维汉机器翻译中可变权重的编辑距离语言模型语料领域适应*

朱少林[1,2,3]，杨雅婷[1,2]，米成刚[1,2]，董瑞[1,2]，王磊[1,2]

（1.中国科学院新疆理化技术研究所，新疆 乌鲁木齐 邮编 830011；

2. 新疆民族语音语言信息处理重点实验室，新疆 乌鲁木齐 邮编 830011；

3. 中国科学院大学，北京 邮编 100049）

摘要：本文旨在从大规模的单语语料中选取特定领域语料训练统计机器翻译的语言模型,以提高机器翻译质量。基于不同词对领域的贡献度不同和领域内语料句子在用词、搭配、句式结构等上具有诸多相同的特征两个方面，并通过调研现有的领域适应语料选取技术，本文将词项的 TD-IDF 权重与编辑距离方法相结合应用到语言模型领域适应的语料选取中，以提高机器翻译质量，实验结果表明，本文的方法可以有效的提高机器翻译质量。

关键词：统计机器翻译；领域适应；数据选取；

中图分类号：TP391 **文献标识码：**A

Language model corpus adaptation by variable weights edit distance in Uighur-Chinese machine translation

ZHU Shaolin[1,2,3], YAN Yating[1,2] , MI Chenggang[1,2], DONG Rui[1,2],WANG Lei[1,2]

(1.The Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi, Xinjiang 830011, China ;

2. National key laboratory of speech language information processing of Xinjiang, Urumqi, Xinjiang 830011 , China ;

3. University of Chinese Academy of Sciences, Beijing 100049 , China)

Abstract: The purpose of this paper is to extract specific areas corpus from the large-scale monolingual corpus to train language model of the statistical machine translation, to improve the quality of machine translation. Based on different words in the field of corpus have different contribution and the specific field sentences have many of the same characteristic which are sentence and field on collocation, sentence structure etc. This article will appliy the edit distance and TD-IDF weight to the field of adaptation in the selection of corpora, to improve the quality of machine translation. The experimental results show that the method can effectively improve the quality of machine translation.

Key words: the statistical machine translation； data selection； field of adaptation

* **收稿日期：** **定稿日期：**

基金项目：中国科学院先导科技专项（XDA06030400）；新疆维吾尔自治区青年自然科学基金项目（2015211B034）；新疆维吾尔自治区重点实验室开放课题项目（2015KL031）

作者简介：朱少林（1989—），男，硕士研究生，自然语言处理、机器翻译；杨雅婷（1985—），女，副研究员，自然语言处理、机器翻译；米成刚（1986—），男，助理研究员，自然语言处理、机器翻译。

1 引言

统计机器翻译（Statistical Machine Translation，简称 SMT）的翻译质量严重依赖训练语料，当测试集和训练集分布不同，且属于不同的领域时，翻译质量有着明显的下降，然而当测试集和训练语料属于同一领域且具有较高的同分布时，翻译质量有着极大的提高[1-2]。

SMT 是基于数据驱动的翻译系统，训练语料越多，就能得到更加准确的翻译效果，但是对于某些使用规模较小的语言，大规模平行语料的获取并不是一件容易的事，有时这种获取是极大的消耗人力、财力，比如维汉机器翻译，维汉双语语料的大规模获取还存在一定的难度，在本次研究中，基于当前的研究背景，本文突出从大规模的单语语料中选取领域相关的单语语料组成 SMT 的语言模型来提高翻译质量，原因是维汉机器翻译中，用来做训练语言模型的语料是汉语单语，汉语单语语料的获取是极为容易的，例如：网络、书籍、文档等多种途径。

在研究中假设已经有一小部分的领域内的语料可供使用，同时用于挑选的大规模语料中包含特定领域的语料，本文研究的任务是用已有的小部分的特定领域的语料，从大规模的异构的语料中选取和特定领域的语料同分布，同构的单语训练语料，用这部分单语语料训练 SMT 的语言模型，进而提高翻译质量。

在 SMT 中训练语料的一个特征严重影响其翻译质量，当训练语料和待翻译的语料具有同样分布，相同的结构，翻译的质量会有明显的提升，原因是相似的数据具有相同的分布，相同的结构，在措词、用词搭配、句式等方面都较相似。

进行特定领域的语料选取的过程，是一个计算句子间相似度的过程，目前已经提出了多种不同的方法，例如，众多学者将信息检索的 TF-IDF 模型和向量空间模型应用到计算相似度中，模型的优点在于考虑了词的权重，根据权重的高低进行相似数据的选取，但是空间向量模型没有考虑词的位置顺序，仅仅考虑两个句子的单词重合数量，造成的结果是忽视句子的词的搭配、用词习惯和句子结构等重要信息。

Wang 等作者在 2013 年提出的基于编辑距离模型的方法计算相似度，该方法的优点是编辑距离的模型考虑的词与词之间的顺序，词与词之间的搭配关系，能更好的进行相似度的计算，但是他们没有考虑词与词之间的不同，他们将编辑操作的代价设为相等权重，但是在根据信息检索理论，普遍被人公认的是，不同的词对文档的贡献度是不同的，也就是说不同词对于相似度的计算是不同，对于领域的划分贡献度也是不同，也就是在进行相似度计算时应该考虑词的权重。

本文的方法是根据两种方法的优缺点，将相似度的计算转换为编辑距离，但不同于之前 Wang 等人的方法，本文在计算编辑距离时，根据词的贡献度不同，区别不同词的贡献度进行计算。

本文的组织结构如下：第二部分介绍现阶段该领域的主要研究及数据选择的主要相关模型，第三部分介绍本研究的方法，第四部分介绍进行实验的对比及结果的讨论，最后是总结与展望。

2 相关工作

语言模型的领域适应技术首先被应用到语音识别中，随后被引用到机器翻译中，Beeferman 和 Huang 在 1999 年将信息检索中向量空间模型用到机器翻译的语言模型领域适应中。向量空间模型将词项的权重作为选取句子的重要因子，原因是领域内句子的词的权重一定比领域外的权重值更高。随后 Matthias Eck、Stephan Vogel 等人将这种信息检索技术应用到汉语到阿拉伯语的机器翻译中，他们选择用于训练语言模型的单语语料，2005 年 Hildebrand 等人将这一技术应用到了西班牙语到英语的 SMT 的翻译模型的领域适应语料选取中，对机器翻译的质量有个明显的提升。Lv 等人提出在不增加任何训练语料的条件下，通过从现有的双语语料中选择与待翻译文本相似的数据集来生成自适应的翻译模型，一定程度上提高了翻译质量。

近几年一些学者提出将编辑距离这一用于单词拼写矫正的技术应用到 SMT 领域适应的模型训练语料中，Leveling 等在 2012 年将这一技术应用到基于实例的机器翻译中。而 Wang 等首先将编辑距离应用到了 SMT 的语料选取中，他们使用该技术的出发点是判断句子间的领域相似是根据两个句子间的单词重合和单词的顺序，从这一点看计算编辑距离的技术正好符合这一标准，Wang 等通过实验证明可以减少训练数据的 99%，同时 BLEU 得到 1.8 的提高，明显可以提高翻译质量。

本文的方法是将编辑距离和 TF-IDF 技术结合来进行领域相关的语言模型语料选取，优点在于计

算两个句子间的编辑距离时将单词的权重视为可变的，这样更加符合事实，原因是不同的单词对于分类贡献度是不同的，这点在信息检索中已经被明确的证明，本文根据单词的不同权重计算编辑距离，这里还将编辑距离归一化，使得计算出来的结果是句子的相似度得分，同时为了减小数据稀疏的问题还采用数据平滑技术。

3 基于 TD-IDF 编辑距离的 SMT 语言模型领域适应语料选取

通过统计与分析，能够发现相同领域的语料在单词顺序、词语搭配等多个方面具有共同之处，而且我们会发现句子中的某些词就能将特定领域区别出来，即不同词对于领域的划分有着不同的贡献度，例如在新闻领域，如果句子中出现“据报道”，我们就能知道这句话是来自新闻领域。基于此，本文提出了使用 TF-IDF 和编辑距离模型结合的方法进行领域相关的语料选取。

3．1 基于编辑距离的模型

该方法首先用到单词的拼写校对，编辑距离定义为对于两个字符串 S1，S2 将 S1 替换为 S2 的最小编辑距离。通常，这样的操作包括：(i)将一个字符插入字符串；(ii)从字符串中删除一个字符；(iii)将字符串中的一个字符替换成另一个字符。通过定义可以看出，两个字符串间的编辑距离越小，说明两者间的相差越小，也就是说明两者间更相似。而计算两个句子相似度的方法是将得分 FMS 定义如下：

$$\mathrm{FMS} = 1 - \frac{LED_{word}(S_G, S_R)}{MAX(|S_G|, |S_R|)} \quad (1)$$

LED_{word}是一个计算S_G和S_R间的最小编辑距离，$|S_R|, |S_G|$是两个句子的长度，也就是单词的个数。可以用下四个例句来说明句子间的编辑距离：

例句 1：今天/，宁泽涛和孙杨/赢得/了/游泳/比赛/的/第一名。

例句 2：今天/,小明和小李/赢得/了/书法/比赛/的/第一名。

例句 3：今天/,博尔特/赢得/了/百米/比赛/的/第一名。

例句 4：今天/,李丽和李玲/赢得/了/百米/和/游泳/比赛/的/第一名。

句子中的斜线表示将句子分词，在上面的 4 个例句中，假设句子 1 是已经确定的特定领域的，那么按上述方法计算得到，句子 1 和句子 2、3 的编辑距离是相同的，编辑距离都是 1，而句子 4 的编辑距离却是 2，可以明显的看出句子 2 和句子 1 是来自不同的领域，而句子 3、4 是和句子 1 来自相同的领域，但是，如果按照上述方法，句子 2 会被认为领域相同的句子，而句子 4 却会被排除在外。

3．2 TD-IDF 模型

TF-IDF 是一种广泛被应用在信息检索领域的方法，结合向量空间模型(Vector Space Model，简称 VSM)共同进行数据相似的计算。对于一篇文档，这里指一个句子 D_i 被影射为一个向量，每一个向量代表句子中一个词项的权重(TF-IDF 值)，那么句子的向量化：

$$D_i = (W_{i1}, W_{i2}, \ldots\ldots W_{in}) \quad (2)$$

n 是一个句子中不重复单词的个数，W_{in}是词项的权重，计算方法如下：

$$W_{ij} = tf_{ij} \times \log(idf_{ij}) \quad (3)$$

tf_{ij}是词项的频率，表示一个单词在整个语料中出现的次数，idf_{ij}是单词的逆文档频率。

使用上述方法，计算下面四个例句中的单词的权重值：

例句 1：今天/，宁泽涛和孙杨/赢得/了/游泳/比赛/的/第一名。

例句 2：今天/,小明和小李/赢得/了/书法/比赛/的/第一名。

例句 3：今天/,博尔特/赢得/了/百米/比赛/的/第一名。

例句 4：今天/,李丽和李玲/赢得/了/百米/和/游泳/比赛/的/第一名。

将上述四个例句视为一篇文档，可以计算该文档中各个单词的 TD-IDF 权重，利用上述公式，可以计算得到词语“书法”的权重为 2.4，“百米”和“游泳”的权重为 0.48，其余词语的权重为 0。需要指出的是这里将命名实体当成同一个词，原因是为了计算方便。

3．3 改进的编辑距离模型

本文提出的方法是结合信息检索中的 TF-IDF 模型和编辑距离模型的优势，但不是简简单单的利用线性插值的方法，该方法来自于拼写校正的研究方法，在英语单词的拼写校正中，由于键盘上字符的位置不同，各个字符导致单词的一个字母被拼写成另一个的概率也不同，具体到实际就是在计算编辑距离时插入，删除或者替换的权重也是不相同的。

在本文的方法中，使用归一化的方法将两个句子相似度的计算转化为编辑距离的计算，具体的转化方法是在计算编辑距离时，对于插入、删除和替换操作的每一步均采用归一化的单词的词项 TF-IDF 值，具体操作如下公式 4：

$$\text{cost} = \frac{1}{W_{ij}+1} \qquad (4)$$

公式使用加一的原因是为了减少数据稀疏，cost 代表每一步插入、删除或者替换所需要的不同代价，W_{ij}表示一个句子中第i 个单词的词项权重 。

具体的算法伪代码如下图 1：

```
int d[|s1|+1][|s2|+1]=0
for i ← 1 to|s1|
do d[i][0]= 1/(w1[i-1]+1)
for j←1 to |s2|
do d[0][j] = 1/(w2[j-1]+1)
for i←1 to |s1|
do for j ←1 to |s2|
        do if s1[i]==s2[j]
           cost =0
        else cost = min(1/(w1[i-1]+1),1/(w2[j-1]+1))
d[i][j]=min{min(d[i-1][j]+1/(w1[i-1]+1),
    d[i][j-1]+1/( w2[j-1]+1)),d[i-1][j-1]+cost}
return d[|s1|][|s2|]
```

图 1 计算两个句子相似度的伪代码

算法的基本思路是首先开设一个二维数组 d[|s1|+1][|s2|+1],|s1|和|s2|分别表示句子 s1 和 s2 的长度，数组 d 用来存贮两个句子间的编辑距离，本算法的改进之处在于计算插入、删除和替换的编辑操作采用可变的权重，并且将单词的 TF-IDF 成功的融入转化成这种可变的权重，算法的重点在于如何将 TF-IDF 应用到具体的编辑操作中，并用数组 d 记录下来，可以用下列公式说明这一问题：

$$\mathrm{d}[i][\mathrm{j}] = \frac{1}{W[i]+1} \quad j = 0 \; and \; i \neq 0 \qquad (5)$$

$$\mathrm{d}[i][\mathrm{j}] = \frac{1}{W[j]+1} \quad i = 0 \; and \; j \neq 0 \qquad (6)$$

$$\mathrm{d}[i][\mathrm{j}] = \mathrm{Min}\begin{cases} d[i-1][j] \\ d[j-1][i] \\ 0 \quad S1[i] = S2[j] \\ \mathrm{Min}\left(\frac{1}{(W_i,W_j)+1}\right) \quad S1[i] \neq S2[j] \end{cases} \qquad (7)$$

公式 5 表示的是图 1 中的伪代码在第一个循环时执行的操作，公式 6 是上图伪代码在第二个循环时执行的操作，公式 7 中的$S1[i]$表示句子$S1$中的第 i 个单词，$S2[j]$表示句子$S2$中的第j 个单词，Min表示求最小值，公式 7 在伪代码的双重循环中进行，在我们的算法中i 和j 从 1 开始循环，直到i 大于句子 s1 的长度，j 大于句子 s2 的长度，此时得到的 d[|s1|][|s2|]便是句子 s1 和 s2 的编辑距离，同时也是他们的相似度得分。

下面按照本文的方法对下面的四个例句进行领域的划分，并指出该方法优于已有方法之处。

例句 1：今天/，宁泽涛和孙杨/赢得/了/**游泳**/比赛/的/第一名。

例句 2：今天/，小明和小李/赢得/了/**书法**/比赛/的/第一名。

例句 3：今天/，博尔特/赢得/了/**百米**/比赛/的/第一名。

例句 4：今天/，李丽和李玲/赢得/了/**百米**/和/**游泳**/比赛/的/第一名。

通过统计分析 4 个例句，会发现最能区别 4 个例句的是上述用黑体表示的单词，即这些单词的权重在不同的领域是变化的，这点在信息检索领域已被充分的证明，计算 4 个例句的相似度得分按照现有的编辑距离方法得到的结果已在上述 3.1 节给出，这里采用本文的方法给出各个句子的相似度得分，句子 1 和句子 3、4 的相似度得分都是 0.67，而和句子 2 的得分却是 0.29 这样就能将句子 2 和其它几个句子区别出来。这样的计算结果符合实际的情况，句子 2 出现的“书法”一词表明其所在的领域

不同于其它三个句子。

4 实验与分析

本次实验使用的基本的翻译系统是现今使用广泛的 Moses 机器翻译系统，使用 GIZA++进行翻译过程的词对齐，使用 SRILM 进行语言模型的训练，采用 4-gram 语言模型，同时为了提高语言模型的质量采用平滑等技术对语言模型进行优化，翻译模型采用基于短语的翻译模型。

对于语言模型的领域适应需要三类语料，首先是大规模的单语训练语料，它是一个包含法律、体育、教育等各个领域的大规模语料，本文使用的这个语料是 400 万的目标语言单语语料[3-4]，领域内的小规模语料使用的是教育和新闻两个领域内的语料，使用不同领域的语料是为验证本文方法的有效性，所有的实验数据如下表 1 所示：

表 1 实验语料统计

数据集	句子数量	平均长度
测试集	1000	9.12
领域内语料	5000	10.46
训练语料	4000000	10.32

4．1 基准系统

首先用基准的双语语料进行翻译系统的训练，本文使用的维汉双语语料来自 2013 年 CWMT 评测所提供的语料，训练一个 11 万句的维汉翻译系统。同时在训练机器翻译系统时，本文用单语的汉语语料训练一个单独语言模型，来说明语言模型对翻译效果的影响。本文设置了如下 2 个基准实验进行对比：

Baseline1：使用 11 万的维汉语料训练统计机器翻译系统。

Baseline2：使用 11 万的维汉语料训练统计机器翻译的翻译模型，而使用 400 万句的单语汉语和 11 万的维汉双语中的汉语单语组合的语料训练语言模型。

实验的结果如表 2 所示：

表 2 基准实验结果

训练语料	新闻测试集 BLEU	教育测试集 BLEU
Baseline1	33.70	32.89
Baseline2	34.08	33.26

从上表首先可以得出结论，在不增加平行训练语料在前提下，单独增加单语语料来提高语言模型可以提高机器翻译的质量，上表的实验结果显示可以提高最多 0.43 个百分点。再者从表中可以看出对于不同领域的测试集，翻译的效果不相同，说明对于训练语料进行领域选取是好的策略，本文所使用的 400 万的单语语料进行语言模型的训练，由于这些单语语料是包含各个领域的通用语料，如果选取出来其中和测试集所在领域相同的子集，除掉一些噪音影响对翻译效果会有更好的提高，下面本文给出的实验更能说明这一点。

4．2 数据优化实验

本文的优化实验主要包括两个部分：（1）通过三种不同的方法选取不同比例的领域适应的语言模型训练语料，便于进行对比试验；（2）将选取出来的语料训练 SMT 的语言模型，然后测试翻译效果。

4.2.1 数据选择过程

语言模型的领域适应数据选取的过程就是根据一种策略选取特定领域的训练语料，具体的数据选择过程如下图 2：

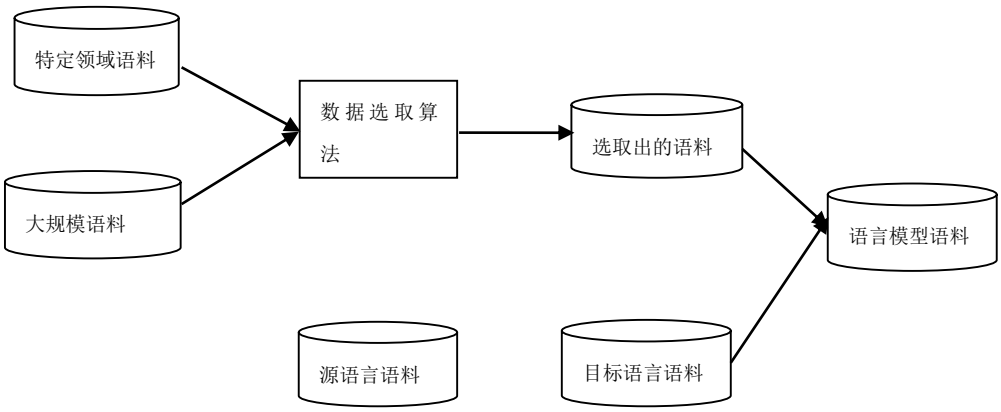

图 2 语言模型语料的选取过程

图 2 中特定领域的语料是人工提前选取好的某一领域或者测试集的小规模单语语料，大规模语料是包含各个领域的混合单语语料，具体的数据选择过程可以总结为三个步骤：

步骤 1：根据数据选取算法，计算大规模单语语料和特定领域语料的相似度得分，设置相似度得分的阈值，将高于该值的语料选取出来。

步骤 2：将选取出来的语料与用于训练 SMT 系统双语平行语料的目标语料组合，来训练语言模型。

步骤 3：根据 SMT 的翻译效果，选取不用的阈值，重新选取训练语料，重复上述步骤，直到得到一个最优化的翻译系统。

4.2.2 实验结果分析

本文做了一系列的实验来验证三种不同的数据选取方法，这三种方法分别是基于 TF-IDF 的数据选取方法（简称 TD-IDF）、基于编辑距离的数据选取方法(简称 ED)和本文所提出的方法(简称 TDE)。每一种方法分别从 400 万句的通用语料中选取 1.0%，1.25%，1.5%，1.75%，2.0%，2.5%,3%,5%的子集，按照上节中的方法组成训练语料,来训练语言模型。需要说明的是在试验中选取相似特定领域的语料，计算出来的相似度语料中含有重复的句子，本文在实际训练中会去掉重复的句子，这样在同等规模的语料中可以减少数据稀疏的问题，提高翻译质量。表 3 给出的是在不同比例下对于维汉翻译系统得到的实验结果：

表 3 不同比例下维汉翻译系统实验结果

语料大小	TD-IDF	ED	TDE
1.0%	33.96/33.05	34.3/33.45	34.78/33.65
1.25%	34.03/33.38	34.15/33.52	35.45/34.28
1.75%	34.17/33.46	34.37/33.78	35.81/34.53
2.0%	35.1/33.76	34.64/34.18	35.25/34.69
2.5%	35.21/33.78	34.27/34.42	35.68/34.49
3%	34.96/34.02	34.23/34.36	35.37/34.36
5%	34.92/34.16	34.15/34.28	35.42/34.47

表 3 中的数据斜线前面的数据是在新闻领域测试集上的 BLEU，斜线后的数据是在教育领域测试集上的 BLEU，通过表 3 的数据对比基准系统，首先可以明显的看出较之基准系统通过选取和测试集领

域相同的数据训练出来的语言模型，对于翻译性能有明显的提升，并且比之使用全部的单语语料训练语言模型，经过选取出领域相关的训练语料子集，不仅大大的减少了模型规模，并且还对翻译的效果有明显的提升。

再者对比表 2 和表 3 中的不同数据，会发现不同的方法对于翻译效果性能的提高也是变化的，但是对比于另外两种方法，本文所提出的方法有更加明显的提升，本文所提出的方法比另外两种方法在语料规模相同的情况下，BLEU 有更高的提升，而且比另外两种方法更加快速的收敛，使得 BLEU 有更大的提升。表 4 更能说明这一点。

最后分析表中的数据会发现，随着选取出来的语料子集的比例增加，对 BLEU 的提升逐渐的不明显，并且随着比例的不断增加，反而会使得 BLEU 的值下降，原因就是随着比例的不断的增加，选取出来的句子逐渐增加，最相似的句子选取出来后，次相似的句子随着比例的不断增加，也会被选取出来，这样的会使得次相似句子的噪音对翻译性能的影响愈加凸显，这说明在选取领域相关的语料时在选取相似的句子时也要注意噪声对性能的影响，选取出来更加与领域相似的训练语料会更有效的提高了 BLEU，这也说明了本文方法更优的原因，本文给出的方法对于相似的计算不是将所有的词视为相同权重，而是根据其对分类贡献度决定其权重，在通过编辑距离计算相似度，这样会将与领域真正相关的语料选取出来。

表 4 三种方法最好结果下选取数据的比例

数据选择方法	选取语料比例	BLEU
基准系统	0	33.70
TF-IDF	2.5%	35.21（+1.51）
ED	2%	34.84（+1.14）
TDE	1.25%	35.81（+2.11）

对比表 4 中是三种方法的翻译效果最好的情况下选取的语料比例大小，通过对比我们可以看出，不同方法对于噪音的消除效果是不同的，本文的方法能更加有效的减少噪音，更快的是系统收敛到最优化的效果。

5 总结与展望

本文提出了一种基于信息检索的 TF-IDF 权重和编辑距离相结合的方法。该方法是从大规模的单语语料中选择与特定领域相似的训练语料，训练一个优秀的语言模型以提高翻译质量。该方法的重点在于计算编辑距离时，用可变的权重表示编辑操作，对于具体的编辑距离的计算，将每一个编辑操作，都用单词的权重确定插入、删除和替换操作需要的代价，实验结果表明该方法可以明显提高机器翻译的质量。

下一步我们将首先分析停用词和命名实体对领域相关的训练语料选取的影响，然后进一步研究使用源语言选取目标语言的训练语料，为 SMT 选取更加准确的训练语料构建翻译系统。

参考文献

[1] Axelrod, X. He, and J. Gao. Domain adaptation via pseudo in-domain data selection[C]//in Proceedings of the Conference on Empirical Methods in Natural Language Processing，2011: 355–362.

[2] Y. Lü, J. Huang, and Q. Liu. Improving statistical machine translation performance by training data selection and optimization [C]// in Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning，2007: 343–350.

[3] UM-Corpus by NLP2CT is licensed under a Creative Commons Attribution NonCommercial NoDerivatives 4.0 International License[EB]. http://creativecommons.org/licenses/by-nc-nd/4.0/deed.en_US.

[4] OpenSubtitles2013[EB]. http://opus.lingfil.uu.se/OpenSubtitles2013.php

[5] Wang L and Derek.F. Edit Distance: A New Data Selection Criterion for Domain Adaptation in SMT[C]//Proceedings of Recent Advances in Natural Language Processing，2013: 727–732.

[6] 黄瑾,吕雅娟,刘群. 基于信息检索方法的统计翻译系统训练数据选择与优化[J]. 中文信息学报，2011，25(002)：72-77.

[7] J. Civera and A. Juan. Domain adaptation in statistical machine translation with mixture modeling [C]// inProceedings of the 2nd ACL Workshop on Statistical Machine Translation，2007: 177–180.

[8] A. S. Hildebrand, M. Eck, S. Vogel, and A. Waibel. Adaptation of the translation model for statistical machine translation based on information retrieval [C]// in Proceedings of the 10th Annual Conference on European Association for Machine Translation，2005: 133–142.

[9] S. C. Lin, C. L. Tsai, L. F. Chien, K. J. Chen, and L. S. Lee. Chinese language model adaptation based on document classification and multiple domain specific language models[C]// in Proceedings of the 5th European Conference on Speech Communication and Technology，1997: 355–362.

[10] R. C. Moore and W. Lewis. Intelligent selection of language model training data [C]// in Proceedings

of the 48th Annual Meeting of the Association for Computational Linguistics，2010：220-–224.
[11] Matthias Eck, Stephan Vogel, Alex Waibel. Language Model Adaptation for Statistical Machine Translation based on Information Retrieval [C]// The International Conference on Language Resources and Evaluation，2004：327–330.
[12] George Foster and Cyril Goutte and Roland Kuhn. Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation [C]//in Proceedings of the Conference on Association for Computational Linguistics，2010：451–459.
[13] Matsoukas, A. V. I. Rosti, and B. Zhang. Discriminative corpus weight estimation for machine translation [C]// in Proceedings of the Conference on Empirical Methods in Natural Language Processing，20109：708–717.
[14] Keiji Yasuda, Hirofumi Yamamoto and Eiichiro Sumita. Method of Selecting Training Sets to Build Compact and Efficient Statistical Language Model [C]// in Proceedings of the Conference on Association for Computational Linguistics，2007：31–37.
[15] ALe Liu Yu Hong, Hao Liu Xing Wang, Jianmin Yao. Effective Selection of Translation Model Training Data [C]//in Proceedings of the Conference on Association for Computational Linguistics，2011：569–573.