，2013，30(7)：2044-2046.

[6] 晋耀红. 一种混合策略的专利机器翻译系统研究[J]. 计算机工程与应用，2012，48(4)：29-32.

[7] MENG F，XIONG D，JIANG W，et al. Modeling Term Translation for Document-informed Machine Translation [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Doha，Qatar：[s.n.]，2014：546-556.

[8] KOEHN P，OCH F J，MARCU D. Statistical Phrase-based Translation [C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Michigan，USA：Association for Computational Linguistics，2002：48-54.

[9] CHIANG D. A Hierarchical Phrase-based Model for Statistical Machine Translation [C]//Proceedings of Meeting of the Association for Computational Linguistics. Michigan，USA：Association for Computational Linguistics，2005：263-270.

[10] BLEI D M，NG A Y，JORDAN M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research，2003，3：993-1022.

[11] DARROCH J N，RATCLIFF D. Generalized Iterative Scaling for Log-linear Models [J]. Annals of Mathematical Statistics，1972，43(5)：1470-1480.

[12] KOEHN P，SCHROEDER J. Experiments in Domain Adaptation for Statistical Machine Translation [C]//Proceedings of the Workshop on Statistical Machine Translation. Michigan，USA：Association for Computational Linguistics，2007：224-227.

[13] BROWN P F，PIETRA S A D P，PIETRA V J D. The Mathematics of Statistical Machine Translation：Parameter Estimation [J]. Computational Linguistics，1993，19(2)：263-311.

[14] OCH F J. Minimum Error Rate Training in Statistical Machine Translation [C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Michigan，USA：Association for Computational Linguistics，2003：160-167.

[15] KOEHN P，HOANG H，BIRCH A，et al. Moses：Open Source Toolkit for Statistical Machine Translation [C]//Proceedings of Meeting of the ACL on Interactive Poster and Demonstration Sessions. Michigan，USA：Association for Computational Linguistics，2007：177-180.

[16] PAPINENI K，ROUKOS S，WARD T，et al. BLEU：A Method for Automatic Evaluation of Machine Translation [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Michigan，USA：Association for Computational Linguistics，2002：311-318.

编辑　金胡考