文章编号：1003-0077(2008)02-0040-07

基于信息检索方法的统计翻译系统训练数据选择与优化

黄 瑾[1,2,3]，吕雅娟[1,3]，刘 群[1,3]

(1. 中国科学院 计算技术研究所，北京 100080；2. 中国科学院 研究生院，北京 100080；
3. 中国科学院 智能信息处理重点实验室，北京 100080)

摘 要：双语平行语料库是构造高质量统计机器翻译系统的重要基础。与传统的通过扩大双语平行语料库规模来提高翻译质量的策略不同，本文旨在尽可能地挖掘现有资源的潜力来提高统计机器翻译的性能。文中提出了一种基于信息检索模型的统计机器翻译训练数据选择与优化方法，通过选择现有训练数据资源中与待翻译文本相似的句子组成训练子集，可在不增加计算资源的情况下获得与使用全部数据相当甚至更优的机器翻译结果。通过将选择出的数据子集加入原始训练数据中优化训练数据的分布可进一步提高机器翻译的质量。实验证明，该方法对于有效利用现有数据资源提高统计机器翻译性能有很好的效果。

关键词：人工智能；机器翻译；统计机器翻译；平行语料库；信息检索；数据选择

中图分类号：TP391 **文献标识码**：A

Corpus Selection and Optimization for Statistical Machine Translation System Based on Information Retrieval Method

HUANG Jin[1,2,3], LV Ya-juan[1,3], LIU Qun[1,3]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;
2. Graduate University of Chinese Academy of Sciences, Beijing 100080, China;
3. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: Parallel corpora are an indispensable resource for translation model training in statistical machine translation (SMT) system. Instead of collecting more and more parallel training corpora, this paper aims to improve the performance of SMT system by exploiting full potential of the existing parallel corpora. We propose an approach to select and optimize training corpus by using information retrieval method. First, sentences similar to the test text are selected to form a small and adapted training data. This allows us to get a comparable or even better performance with only a subset of the total data and the less hardware need. Second, we add the selected subset to the entire corpus to optimize the data distribution and get a better result. The experiments show that this method can effectively improve the performance of SMT system.

Key words: artificial intelligence; machine translation; statistical machine translation; parallel corpus; information retrieval; data selection

1 引言

统计机器翻译系统的训练需要使用大规模的双语平行语料库，一般来说，使用越多的数据来估计模型的翻译参数，训练得到的参数越稳定、越接近于真实情况，翻译质量越高。然而，大规模高质量的双语平行语料库并不容易获得，现有的多数平行语料库

收稿日期：2007-03-22 **定稿日期**：2007-05-17

基金项目：国家自然科学基金资助项目(60603095；60573188)

作者简介：黄瑾(1982—)，女，硕士，研究方向为自然语言处理、机器翻译；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；刘群(1966—)，男，研究员，博导，主要研究方向为自然语言处理和机器翻译。

中还包含着大量的噪声，包括拼写错误、错别字、错误的译文、构造语料库时段落对齐或者句子对齐错误而导致的翻译错误等，这些错误都会影响训练结果的可靠性；现有的语料库往往来源于不同的领域，例如 LDC 提供的平行语料库就来自香港法律、香港议会议事录、香港新闻等相差很大的若干领域[1]。实验表明，简单地将不同领域的数据资源混合进行训练并不总是能够提高统计机器翻译的性能。从另外一方面说，即使可以获取海量的数据资源，训练这些资源也还需要巨大的计算能力，这使得统计机器翻译可能无法适用于仅有有限计算资源的应用领域；而且随着所使用数据的不断增加，通过扩大数据规模可获取的性能提升也会越来越小。因此，在继续扩大双语平行语料库规模的同时，研究如何挖掘现有资源的潜力进一步提高统计机器翻译的性能具有积极的意义。

相对于统计机器翻译模型和算法方面的研究来说，目前关于如何有效利用双语语料库的研究还比较少。针对双语语料库的质量问题，陈毅东等人使用排序策略，利用一些预先选定的特征将已有的平行语料进行打分排序并只选取分数靠前的部分组成训练数据，这种策略采用的特征与待翻译文本无关，是一种判断语料本身“质量”好坏的方法，简单易行，有助于进行平行语料库的筛选工作[2]。信息检索模型先是被应用于语音识别的自适应语言模型中并取得了一定的效果[3]，后来这一思想被 Eck 和 Zhao 等人引入统计机器翻译领域，将首次翻译得到的候选翻译结果视为信息检索中的查询条件，在海量的单语语料库中检索出相似的数据子集并根据其训练得到自适应的语言模型，使用该语言模型重新进行翻译可显著地提高统计机器翻译的质量[4,5]。Almut 等人则尝试使用领域自适应的翻译模型，根据待翻译文本用信息检索工具从大规模其他领域的双语语料中筛选出部分相似数据，加入到已有的小规模同领域数据集合中进行训练，在一定程度上提高了翻译性能[6]。

本文使用信息检索模型，在不增加任何训练语料的条件下，通过从现有双语平行语料库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。数据选择方法使用相似数据集合生成小规模的自适应翻译模型，数据优化方法使用相似数据对原始语料库进行分布优化，使得选择与优化后的训练数据与待翻译文本具有更为相似的领域、主题及用词，得到更好的翻译模型，以达到提高统计机器翻译质量的目的。文章的第二部分介绍了信息检索模型的相关概念和相似数据检索的基本方法；第三部分和第四部分分别讨论了数据选择和数据优化的策略；第五部分给出了数据选择与优化策略的通用模型；实验与分析在第六部分给出，讨论了一系列汉英翻译实验的结果，分析不同的数据选择与优化方法对系统性能的影响，最后一部分进行了总结。

2　基于信息检索模型的相似数据检索

信息检索(Information Retrieval)是一个从文档集合中返回满足用户需求的相关信息的过程，信息检索模型主要关注如何表示用户查询(Query)和现有文档(Document)并对它们进行相似度的计算。具有代表性的检索模型主要有布尔模型(Boolean Model)、向量空间模型(Vector Space Model，简称 VSM)、概率模型(Probabilistic Model)等，这些模型从不同角度使用不同的方法对查询和文档之间的相关度进行建模，这里使用最常用的向量空间模型及 TF-IDF 相似度计算方法来实现相似数据的检索。

向量空间模型将用户输入的查询和系统中的文档都使用向量表示，假设共有 n 项(term，一般为单词)，则每篇文档(或查询) D_i 都可视为一个 n 维向量 $(w_{i1}, w_{i2}, \cdots, w_{in})$，此处的 w_{ij} 表示文档 D_i 中的第 j 维的权值，可按如下方法计算：

$$w_{ij} = tf_{ij} \times \log(idf_j) \tag{1}$$

其中，tf_{ij} 是指项 j 在文档 D_i 中出现的频次，tf_{ij} 的值越大，表示项 j 对于文档 D_i 越重要；而 idf_j 称为“逆文本频率指数”，为包含有项 j 的文档数目的倒数，计算时一般使用文档总数除以含有项 j 的文档数。idf_j 越小，包含项 j 的文档数目越多，表示项 j 在衡量文档相似性方面的作用越低。当用户输入查询条件时，检索系统通过计算查询与所有文档向量之间的相似度评价结果并依此进行排序，计算时常采用向量之间的夹角余弦或者内积来表示相似度的大小。

基于上述信息检索模型，我们将已有的双语平行语料库的源语言部分视为待查询的文档集，每一个源语言句子都视为一篇文档并建立索引，将待翻译文本的部分实例或与待翻译文本相似的句子作为查询条件，通过计算查询条件与文档之间的相似度，可获得一个排序的与待翻译文本汉语部分相似的数据子集。与原始数据集相比，选择出的数据子集与待翻译文本的领域、风格及用词可能更加接近，使用这种与待翻译文本更加相似的语料进行统计机器翻

译参数训练，应该有助于提高机器翻译的质量。具体的数据选择步骤如下：

1）对双语平行语料库的汉语部分建立索引；

2）对待翻译实例中的每个汉语句子，利用信息检索模型在上述语料库中进行检索；

3）根据相似数据的选择策略，选取每个句子检索结果中的位置排前的若干句子及其对应的译文组成新的平行语料库。

3 基于相似数据的训练语料选择

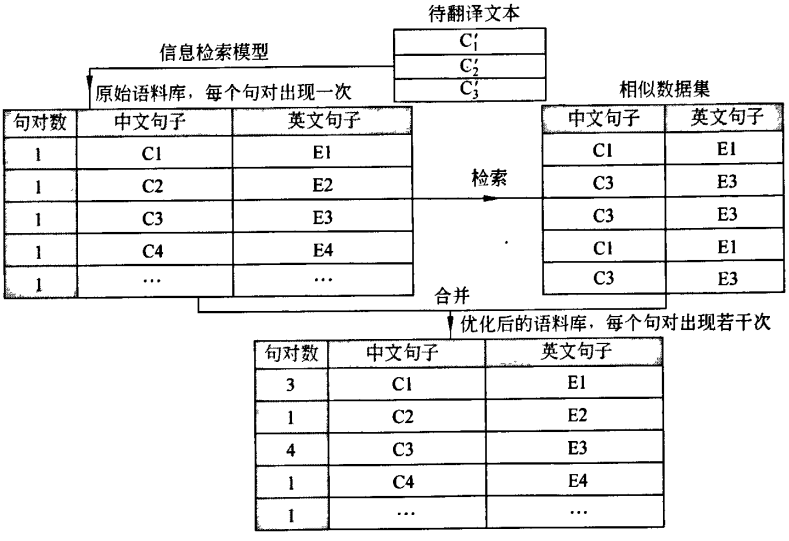
最简单的相似数据选择策略是直接使用检索出的前 n 个句子对组成新的训练语料，另一种方法是根据检索工具为每个候选结果打出的分数，选取某个阈值 γ 以上的句子及其译文构成训练语料。无论使用哪种策略，我们都将所有选择出的语料合并起来而不是单独训练每个句子的翻译模型，这是因为，统计机器翻译方法训练需要一定规模的数据，仅由几百个句子训练出来的模型会存在严重的数据稀疏现象。这样做的一个前提条件是，一般情况下，一个待翻译文本中的源语言句子往往具有较为一致的遣词造句特征，属于同一个领域，因此可以将这些选择出的数据合并成为一个整体进行训练。如果事实上已知待翻译文本来源于不同的领域具有较大的差异，应该将其分割开来分别进行训练并使用相应的参数来进行翻译。

需要考虑的一个问题是，无论使用上述哪种策略进行相似数据的选择，根据待翻译文本选择出的数据中均可能含有重复的句子。一般情况下，阈值 n 越大（γ 越小），选出的用于训练的句子总数越多，其中重复句子所占的比例也越大。重复的句子倾向于将翻译模型引向那些出现频次较多的词条及句型结构，而我们使用信息检索模型已经考虑了此类问题，保留这种重复对于训练模型提高翻译性能有所帮助。实验部分将删除相似数据中的重复的句子进行对比实验来说明这种重复对翻译性能的影响。

此方法在进行相似数据选择时只使用了待翻译文本本身（即源语言部分），相比于 Zhao 等人的语言模型自适应策略，此方法不需要预先翻译待译文本，简化了处理的过程[5]。另外，由于我们根据阈值只选取部分更为相似的数据组成训练语料，实际上所使用的数据规模要远小于原始语料，缩小了模型的大小，加快了训练及解码的速度，可缓解计算资源不足造成的数据训练瓶颈。

4 基于相似数据的训练语料优化

使用小部分选择出的自适应语料进行训练可以较大程度地减少实验中硬件条件的限制，但也可能造成一定的数据稀疏问题。基于相似数据的训练语料优化的思想在于，仍然使用全部的双语平行语料，但是人为地调整语料库的数据分布，加大与待翻译文本更为相似的数据集的权重，以达到优化翻译模型的目的。我们将检索得到的相似数据与原始的双语平行语料库进行合并，即将出现在相似数据集中的句子对在原始语料库中的出现次数相应增加，使其具有更大的权重值。使用合并以后的数据进行参数训练并送入解码器进行翻译。



句对数	中文句子	英文句子
1	C1	E1
1	C2	E2
1	C3	E3
1	C4	E4
1	…	…

中文句子	英文句子
C1	E1
C3	E3
C3	E3
C1	E1
C3	E3

句对数	中文句子	英文句子
3	C1	E1
1	C2	E2
4	C3	E3
1	C4	E4
1	…	…

图 1 训练数据优化策略示意图

图 1 为优化合并的数据走向图，原始语料库中与待翻译文本中的源语言句子相似的部分被筛选出来，构成了有重复句对出现的相似数据集，然后与原始语料库进行合并，这里没有新的句对被引入，因此不会增加模型的规模，只是原始语料库中各个句对的出现次数发生了一定的变化，那些与待翻译文本更为相似的句对被赋予了更高的权值，使得训练得到的翻译参数更适应于当前待翻译文本的翻译。由于所使用的模型训练工具支持句子加权的训练模式，这种方法不会增加训练模型所需的时间和空间。

5　基于相似数据的数据选择与优化通用模型

本文中所采用的信息检索方法只是求取相似数据的一种具体实现方法，在应用过程中还可以考虑使用其他的相似数据计算方法，其通用模型可如下表示：

$$w_i = \alpha + \beta \sum_{j=1}^{K} \theta(i,j) i = 1,2,\cdots,M \qquad (2)$$

其中，w_i 表示原始训练语料库中的句子 S_i 在最终训练语料库中的权重。系数 α 用于确定原始训练语料库中的句子本身的权重，系数 β 用于确定选择出的相似数据的权重。M 表示训练语料库中的句子数目，K 表示测试集中的句子数目，$\theta(i,j)$ 表示 S_i 与测试集中的句子 S_j 的某种相似性度量，可以采用任意的相似性计算策略。其中本文提出的基于 topN 的数据选择与优化策略即可相应地表示为：

$$\beta = 1;$$

$$\theta(i,j) = \begin{cases} 1, & \text{如果 } \mathrm{score}(S_i,S_j) \in \underset{N}{\mathrm{top}}\{\mathrm{score}(S_{i'},S_j), \\ & i' = 1,2,\cdots,M\} \\ 0, & \text{否则} \end{cases} \qquad (3)$$

如果按照阈值 γ 进行选择，则公式可表示为：

$$\beta = 1;$$

$$\theta(i,j) = \begin{cases} 1 & \text{如果 } \mathrm{score}(S_i,S_j) \geqslant \gamma \\ 0, & \text{否则} \end{cases} \qquad (4)$$

其中 $\mathrm{score}(S_i,S_j)$ 表示使用信息检索方法求得的训练语料库中的句子 S_i 与测试集中的句子 S_j 的分值。当 α 取值为 0 时为数据选择方法；当 α 取值为 1 时为数据优化方法。

我们还初步尝试了其他的相似度计算方法，例如使用 Dice 系数作为上述相似度评测量度，Dice 系数通过统计两个句子中共现词的个数来表示句子相似性，计算公式如下：

$$\mathrm{Dice}(S_i,S_j) = \frac{2 \times |S_i \cap S_j|}{|S_i| + |S_j|} \qquad (5)$$

通用公式可表示如下：

$$\beta = \frac{10}{K};$$
$$\theta(i,j) = \mathrm{Dice}(S_i,S_j) \qquad (6)$$

上述 K 值为测试集句子数目，β 的设置策略是将所有测试句子与当前训练语料库中的句子的 Dice 系数值求平均，由于 Dice 系数的取值范围在0～1之间，我们根据经验为这个平均值乘以了一个系数 10。

6　实验与分析

实验数据使用 LDC 发布的双语平行语料库，根据其来源不同数据可分为如下几个部分：

表 1　训练语料描述

训练语料	LDC 编号	说　明	句对
FBIS	LDC2003E14	FBIS Multilanguage Texts 新闻语料	200 000
HK_Hansards	LDC2004T08	随机选自 Hong Kong Parallel Text 香港议会记录	200 000
HK_News	LDC2004T08	随机选自 Hong Kong Parallel Text 香港新闻	200 000
Baseline	—	以上全部数据	600 000

上述各个子领域语料规模相当，将所有子语料合并形成的全部数据称为 Baseline。实验中所使用的翻译系统为中科院计算技术研究所多语言交互技术实验室开发的基于短语模型的统计机器翻译系统①，系统实现采用对数线形模型[7,8]。其中，词汇对齐训练采用了 GIZA++ 工具[9]，短语抽取使用了文献[10]中的方法，对数线性模型的参数训练使用最小错误率训练[11]，语言模型为使用 Gigaword Xinhua 语料库②的英文部分训练而成的一个四元语言模型（语言模型使用 SRI 语言模型工具[12]训练）。实验中使用 NIST2002 的测试集作为开发集来训练解码器参数，NIST2005 年的测试集作为测试文本，机器翻译性能的评测标准为 BLEU4[13]，评

① 系统演示见：http://mtgroup.ict.ac.cn/mtdemo/

② 语料库说明见：http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T09

测工具为 mteval-v11b.pl[①]，大小写敏感。所有的信息检索任务由 Lemur[14]完成。

6.1 Baseline 实验

首先使用各个子领域的语料分别训练翻译模型参数并利用开发集训练对数线性模型的特征权值。实验结果如表 2 所示。

从表 2 中可以看到，尽管 3 个子领域的训练数据规模相当，利用它们训练出来的翻译模型的翻译效果却相差很大。由于 FBIS 和测试语料比较相似，因此训练出的翻译模型的效果远远好于其他两个语料库。由此可以看出选择适合的训练语料库的重要性。将所有数据混合训练出来的 Baseline 模型好于各个子模型，这是由于训练语料规模的增加有利于解决数据稀疏问题。但是相对于 FBIS 模型来讲，尽管 Baseline 模型的训练语料规模增加了三倍，但是翻译结果的提高并不是特别显著。这说明将不同领域的语料库进行简单的合并并不一定能提高翻译质量。特别是当训练数据达到一定规模后，盲目扩大训练语料库规模将加重系统训练和解码算法的计算资源负担，但是对于改善翻译质量却不一定能达到理想的效果。因此，研究有效的数据选择与优化策略是非常必要的。

表 2 Baseline 系统实验结果

训练语料	开发集 BLEU 值	测试集 BLEU 值
FBIS	**0.261 4**	0.233 1
HK_Hansards	0.167 9	0.162 4
HK_News	0.174 8	0.160 8
Baseline	0.256 5	**0.236 3**

6.2 基于信息检索模型的相似数据选择实验

首先，我们为全部数据的汉语部分建立索引，使用开发集和测试集中的汉语句子作为查询条件送入信息检索工具 Lemur 中，系统为每个句子输出若干个相似候选；根据输出结果依次选出最好的 n＝100，200，500，1 000，2 000 个句子组成训练语料子集，将其送入训练工具中进行参数训练，并使用训练得到的参数进行统计机器翻译。随着 n 值的增大，所选出的相似数据集中重复的句子所占比例也增大。

需要指出的是检索工具返回的检索句子数有时候会少于指定的 n 值，一般情况下较短的句子返回的结果会更少一些。表 3 显示的是对不同的 n 值实验得到的结果，这里分别给出了对选择出来的语料库不去重复句子和去掉重复句子的实验结果。

表 3 训练语料选择实验结果

	全部训练语料句对数	语料中不重复句对数	翻译模型规模	不去重的翻译结果 BLEU 值	去重的翻译结果 BLEU 值
Baseline	600 000	600 000	2.41G	0.236 3	0.236 3
Top100	196 000	91 804	0.43G	0.230 6	0.234 6
Top200	392 000	150 619	0.73G	0.236 0	0.234 5
Top500	980 000	261 003	1.28G	0.241 5	0.237 0
Top1000	1 960 000	357 337	1.74G	**0.246 3**	**0.237 6**
Top2000	3 920 000	445 890	2.11G	0.235 1	0.234 6

实验结果表明：

1. 基于相似数据的训练语料选择策略，只使用全部语料的部分子集即可达到比使用全部数据更好的结果，说明更相似的数据是更好的数据。当 n 的值为 200 时，仅使用整个语料的四分之一数据就已经可以达到使用全部语料的翻译性能，模型规模也有很大程度的减少。模型的大小对于一些对资源要求很高的翻译应用来讲是非常重要的。

2. 统计机器翻译的质量随着所使用的相似句子数目 n 值的变化存在一个“峰值”，在达到这个峰值之前，随着 n 值的增大，所使用的语料越多，统计机器翻译的质量越好。当相似句子数目 n 超过一定规模时，系统翻译质量不升反降，这是由于随着相似数据规模的增加，一些不很相似的噪声语料也逐渐被加入进来，造成了数据质量的下降。在实际应用中可以通过开发集的翻译结果来确定 n 值的大小。

3. 总体看来，相似数据不进行排重的效果要优于排重的效果。这也说明对于特别相似的数据增加

① 评测工具下载见：http://www.nist.gov/speech/tests/mt/resources/scoring.htm

权重有利于提高翻译模型训练的质量，验证了我们的训练数据优化策略思路的正确性。

我们进行了其他的数据选择策略实验，只选择检索到的分值在某个阈值 γ 以上的数据生成子集，其实验结果见表 4。

表 4　使用分值作为阈值的训练语料选择实验结果

	全部训练语料句对数	语料中不重复句对数	翻译结果的 BLEU 值
Baseline	600 000	600 000	0.236 3
Top0.4	135 890	71 603	0.218 5
Top0.3	474 989	183 114	0.235 9
Top0.2	1 763 152	350 976	**0.243 7**

使用检索工具给出的分值作为选择相似数据的阈值，同样可以达到使用少部分训练数据即得到更优的翻译结果的目的，而且使用分值作为阈值可避免引入过多“不相似”数据，在一定程度上避免了过度训练情况的出现。但是，从使用的不重复句对的数目上看，Top0.2 与 Top1000 使用数据规模类似时，翻译质量略差于后者。

6.3　基于相似数据的训练数据优化实验

我们将数据选择实验中检索出的最优的 $n=100,200,500,1\,000$ 及 2 000 个句子加入到全部数据中改变原有数据的分布，重新训练翻译模型，实验结果如表 5 所示：

表 5　训练数据优化实验

	语料中不重复句对数	翻译结果的 BLEU 值
Baseline	600 000	0.236 3
Top100+	600 000	0.238 7
Top200+	600 000	0.244 3
Top500+	600 000	**0.246 1**
Top1000+	600 000	0.243 1
Top2000+	600 000	0.235 5

数据优化策略并没有增加所使用数据的规模，只是通过数据加权优化了整个语料中的数据分布。上面实验结果表明，使用数据优化的方式可以进一步提高系统翻译的性能。在 n 值相同的情况下，数据优化方法取得的结果要优于直接使用子集进行训练得到的结果。这是由于使用大规模数据有利于解决训练中的数据稀疏问题，在适应领域翻译的同时，使得到的模型更加稳定。

为了避免单一测试集可能存在的数据偏向，我们使用 NIST2004 年的测试数据集重复了上述实验。实验证明，在不同的测试集上使用数据选择与优化方法同样可提高系统的翻译质量。

表 6　训练数据选择实验（NIST04 测试集）

	语料中不重复句对数	NIST04 BLEU 值
Baseline	600 000	0.268 7
top100	119 867	0.266 5
top200	189 004	0.269 0
top500	306 672	0.275 9
top1000	398 037	**0.276 0**

表 7　训练数据优化实验（NIST04 测试集）

	语料中不重复句对数	NIST04 BLEU 值
Baseline	600 000	0.268 7
top100+	600 000	0.276 7
top200+	600 000	0.276 9
top500+	600 000	0.277 0
top1000+	600 000	**0.277 9**

另外，我们初步实验了使用 Dice 系数作为相似度度量的数据选择与优化策略，在 NIST2005 测试集上的实验结果如表 8 所示。

从表 8 中可以看出，以 Dice 系数作为句子相似度计算方法同样可以提高机器翻译系统的翻译质量，但是取得的效果不如信息检索模型 TF-IDF 方法。这一方面是因为 TF-IDF 计算方法中的词语的逆文档频率综合考虑了不同的词在相似度计算中的权重，另一方面是由于两种方法在计算权重时采取的策略稍有差别，前者采用句子计数的方式，后者采用累加分值的策略。表中所示的数据优化结果略差于数据选择结果，这是因为通过计算 Dice 系数的累加和来确定句子的权重，使得训练语料中的句子区分度降低，从上表中所使用的“不重复的句子对数目”即可看出，初步尝试的这种相似数据选择方法可能引入过多对提高语料质量贡献并不大的句对，再进行合并时弱化了数据优化策略对训练语料数据分布的影响，降低了数据选择以及优化方法对系统质量的提高效果。

表8 基于Dice系数的数据选择与优化实验

	语料中不重复句对数	NIST05 BLEU值
Baseline	600 000	0.236 3
Dice系数	587 419	0.239 3
Dice系数+	600 000	0.238 7

7 结论

本文提出了一种基于信息检索模型的统计机器翻译训练语料选择与优化的方法。利用信息检索模型在已有的双语平行语料库中选择出与待翻译文本相似的数据构造自适应的训练语料。在此基础上通过加权调整已有资源中的数据分布，在不增大数据规模的基础上可生成更为优化的模型参数。实验证明，文中提出的方法使用小规模数据即可达到甚至超过使用全部数据可获得的机器翻译质量，通过优化数据分布可进一步提高翻译的性能。本文的实验也表明，在增大训练语料规模的同时，选择适合的训练语料对于提高机器翻译的质量也很重要。

本文中所采用的信息检索方法只是求取相似数据的一种具体实现策略，在该方法的应用过程中还可以考虑使用其他的相似性计算方法，文中给出了该方法的通用模型并在实验中初步尝试了使用Dice系数的相似数据度量。另外，本方法适用于待翻译文本或其领域已知的情况，针对需要实时地进行领域适应的真实翻译场景，我们在另外一篇文章中提出了一种基于信息检索模型的在线模型参数优化方法，可实时地选择优化翻译模型参数，提高系统翻译的质量[15]。

下一步的工作中，我们将继续完善数据选择和加权优化的方法。尝试信息检索模型中其他的相似度算法；针对信息检索相似度模型对单词出现顺序不加考虑的特点，引入与词序有关的模型方法对检索出的相似数据集进行重排序；按照检索结果的得分或者顺序进行更为复杂的加权处理方法等；此外，我们还将考虑进一步引入自适应的语言模型。

参考文献：

[1] LDC (Linguistic Data Consortium) [EB/OL]. http://www.ldc.upenn.edu/.

[2] 陈毅东，史晓东，周昌乐. 平行语料库处理初探：一种排序模型[J]. 中文信息学报，2006，增刊：66-70.

[3] Milind Mahajan, Doug Beeferman, X. D. Huang. Improved Topic-Dependent Language Modeling Using Information Retrieval Techniques [A]. IEEE International Conference on Acoustics, Speech and Signal Processing[C]. 1999, Volume 1: 541-544.

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

[5] Zhao Bing, Matthias Eck, Stephan Vogel. Language Model Adaptation for Statistical Machine Translation with structured query models [A]. The proceeding of The 20th International Conference on Computational Linguistics[C]. 2004.

[6] Almut Silja Hildebrand et al, Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval [A]. EAMT 2005 Conference Proceedings[C]. 2005.

[7] Franz Josef Och, Hermann Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation [A]. In: ACL 2002: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics[C]. 2005. 295-302.

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

[9] The Giza++ Toolkit [EB/OL]. http://www.fjoch.com/GIZA++.html.

[10] Richard Zens, Franz Josef Och, Hermann Ney. Phrase-Based Statistical Machine Translation [A]. M. Jarke, J. Koehler, G. Lakemeyer. KI - 2002: Advances in artificial intelligence. 25. Annual German Conference on AI [C]. KI 2002, Vol. LNAI 2479, 18-32.

[11] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation [A]. Proceeding of the 41st Annual Meeting of the Association for Computational Linguistics[C]. July 2003. 160-167.

[12] The SRI Language Modeling Toolkit [EB/OL]. http://www.speech.sri.com/projects/srilm/.

[13] Papineni kishore et al. BLEU: a Method for Automatic Evaluation of Machine Translation [A]. Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics[C]. 2002. 311-318.

[14] The LEMUR Toolkit [EB/OL]. http://www.cs.cmu.edu/~lemur/.

[15] Lu Yajuan, Huang Jin and Liu Qun. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization[A]. Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning (EMNLP)[C]. 2007.