文章编号：1003-0077(2010)06-0050-07

利用上下文信息的统计机器翻译领域自适应

曹 杰，吕雅娟，苏劲松，刘 群

（中国科学院 计算技术研究所，中国科学院 智能信息处理重点实验室，北京 100190）

摘 要：统计机器翻译系统用于翻译领域文本时，常常会遇到跨领域的问题：当待翻译文本与训练语料来自同一领域时，通常会得到较好的翻译效果；当领域差别较大时，翻译质量会明显下降。某个特定领域的双语平行语料是有限的，相对来说，领域混杂的平行语料和特定领域的单语文本更容易获得。该文充分利用这一特点，提出了一种包含领域信息的翻译概率计算模型，该模型联合使用混合领域双语和特定领域源语言单语进行机器翻译领域自适应。实验显示，自适应模型在 IWSLT 机器翻译评测 3 个测试集上均比 Baseline 有提高，证明了该文方法的有效性。

关键词：统计机器翻译；领域自适应；上下文信息

中图分类号：TP391 **文献标识码**：A

SMT Domain Adaptation Based on Monolingual Context Information

CAO Jie, LV Yajuan, SU Jinsong, LIU Qun

(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Domain adaptation problem will arise when statistical machine translation (SMT) system is used to translate domain-specific texts. When the texts to be translated and the training data come from the same domain, SMT system can achieve good performance. Otherwise, the translation quality will degrade dramatically. In general, domain-specific parallel corpus is limited, while domain-mixed parallel corpus and domain-specific monolingual corpus are easy to obtain. According to the fact, this paper proposed a new translation model which utilized domain-mixed parallel corpus and domain-specific monolingual corpus to improve the domain translation quality. Experiments show that the proposed method improves translation performance in three IWSLT evaluation tests significantly.

Key words: statistical machine translation; domain adaptation; context information

1 引言

近年来，统计机器翻译研究得到了迅猛的发展，提出了很多新的模型和方法并取得了很好的效果[1-3]。一些先进的统计机器翻译系统已经得到了实际应用，如 Google 的在线翻译和跨语言信息检索系统。统计机器翻译的实用价值逐渐得到体现。

当前主流的统计机器翻译系统都需要在大规模的双语语料库上进行训练得到翻译模型和语言模型。训练得到的模型在翻译同一领域的文本时通常会得到质量较高的译文，但翻译其他领域文本时，翻译质量明显下降。对于某些领域而言，获取大规模的平行语料是非常困难的。研究有效的领域自适应策略是一个可行办法。

一般说来，获取某个领域的单语语料库比双语平行的语料库要容易很多，而不同领域的单语上下文中包含着与领域相关的信息，有效利用这些领域信息会对统计机器翻译自适应研究有所帮助。

本文提出了一种领域特征计算方法，领域特征的计算中引入了单语上下文信息。相比于基于短语的模型而言，能够利用词性信息、长距离上下文等更

收稿日期：2009-12-21 **定稿日期**：2010-04-19

基金项目：国家自然科学基金资助项目(60873167)

作者简介：曹杰(1981—)，男，硕士生，主要研究方向为自然语言处理；吕雅娟(1972—)，女，博士，副研究员，主要研究方向为自然语言处理、机器翻译；苏劲松(1982—)，男，博士生，主要研究方向为自然语言处理、机器翻译。

丰富的上下文信息。

2　相关工作介绍

领域自适应问题在语音识别领域已有较多研究，但由于机器翻译问题的复杂性，在机器翻译模型尚不完善的阶段，领域自适应的研究较少。随着近几年统计机器翻译模型的不断完善，越来越多的学者注意到领域自适应问题在机器翻译中的重要性，相关研究也开始增多。

目前在机器翻译领域，自适应的研究按照自适应对象模型的不同可以分为翻译模型自适应和语言模型自适应。

对语言模型自适应的研究思路基本上都是采取构建信息检索模型，从单语语料库中检索与待翻译领域相似的句子，用这些句子构建自适应的语言模型以提高翻译效果。语言模型自适应先是被应用于语音识别领域，并取得了一定的效果，Eck et al. 和 Zhao et al. 等人将这一思想引入统计机器翻译领域[4-5]，将首次翻译的得到的候选翻译结果视为信息检索模型中的查询，在海量的单语语料库中检索出相似的数据，根据检索数据训练得到自适应的语言模型，可以明显的提高统计机器翻译的质量。

翻译模型的自适应研究中，Hildebrand et al. 提出一种方法，从双语平行语料库中检索与测试集相似的句子，在检索返回的句子上训练自适应的翻译模型，将自适应的翻译模型与原翻译模型联合使用将会提高翻译质量[6]。Ueffing et al. 提出一种在机器翻译中使用半监督学习的算法[7]：首先利用双语语料库训练一个初始的翻译系统，然后对 in-domain 的源语言单语进行翻译并对翻译结果进行打分，选择分数较高的译文与源语言单语构成人工构造的双语语料库，将构造出的双语语料库与原平行语料库合并进行训练，重复该过程，直到到达一定的轮数。这时候得到的增强的模型翻译效果比初始翻译系统要好。Yajuan Lü et al.[8] 提出通过离线的数据选择和在线的模型优化的策略进行翻译模型的自适应。本质上看，模型优化是对多个短语表的插值使用，插值的系数由检索到的隶属子模型的句子在整个检索结果集中的比例决定。

综上所述，当前统计机器翻译领域自适应的研究主要集中在利用信息检索工具或者半监督等学习方法扩大训练集规模上，而对单语信息的利用并不充分。目前对单语的使用方式主要有以下两种：一是作为查询条件从双语语料库中检索相似句子作为自适应训练集，另外一种是用于半监督学习，通过初始系统对源语言翻译，然后选择较好的译文得到人工构造的双语，与原有的训练数据一起训练新的模型。

我们认为以上两种使用单语的方法没有充分挖掘单语内部包含的领域信息，本文提出一种有效利用单语上下文信息引入领域相关(Domain-specific)特征的方法。主要包括两步：一是从领域混杂的语料库中检索出与待翻译文本领域上接近的平行句对以扩大训练集规模，一是挖掘该领域的单语上下文信息，作为新特征引入对数模型框架内，使得与该领域相关的短语译文更有可能在机器翻译解码过程中被选择到。

3　利用上下文信息的领域自适应

3.1　基本思想介绍

领域自适应问题研究中，领域的表示是一个重要问题。我们认为，领域单语中包含着领域信息，上下文信息可以认为是表示领域的一个重要特征。如果能有效的融合领域信息和翻译模型，那么对领域自适应的研究将是很有帮助的。下面的例子显示了单语上下文信息对机器翻译的帮助。

我们要翻译关于功夫电影的文本，但我们只有经济领域的双语语料来训练翻译系统，此外，我们还有大量关于功夫电影领域的单语文本。假设语料如表 1 所示，词对齐后，中国分别对齐到 China 和 Chinese。

表 1　包含“中国”一词的平行句对和单语句子

经济领域中包含“中国”一词的平行句对	中国 经济 发展 迅速 The economy of China develops fast 中国 经济 迅速 崛起 The economy of China grows up rapidly 中国 人 富裕 起来 Chinese people becomes richness
功夫电影领域包含“中国”一词的源语言单语	中国 人 中 懂 功夫 的 很多 中国 人 李小龙 将 功夫 推向 世界 很多 人 喜欢 中国 功夫

在经济领域语料中，不考虑短语扩展的情况下，“中国”一词的翻译概率如表 2 所示。

我们要翻译来自功夫电影领域的汉语句子“我喜欢 中国 功夫”，由经济领域的平行语料训练的翻

表2 “中国”一词的翻译概率

原文	译文	同现次数 Count(c,e)	翻译概率 P(e|c)
中国	Chinese	1	1/3
	China	2	2/3

译系统，将“中国”翻译为 China 的概率要大于 Chinese，翻译结果可能是“I love China Gongfu”。

在功夫电影领域的源语言单语中，“中国”一词出现时，后面经常接“人”这个词，这就提供了有用的领域信息：在功夫的电影领域，“中国”后面经常接“人”这个词。本文提出的方法可以将上下文信息转化为领域特征引入翻译模型中，提高“中国”翻译为 Chinese 的概率，在翻译“我 喜欢 中国 功夫”一句时，可以翻译得到“I love Chinese Gongfu”。翻译效果要好于直接使用经济领域双语训练出的模型。

3.2 领域特征介绍

统计机器翻译的对数线性模型中，翻译的过程被建模为寻找最大概率译文 e_{best} 的过程：

$$e_{best} = \arg\max_e p(e \mid f) \approx \arg\max_e \sum_e \lambda_m h_m(e,f) \quad (1)$$

其中，$h_1(e,f)\cdots h_m(e,f)$ 是建立在源语言 f 和目标语言 e 上的 m 个特征函数，$\lambda_1 \ldots \lambda_m$ 是其对应的特征值。对数线性模型中可以方便的扩充新的特征，在此，我们引入带上下文信息的领域翻译概率 $P_D(e|f)$，其计算公式为：

$$P_D(e \mid f) = \sum_{context} P_D(e, context \mid f) = \sum_{context} P_D(e \mid f, context) P_D(context \mid f) \quad (2)$$

在 $P_D(e|f)$ 的计算过程中，我们引入了隐变量 $context$，代表上下文特征。其中，$P_D(context|f)$ 可以从领域 D 的单语计算得到，代表了一定的领域信息。对于 $P_D(e|f,context)$，缺乏 D 的双语语料，无法准确计算，我们采用信息检索的办法从大规模的混合领域双语中检索出与领域 D 接近的语料作为双语训练语料。检索得到的双语语料在领域上与 D 接近，从近似语料中计算的概率分布 $P_{D\text{-}similar}(e|f,context)$ 与领域 D 上的概率分布 $P_D(e|f,context)$ 比较接近。在此，我们用 $P_{D\text{-}similar}(e|f,context)$ 代替 $P_D(e|f,context)$，得公式(3)：

$$P_D(e|f) \approx \sum_{context} P_{D\text{-}similiar}(e \mid f, context) P_D(context \mid f) \quad (3)$$

总结起来，领域特征的计算过程可分为以下四步：1) 从领域单语中抽取领域相关的单语上下文信息。2) 检索出一批领域接近的双语语料作为新的训练语料 3) 从训练语料的双语词对齐结果中抽取带上下文信息的短语翻译对。4) 用 1)和 3)的结果计算领域特征。

下面以 3.1 节的例子为例，说明本方法起作用的原因。这里，我们取 Context 为下文第一个词，下面以 3.1 节中的“中国”一词为例，分别计算 $P_{D\text{-}similar}(e|f,context)$ 和 $P_D(context|f)$。在经济领域中，带有 Context 的翻译概率如表 3 所示。

表3 经济领域双语中“中国”下文词的统计信息

原文	Context 下文第一个词	译文	同现次数 count(c,e,context)	翻译概率 $P_{D\text{-}similar}(e\|f,context)$
中国	人	China	0	0
		Chinese	1	1
	经济	China	2	1
		Chinese	0	0

在功夫电影领域单语中，“中国”的下文词统计信息如表 4 所示。

取 Context 为下文第一个词时，根据表 3，4 的统计信息，使用功夫电影领域单语进行自适应后的领域特征的计算如表 5 所示。

表4 功夫电影领域“中国”下文词的统计信息

原文	Context 下文第一个词	同现次数 count(c,context)	翻译概率 $P_D(context\|f)$
中国	人	2	2/3
	功夫	1	1/3

表 5　领域特征的计算

原文	译文	Context（取下文第一个词）	$P_{D\text{-}similar}(e\|f,context)$	$P_D(context\|f)$	$P_D(e\|f)$	归一化
中国	Chinese	经济	0	0	$0*0+1*2/3=2/3$	1
		人	1	2/3		
	China	经济	1	0	$1*0+0*2/3=0$	0
		人	0	2/3		

表 5 中，“中国”翻译为 Chinese 领域特征值大于翻译为 China。加入新特征后的模型将“我 喜欢 中国 功夫”翻译为“I love Chinese Gongfu”，与正确译文更加接近。

上述推导过程可以看出本方法充分利用了单语体现出的“中国”后面经常接“人”这个词的特征。本例子用到的特征是源语言单语的下文第一个词，同理，其他上下文特征也能起类似作用。

基于短语的翻译模型本身具备一定的上下文翻译能力，但对于词性等上下文信息没有处理能力，本文的方法可以应用词、词性、长距离上下文等多种上下文信息，比短语内部包含的上下文要丰富很多。而且，这里的上下文来源于领域单语，这也是与短语模型上下文的区别。

理论上，任何能有效表示该领域上下文的特征都可以转化为领域特征融入对数线性模型，本文提出的模型在上下文特征的选择上具有很强的扩展性。

4　实验

4.1　实验设置

我们采用著名的开源工具 Moses① 作基线系统，所使用的特征如表 6 所示。

表 6　对数线性模型的特征

特征	短语模型特征（7 个）	正、反向翻译概率；正、反向词汇化翻译概率；语言模型；句子长度；短语个数
	领域特征（4 个）	上文第一个词；下文第一个词；上文第一个词的词性；下文第一个词的词性

语言模型训练工具采用 SRILM Toolkit[9]，评测工具使用 mteval-v11b. pl②，评测指标采用 BLEU4[10]，大小写不敏感。另外，使用了 Lemur 作为检索语料的工具。

在 IWSLT③ 评测的汉英翻译任务上进行实验，IWSLT 评测语料主要由面向旅游领域的口语对话组成，领域特征比较明显，适合进行领域自适应的研究。实验语料详见表 7。

4.2　实验结果

实验中，我们设置了三组 baseline：以 FBIS 做训练集的 Baseline1、以混合语料做训练集的 Baseline2、FBIS 与混合语料合并后的 Baseline3。采用本文第 3 节提出方法利用上下文信息进行领域自适应，具体做法分如下步骤：

1. 合并 FBIS 语料和领域混杂语料，记为 T，用信息检索工具 Lemur 在 T 上建索引。

2. 计算 T 中的每个句子与开发测试集合每个句子的相似度分数，并按照相似度分数对 T 进行排序。这一步耗时较长，尤其当训练数据规模较大时。数据选择的策略还有很多，不是本文研究的重点，这里我们采用了这种比较简单的方式。

3. 从 T（共 700k）中选取 topN（N = 100k，200k，300k，…）平行句对作为新的训练集，进行词对齐，并抽取带上下文特征的短语表，即公式(3)中的 $P_{D\text{-}similar}(e|f,context)$。

4. 根据领域单语语料，计算 $P_D(context|f)$。

5. 根据公式(3)，计算的得到领域特征 $P_D(e|f)$，重新训练并记录翻译 BLEU 值。

采用本文提出的方法进行领域自适应的实验结果如表 8 所示。

① http://www.statmt.org/moses/

② http://www.nist.gov/speech/tests/mt/resources/scoring.htm

③ http://www.is.cs.cmu.edu/iwslt2005/

表7 实验语料情况

语料类型	来 源	描 述	规模(句对)
双语语料	FBIS	FBIS Multilanguage Texts	200k
混合语料	BTEC	旅游口语领域	45k
	HK_News	香港新闻	200k
	2004-863-008	Chinese LDC 中的对话语料	51k
	HK_Hansards	香港议会记录	200k
领域单语语料	IWSLT2007	IWSLT2007 训练语料汉语	40k
	2004-863-008	Chinese LDC 中的对话语料	51k
	CLDC-LAC-2003-004	CLDC 对齐语料	199k
	CLDC-LAC-2003-006	CLDC 对齐语料	299k
开发集	IWSLT06 开发集	IWSLT06 提供的开发集	489
测试集	IWSLT05 测试集	IWSLT05 提供的测试集	506
	IWSLT06 测试集	IWSLT06 提供的测试集	500
	IWSLT07 测试集	IWSLT07 提供的测试集	489

表8 IWSLT 上的实验结果

		IWSLT06-dev	IWSLT05-test	IWSLT06-test	IWSLT07-test
Baseline1		0.095 648	0.155 4	0.088 0	0.063 0
Baseline2		0.187 492	0.409 1	0.173 8	0.231 5
Baseline3		0.190 749	0.413 0	0.175 4	0.227 0
Adapt + topN	100k	0.187 136	0.359 2	0.157 3	0.191 3
	200k	0.188 017	0.380 2	0.173 0	0.203 6
	300k	0.189 100	0.400 3	0.175 7	0.215 9
	400k	0.190 462	0.410 7	0.167 8	0.218 6
	500k	**0.195 362**	**0.417 5**	**0.181 1**	**0.235 7**
	600k	0.195 269	0.429 4	0.182 2	0.230 7
	700k	0.194 780	0.436 5	0.179 3	0.233 0

表8说明以下问题：

1. Baseline1 的训练语集是 FBIS 语料，属于新闻领域，而开发测试集属于旅游领域，领域差别较大，所以翻译效果较差。这也说明了进行领域自适应研究的必要性。

2. Baseline2 的训练集是各领域混杂的语料，其中也包括了旅游领域的语料。语料规模较大，所以开发测试集的许多短语能在混合语料中找到正确的译文，BLEU 值比 Baseline1 要高很多。将两者混合后的 Baseline3 因为语料规模的增大比 Baseline1、Baseline2 都要好。

3. 自适应模型的 BLEU 值随着选取语料规模 N 的变化而变化。基本的变化规律是：当 N 较小时，随着 N 的增大，BLEU 一直增大，增大到一定程度后，再继续增大 N，BLEU 值不稳定，且有下降的趋势。

我们分析其原因是：当 N 较小时，有许多短语没有学习过，解码器找不到对应的译文，增大语料规模，可以使得更多的本领域短语被学习到。增大到一定程度以后，继续增大 N，排名靠后的语料与开发测试集合的领域差别较大，对译文选择起到干扰作用。

N＝500k 时，在开发集上 BLEU 值最高(0.195 362)，在测试集 IWSLT07 上面 BLEU 值也是最高的。我们以 N＝500k 作为实验结果。自适应模型相比 Baseline2 在三个测试集合都有不同程度的提高：IWSLT05 上提高 0.84 个点，IWSLT06 上提高 0.73 个点，IWSLT07 上提高 0.42 个点。相比 Baseline1，自适应模型提高效果更加显著。

4.3　词特征与词性特征的比较

为了进一步分析领域特征带来的影响，我们在 topN＝500k 的基础上分别对词特征、词性特征进行实验对比。

当不使用单语领域信息时，自适应模型退化为标准的基于短语的翻译模型，我们以此为 Baseline。加入不同特征时的翻译 BLEU 对比见表 9。其中 W_{-1} 代表下文第一个词、W_{+1} 代表上文第一个词、POS_{-1} 代表下文第一个词的词性、POS_{+1} 代表上文第一个词的词性。

表 9　采用不同上下文特征对自适应效果的影响

	IWSLT05 (BLEU-4%)	IWSLT06 (BLEU-4%)	IWSLT07 (BLEU-4%)
Baseline	41.09	17.95	22.20
W_{-1}	41.17	18.08	22.25
W_{+1}	41.14	18.07	22.31
POS_{-1}	41.66	18.14	23.43
POS_{+1}	41.39	18.10	23.20
$W_{-1}+W_{+1}+POS_{-1}+POS_{+1}$	**41.75**	**18.11**	**23.57**

表 9 中第二行是不使用任何上下文的 Baseline 值，第三到第六行代表分别加入不同的上下文特征进行自适应，第七行是加入所有特征的结果。从表 9 可以看出：

1. 使用多个上下文特征要好于使用单个特征。

使用多个上下文特征可以产生多个领域特征，短语译文选择时候可以利用更多的信息源，从而做出更加正确的判断。

2. 词性特征作为上下文要明显好于词特征。

从 BLEU 值看，在三个测试集上，使用上下文词性特征普遍比使用上下文词特征效果要好。我们分析原因是使用词特征时，数据稀疏问题影响要比使用词性特征严重。我们使用的判别式词性标记工具采用了北大语料库加工规范标准，词性集有 40 多个[11]。

为了比较词特征与词性特征的作用，我们统计了短语表中分别被词特征和词性特征赋予领域概率的短语对数目，如表 10 所示。

表 10　使用词和词性特征被赋予领域特征的短语对数目比较

特征	# number of phrase pairs adapted	Ratio of phrase pairs adapted	Total
POS_{+1}	2 075 166	9.87%	21 019 258
POS_{-1}	1 801 445	8.57%	
W_{+1}	1 327 892	6.31%	
W_{-1}	1 101 238	5.24%	

从表 10 可以看出，被词性特征赋予领域概率的短语对数目要大于被词特征赋予领域概率的短语对数目。以词性作为上下文信息，数据稀疏问题远没有以词为特征时严重。

4.4　单语规模的影响

图 1 是在 topN＝500k 的自适应实验中只改变单语数量，保持其他因素不变的情况下比较翻译结果。可以看出，随着单语规模的增大，自适应效果会越来越好，各个测试集的 BLEU 值一直处于上升趋势。

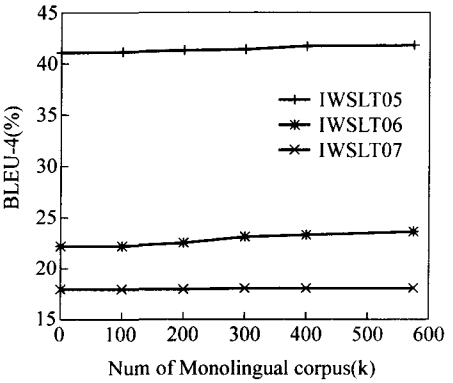

图 1　单语规模与 BLEU 值关系

领域特征的计算受单语规模影响较大。当单语规模较小时，数据稀疏问题会变得非常严重，这时只能对较少的部分短语对赋予领域翻译概率，翻译质量的改善将很不明显。如果单语数量变为 0，则所有短语对的领域概率将变为 0，领域特征将不起作用，自适应模型变退化为标准的基于短语的翻译模型。

对某个领域来说，单语的获取要比双语获取容

易得多，数量也大得多，本文提出的利用单语进行自适应的方法有应用价值。利用单语进行自适应研究的方法可以充分挖掘单语内部的领域信息，随着单语规模的增大，我们相信该方法会起到更大的作用。

5 总结与下一步的工作

本文提出了一种基于单语上下文信息的自适应方法，在对数线性模型框架内引入领域特征。领域特征的计算中，一方面利用检索模型从混合语料中检索领域类似语料以更准确的估计本领域的翻译概率，另一方面从领域单语中挖掘单语的上下文信息并用来计算领域特征。

从实验结果与分析可以看出，利用单语上下文信息能够对统计机器翻译领域自适应有所帮助的。从理论上，该方法既可以使用上下文词、词性等局部上下文信息，也可以使用长距离的上下文信息。如果不考虑任何上下文信息，所有短语对的领域特征值变为0，便退化为标准的基于短语的翻译模型。

当单语规模较小时，新模型会存在数据稀疏的问题。这时用单语上下文信息的方法只能对较少的短语对赋予领域特征。随着单语规模的增大，自适应的短语会越来越多，新模型的效果会越来越好。一般而言，单语的获取要比双语容易得多，本文的方法是有应用价值的。

下一步工作我们将寻找解决数据稀疏问题的办法，并尝试引入更多的上下文特征，还将考虑多元的上下文特征。

参考文献

[1] Peter. F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, The Mathematics of Statistical Machine Translation: Parameter Estimation [J]. Computational Linguisitics, 1993, 19(2): 263-312.

[2] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translateion [C]//Proceedings of HLT-NAACL 2003: 127-133.

[3] Franz Josef Och and Hermann Ney. Discrimitive training and maximum entropy models for statistical machine translation [C]//Proceedings of ACL 2002, 2002: 295-302.

[4] Matthias Eck, Stephan Vogel, Alex Waibel. Language model adaptation for statistical machine translation based on information retrieval[C]//International Conference on Language Resources and Evaluation, 2004.

[5] Bing Zhao, Matthias Eck, Stephan Vogel. Language Model Adaptation for Statistical Machine Translation ria structured query modes [C]//Proc. of COLING, 2004: 411-417.

[6] Almut Silja Hildebrand et al, Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval [C]//Proc. of EAMT 2005, 2005: 133-142.

[7] Nicola Ueffing, Gholamreza Haffari and Anoop Sarkar. Semi-superivesed Model Adaptation for Statistical Machine Translation [J]. Machine Translation, 2008, 21(2): 77-94.

[8] Yajuan Lü, Jin Huang. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization [C]//International Conference on Empirical Methods in Natural Language Processing (EMNLP), 2007: 343-350.

[9] A. Stolcke. 2002. SRILM-an extensible language modeling toolkit[C]//Proc. of ICSLP, 2002: 901-904.

[10] Papinensi, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation[C]//Proc. of the 40th Annual Meeting of the Association of Computational Linguistics, 2002: 311-318.

[11] 俞士汶，段慧明，朱学锋，孙斌，常宝宝. 北大语料库加工规范：分词 词性标注 注音 [J]. Journal of Chinese Language and Computing, 2002, 13(2): 121-158.