统计机器翻译中多分词结果的融合*

马永亮　赵铁军

哈尔滨工业大学教育部—微软语言语音重点实验室 哈尔滨 150001

E-mail: {ylma, tjzhao}@mtlab.hit.edu.cn

摘　要：汉英统计机器翻译中，汉语语料通常需要使用中文分词将句子切分成词序列。然而中文分词不是为统计机器翻译而开发的技术，它的分词结果不能保证对统计机器翻译的优化。近些年，一些研究试图改进中文分词方法从而达到对统计机器翻译的优化。在本文中，我们将从另外的角度研究中文分词对统计机器翻译的影响。我们的基本思想是利用多分词结果作为额外的语言知识，提出一种简单而有效的方法使这些知识为统计机器翻译所用，我们使用了一系列策略融合多分词结果，并将融合结果应用在统计机器翻译系统中。实验结果表明这种方法比没有使用多分词结果融合的系统提高 1.89 个 BLEU 分数。

关键词：统计机器翻译，中文分词，混合模型，翻译模型特征插值，多策略特征融合

Combining Multiple Chinese Words Segmentation Results for Statistical Machine Translation

Ma Yong-liang, Zhao Tie-jun

MOE-MS Key Laboratory of Natural Language Processing and Speech, Harbin Institute of Technology, Harbin 150001

E-mail: {ylma, tjzhao}@mtlab.hit.edu.cn

Abstract: In Chinese-English statistical machine translation (SMT), Chinese texts always need Chinese word segmentation (CWS) which segments sentences into words. However, CWS is not developed for SMT, its results are not necessarily optimal for SMT. In recent years, many investigations have been performed concerning making CWS suitable for SMT, but we explore it from another direction. In this paper, our basic idea is to use multiple CWS results as additional language knowledge source and we present a simple and effective approach to use multiple CWS results for SMT. We also give experiment results over rang of strategy, and the best result shows we gain 1.89 BLEU percentage points over a start of the art SMT system without using multiple CWS results.

Key Words: Statistical machine translation, Chinese word segmentation, mixture modeling, feature interpolation of translation model, multi-strategy feature blending of translation model.

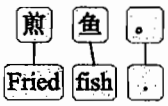
1　引言

汉-英统计机器翻译通常会遇到一个问题：中文语料是由一个个以字为基本单位的句子组成。这与大多数欧洲语言是完全不同的，它们的句子是由自然切分的单词组成的。这意味着，在原始的中文语料中并没有“词”。有研究人员进行了实验[1]，基于单个字切分的汉语语料不能使统计机器翻译获得最好的性能。对于实际的汉-英统计机器翻译系统，常常需要使用中文分词工具对汉语语料进行预处理。

广泛使用的中文分词方法较多，最早有一些基于最大匹配或者基于词典的方法，随后出现了基于层次隐马尔可夫的中文分词方法[2]。最近，许多 state of the art 机器学习算法被广泛应用在自然语言处理领域，比如最大熵，支持向量机，条件随机域等。基于这些机器学习算法，研究人员们提出了相应的中文分词方法[3-5]。除了众多中文分词方法，中文分词还使用了几种不同的中文分词标准，比如 CTB，PKU，MSR 等等。这些标准定义了不同的分词原

*本文研究受国家自然科学基金项目（60736014）和国家 863 计划项目（2006AA010108）的支持。

则，这使得同样的中文语料在使用不同标准的分词工具的处理后分词结果也有较大的差别。例如，图 1 给出一个简单句对的两个分词结果和相应的词对齐结果。



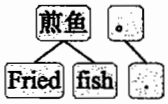


图 1 分词结果和词对齐

图 2(a) 分词结果和词对齐

图 2(b)分词结果和词对齐

无疑这两种分词结果都是正确的，但是某些情况下我们更希望是第一个分词结果。因为，在这种情况下我们能够从句对中获得词对齐信息“煎—Fried”。这样当我们就从句对中学习到了有用的翻译知识。再给出一个对中文分词来说更加难以面面俱到的例子。我们的目标是翻译“能把腰部弄短吗？”，当翻译结果是 “Can you shorten the waist ?”，图 2(a)中的分词和对齐结果没有问题。但是，如果翻译结果是 “Can you make the waist a little shorter?”, 我们更倾向于使用图 2(b)中的中文分词结果。我们认为基于不同方法和分词标准的中文分词工具产生的结果尽管不同，但是在特定的上下文和语义中他们都可能是正确而有用的。

这些通用中文分词方法并不是为统计机器翻译量身定做的，这使得他们的分词结果对统计机器翻译而言也就不可能是优化的。近些年许多研究者对中文分词在统计机器翻译中的影响进行了广泛而深入的研究，力求能够使中文分词更加适合于统计机器翻译。[1]将中文分词集成到统计机器翻译系统的模型训练和解码过程中，Chang[6]等的研究表明分词的一致性和分词的粒度对统计机器翻译是重要的, 并且提出了一种直接针对统计机器翻译对分词的粒度进行优化的方法。Zhang 等[7]进行了比较综合的研究。 在[7]中, 作者对中文分词和统计机器翻译的关系进行了细致的分析, 包括分词标准、未登陆词、切分错误等对 SMT 的影响。

在[1]和[6]中，作者的主要目是针对统计机器翻译来优化中文分词方法。总的来说，他们要找到单一的一种中文分词方法，从而尽可能使分词结果适合统计机器翻译。但是从我们的观点考虑，即使存在一个对统计机器翻译完全优化的中文分词结果，并且我们能够通过对中文分词方法的改进获得这种中文分词，我们依然可以利用基于不同分词方法和不同分词标准的分词工具的结果，并从中获得对统计机器翻译而言有用的信息。我们研究的出发点与[7]中融合多个不同分词结果训练的模型是一致的。通过进一步研究，我们发现了更有价值的结论。

2 层次短语翻译模型

从 Brown 等人提出统计机器翻译方法[8]，统计机器翻译已经逐渐成为机器翻译的主流技术。多年来，基于短语的统计机器翻译[9]、log-linear 模型[10]、最小错误率训练[11]等较大的提高了统计机器翻译系统的性能，基于句法的统计机器翻译也逐渐发展起来。

本文使用的统计机器翻译系统是新近兴起的一种基于句法的统计机器翻译系统[12]。层次短语翻译模型建立在上下文无关文法的基础之上。给定源语言句子 $\mathbf{f}$ ，一个同步上下文无关文法将存在有多种推导得到源语言端的 $\mathbf{f}$ ，因此也就对应着会有多种可能的翻译结果 $\mathbf{e}$ 。在 log-linear 模型框架下，可以对每种推导如下建模：

$$P(D) \propto \prod_i \phi_i(D)^{\lambda_i} \tag{1}$$

式中 D 是某个推导，ϕ_i 是定义在 D 推导上的特征函数，λ_i 是特征权重。因此我们可以通过找到使 $P(D)$ 最大的推导进而找到此推导对应的目标语言句子 $\hat{\mathbf{e}}$：

$$\hat{\mathbf{e}} = e(\arg\max_{D s.t. f(D)=f}(P(D))) \tag{2}$$

式中 $f(D)$ 是推导 D 对应的源语言句子，$e(D)$ 是推导 D 对应的目标语言句子。

在我们的系统中，我们使用了 7 个特征，短语翻译概率, 反向短语翻译概率, 词汇化权重, 反向词汇化权重，词汇惩罚，规则惩罚，Glue 规则惩罚和目标语言语言模型。这些特征的权重通过最小错误率训练进行优化[11]。

3 多分词结果融合

3.1 三个中文分词工具

HIT 中文分词系统[13]的基本方法是基于词典的。为了获得更高的分词精度，HIT 中文分词系统在全切分的基础上加入了多步处理策略。ICTCLAS[2]是基于层次隐马尔可夫模型的中文词汇分析系统。ICTCLAS 可以同时提供中文分词，词性标注，中文命名实体识别。Stanford Chinese Word Segmenter[6]是基于条件随机域模型的中文分词系统。模型除了使用字符、字形、字符复现等特征外，还加入了词汇化特征。

3.2 翻译模型特征插值

我们将不同分词结果对应的训练语料看成训练数据的几个部分，然后对应每个部分训练一个模型，最后根据目标调整各个模型的权重。通常可以通过两种方式完成这样的工作。

第一种方式是线性的，在实际应用中被广泛采纳，可以表示为：

$$p(x|h) = \sum_c \lambda_c p_c(x|h) \tag{3}$$

式中 $p(x|h)$ 是混合后的模型，$p_c(x|h)$ 是从部分 c 上训练的模型，λ_c 是对应的权重。

第二种是 log-linear 模型，可表示为:

$$p(x|h) = \prod_c p_c(x|h)^{a_c} \tag{4}$$

式中 a_c 是通过全局优化得到的权重。本文中，我们使用了与 Zhang 等[10]相同的方法, 采用线性的方式对多分词结果进行融合，称为翻译模型线性插值。

在层次短语翻译模型中，我们使用了四个翻译特征。翻译模型的特征插值就是对相同类型的特征进行线性插值计算。我们使用前面提到的 3 个中文分词工具对汉-英平行语料中的中文部分进行分词处理，由此得到对应的 3 个经过分词汉-英平行语料。然后这 3 个平行语料分别用于训练各自的层次短语翻译模型。这样，每个训练出的翻译模型就对应一个中文分词工具。特征插值的计算可以描述如下：

$$p(e|f) = \sum_{i=1}^{S} \lambda_i p_i(e|f) \tag{5}$$

式中 $p_i(x|h)$ 是对应第 i 个翻译模型的短语翻译特征，λ_i 是第 i 个翻译模型的权重，S 是翻译模型的总数。因为是线性相加，我们令 $\sum_{i=1}^{S} \lambda_i = 1$。延续 [7]对 λ 的处理，所有的 λ 都取相同的值。其它 3 个特征的计算与短语翻译特征相同，这里就不再详细一一列出了。

3.3 翻译模型特征融合

在[7]中, 特征融合被用于与特征插值进行比较。特征融合首先要定义一个主模型，除主模型外的其它模型被定义为辅助模型。然后，将在辅助模型中的翻译条目加入到主模型中。

与特征插值不同的是，只有那些在主模型中没有的翻译条目才从辅助模型中加入到主模型中，并且每个翻译条目的各个特征值不变。[7]中的实验结果显示特征插值要比特征融合性能更好，但是并没有进一步深入研究特征融合性能不如特征插值的原因。本文中，我们继续使用 [7]中对特征融合的定义，不过我们的研究重点是如何利用多分词结果提高统计机器翻译性能。我们认为翻译模型的特征融合是比翻译模型的特征插值更有潜力的方法，因为这种方法基于一个明显的事实：更多的信息会有效的扩大模型的空间，如果能够有效的控制模型空间的增长与随之带来的噪声问题，那么翻译模型的特征融合技术就有希望获得较好的性能。从基本出发，我们在文章中只探讨两个模型的特征融合。

3.4 特征融合的策略

除去翻译模型中的特征值，模型中的翻译条目由两部分组：源语言部分 f 和对应的目标语言部分 e。为了进一步分析辅助模型在与主模型融合时的特性，我们将辅助模型分为 4 部分：

- $f|||e$, f 是源语言部分， e 是对应的目标语言部分。它们已经在主模型中成对出现。
- $\overline{f}|||e$, $\overline{f}$ 是源语言部分，没有在主模型中出现； e 是对应的目标语言部分，已经在主模型中出现。
- $f|||\overline{e}$, f 是源语言部分，已经在主模型中出现； $\overline{e}$ 是对应的目标语言部分，没有在源语言中出现 。
- $\overline{f}|||\overline{e}$, $\overline{f}$ 是源语言部分，没有在主模型中出现； $\overline{e}$ 是对应的目标语言部分，没有在源语言中出现。

$\overline{f}|||e$、 $f|||\overline{e}$、 $\overline{f}|||\overline{e}$ 将被应用到特征融合中。我们认为这 3 部分以不同的方式影响特征融合，它们的组合也同样在模型融合中表现出不同的性质。为了通过特征融合有效的利用多分词结果提高统计机器翻译系统的性能，我们将对这些融合方式进行实验。$\overline{f}|||e$ 给出了一个新的源语言选项，对应的翻译是主模型中已经存在的目标语言选项； $f|||\overline{e}$ 给出了已经存在于主模型中的源语言选项的另外一种目标语言选项； $\overline{f}|||\overline{e}$ 给出了主模型中完全不存在的翻译选项，对于主模型而言是全新的信息。 在[7]中 $f|||e$ 被用于特征插值， $\overline{f}|||e$、 $f|||\overline{e}$、 $\overline{f}|||\overline{e}$ 被用于特征融合。考虑到被用于特征融合的三个部分的多种组合，我们提出 7 个特征融合的策略。表 1 中是 7 个用于特征融合的策略，[7]中的特征融合采用的是策略 1。从表 1 中我们可以很容易发现策略 1 使用了所有可有的信息。策略 1 尽管使用了最多的额外信息，主模型的模型空间能够得到最大程度的扩展，但是策略 1 也最有可能使主模型受到这些额外信息的干扰和引入噪声。

表 1 特征融合的策略

融合策略	辅助模型中的部分
1	$\overline{f}\|\|\|e, f\|\|\|\overline{e}, \overline{f}\|\|\|\overline{e}$
2	$\overline{f}\|\|\|e, f\|\|\|\overline{e}$
3	$\overline{f}\|\|\|e, \overline{f}\|\|\|\overline{e}$
4	$f\|\|\|\overline{e}, \overline{f}\|\|\|\overline{e}$
5	$\overline{f}\|\|\|e$
6	$f\|\|\|\overline{e}$
7	$\overline{f}\|\|\|\overline{e}$

表 2 IWSLT2004 语料

语料		句子数	平均句长	词数
训练集	中文	20000	9.1	7643
	英文	20000	9.4	8191
开发集	中文	506	6.9	870
	英文	8089	7.5	2435
测试集	中文	500	7.6	893
	英文	8000	8.4	2496

4 实验

4.1 系统

我们的实验平台是一个用 C++实现的层次短语翻译系统[12]。训练过程中使用 GIZA++获得词对齐，使用 SRI language Modeling Tookit[14]训练 modified-KN 平滑的 3-gram 语言模型。

4.2 训练集和测试集

我们在小规模口语汉-英平行语料上进行了实验。实验的语料采用 IWSLT2004 的训练集和测试集。实验采用自动翻译评价，使用了大小写相关的 BLEU-4 自动翻译评价指标[15]。训练集的英语部分用来训练语言模型。为了优化系统的特征权重，使用最小错误率训练来最大化系统在开发集上的 BLEU 得分。最后，系统将使用不同的模型在测试集上进行测试。训练集和测试集的详细信息见表 2。

4.3 实验结果

实验中，我们使用没有加入任何模型优化方法的层次短语翻译系统作为 Baseline 系统。在 Baseline 系统的实验中，所有中文语料均使用 HIT 中文分词系统进行分词。表 3 给出了 Baseline 系统的实验结果。表 4 给出了用 Stanford 和 ICT 中文分词系统对所有中文语料进行分词后系统的实验结果。

表 3 Baseline 实验结果

Baseline	BLEU(开发集)	BLEU(测试集)
HPBT System(HIT)	40.95	40.50

表 4 Stanford and ICT 实验结果

System	BLEU(开发集)	BLEU(测试集)
HPBT (Stanford)	39.25	41.73
HPBT (ICT)	39.62	40.51

我们对所有 7 个特征融合策略进行了两组实验。第一组实验使用 HIT 中文分词系统处理的训练语料作为主模型的训练语料，Stanford 中文分词系统的处理的训练语料作为辅助模型的训练语料；第二组实验使用 HIT 中文分词系统处理的训练语料作为主模型的训练语料，ICTCLAS 处理的语料作为辅助模型的训练语料。两组实验中的开发集和测试集均由 HIT 中文分词系统进行分词。表 5 给出了第一组实验结果，表 6 给出了第二组实验结果。

表 5 HIT+Stanford 实验结果

策略	BLEU(开发集)	BLEU(测试集)
1	41.28	41.33
2	41.46	40.55
3	**41.14**	**42.08**
4	40.58	41.23
5	**40.82**	**42.14**
6	**40.65**	**42.26**
7	40.90	40.90
特征插值	**39.66**	**41.34**

表 6 HIT+ICT 实验结果

策略	BLEU(开发集)	BLEU(测试集)
1	41.21	41.35
2	**40.48**	**42.39**
3	**40.66**	**42.12**
4	40.37	41.41
5	40.58	41.25
6	**41.07**	**41.95**
7	40.76	41.31
特征插值	**40.16**	**41.41**

最后，我们使用Paired Bootstrap Resampling[16]对系统性能差异的统计显著性进行衡量。结果显示，在p=0.05的条件下，HIT+CTB组中的策略6和HIT+ICT组中的策略2对Baseline系统、特征差值策略、HBPT(Stanford)和HBPT(ICT)的BLEU分数提高都具有统计显著性。

5 实验结果分析

实验结果显示通过多策略的特征融合我们可以利用多分词结果有效的改进机器翻译系统的性能。相对与 Baseline 系统，两组实验分别获得了 1.76 和 1.89 的 BLEU 分数增长。

我们观察 Baseline 系统和使用特征融合方法的系统的输出，发现特征融合在改进系统性能方面最为显著的表现就是将 Baseline 系统无法翻译的未登陆词进行了正确的翻译，而这些翻译知识正是从辅助模型中获得的。例如，Baseline 中的翻译结果“I can't see 屏幕.”改进为“I can't see screen very well.”，“How 按摩 脊背?”改进为“Have massage 脊背?”，“一共 is 八十七 dollars.”改进为“八十七 is all together dollars.”等。

前面这种类型的改进可以通过融合多分词结果中合适的分词结果获得正确的词对齐并产生原本存在于汉-英句对中的对应翻译项。借助正确的词对齐，我们可以获得更准确的短语翻译。这些通过融合多分词结果获得的新翻译知识是十分有意义的。这种改进可以通过观察系统输出、分析训练和解码过程方便的认识到，但是我们猜测翻译模型融合方法不会只通过这种方式改进系统性能，其它复杂、深入的原理还不能通过我们的实验直观的反映出来。

还存在一些值得关注的现象。我们知道策略 1 使用了最多的额外信息，辅助模型中 $\bar{f}\,|||\,e$、$f\,|||\,\bar{e}$、$\bar{f}\,|||\,\bar{e}$ 3 部分的翻译条目都被加入了主模型。但是观察表 5、表 6，策略 1 只是获得中等水平的性能。我们在前面已经简单分析过策略 1 可能存在的问题：由于它使用了最多的额外信息，这也使得它可能受到的负面影响最大。一方面，这些额外信息自身之间存在相互影响；另一方面，这些额外信息与主模型信息之间也存在相互影响。在表 5 中，我们从策略 1 种去掉 $f\,|||\,\bar{e}$ 得到策略 3，策略 3 的性能要比策略 1 好一点；从策略 3 中再去掉 $\bar{f}\,|||\,\bar{e}$ 得到策略 5，策略 5 的性能就已经要好于特征插值了。策略 6 反而只使用了 $\bar{f}\,|||\,\bar{e}$ 就获得了这组实验中最好的性能，比 Baseline 高 1.76 个 BLEU 分数，比特征插值高 0.92 个 BLEU 分数。在表 6 中，策略 2、策略 3 和策略 6 的性能都高于特征插值，策略 2 的性能是所有实验结果中最好的，比 Baseline 高 1.89 个 BLEU 分数，比同组的特征插值高 0.98 个 BLEU 分数。所以，我们指出辅助模型的各个部分之间以及辅助模型的各个部分与主模型之间存在复杂相互作用，我们从实验中没能直接确定这种相互作用的形式。但是，这 7 种翻译模型融合的策略给出对于辅助模型中信息一个相对合理的分类框架，从而能够减小复合模型中的噪声影响，取得了较好的实验结果。

因为辅助模型的各个部分之间以及辅助模型的各个部分与主模型之间的复杂相互影响，对应不同中文分词系统的组合，对于固定的训练和测试语料，最优特征融合策略并不固定。因此，当我们将这样的方法应用到大规模语料的处理时，我们必须首先确定一个最优的策略。在没有好的、确定的方法前，这样的工作将是比较耗时的。

6 结论与展望

我们提出了一个融合多分词结果改进统计机器翻译系统性能的方法。这个方法直接利用已有的中文分词系统的分词结果，采用多策略特征融合技术直接得到复合的翻译模型。使用复合翻译模型的统计机器翻译系统性能得到了显著的提高。

在我们的方法中，我们将不同的中文分词工具的分词结果看作是多知识源，力求能够高效的使用这些知识，提高统计机器翻译系统性能。我们为此使用了多策略翻译模型融合，每个策略对应使用辅助模型的一部分信息。这些策略并不能都在测试集上得到性能的显著提高，但是一些策略在实验中表现出了对 Baseline 系统较明显的性能优势和对特征插值的优势。从实验结果和分析中，我们指出在统计机器翻译中有效的利用多分词结果能够显著的提高汉-英统计机器翻译系统的性能。同时，我们也认识到我们提出的方法在处理大规模统计

机器翻译时存在效率问题。在未来的工作中，我们将进一步研究在我们提出的方法中融合策略选择的问题，并且我们还会进一步深入发掘来自不同分词结果的各个模型的内部以及它们之间的作用机制，进一步减小噪声在复合模型中的影响。最后，我们指出结合模型特征插值和模型特征融合有希望达到减小噪声影响的目的，我们也会在未来的工作中研究这种方法的优劣。希望更多的研究者能够从中文分词入手，解决中文分词在统计机器翻译中存在的问题，进一步提高与汉语相关的统计机器翻译的性能。

参考文献

[1] Jia Xu, Richard Zens, and Hermann Ney. Do we need Chinese word segmentation for statistical machine translation? [A] In ACL SIGHAN Workshop 2004[C], July 2004, pages 122-128.

[2] Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. HHmm-based chinese lexical analyzer ICTCLAS[A]. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing[C], pages 184-187, July 2003.

[3] Nianwen Xue and Libin Shen. Chinese word segmentation as LMR tagging[A]. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing[C], pages 176-179, July 2003.

[4] Taku Kudo and Yuji Matsumoto. Chunking with support vector machines. In Second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies[C], pages 1-8, 2001.

[5] Fuchun Peng, Fangfang Feng, and Andrew McCallum. Chinese segmentation and new word detection using conditional random fields[A]. In Proceedings of Coling 2004[C], pages 562-568, 2004.

[6] Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. Optimizing Chinese word segmentation for machine translation performance[A]. In Proceedings of the Third Workshop on Statistical Machine Translation[C], pages 224-232, June 2008. Association for Computational Linguistics.

[7] Ruiqiang Zhang, Keiji Yasuda, and Eiichiro Sumita. Chinese word segmentation and statistical machine translation [J]. ACM Trans. Speech Lang.Process. , 5(2):1-19, 2008.

[8] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation [J]. Comput. Linguist., 16(2):79-85, 1990.

[9] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation[A]. In Proceedings of the 2003 Conference ofthe North American Chapter of the Association for Computational Linguistics on Human Language Technology[C], pages 48-54, 2003.

[10] Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation[A]. In Proceedingsof 40th Annual Meeting of the Association for Computational Linguistics[C], pages 295-302, July 2002.

[11] Franz Josef Och. Minimum error rate training in statistical machine translation[A]. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics[C], pages 160-167, 2003. Association for Computational Linguistics.

[12] David Chiang. Hierarchical phrase-based translation [J]. Comput. Linguist., 33(2):201–228, 2007.

[13] 赵铁军，等. 提高汉语自动分词精度的多步处理策略[J]. 中文信息学报, 2001 , 15(1): 13-18.

[14] Andreas Stolcke. Srilm - an extensible language modeling tookit[A]. In Proceedings of the International Conference on Spoken Language Processing[C], volume 2, pages 901–904, 2002.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation[A]. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics[C], pages 311–318, 2001.

[16] Philipp Koehn. Statistical significance tests for machine translation evaluation[A]. In Proceedings of EMNLP 2004[C], pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.