基于句对质量和覆盖度的统计机器翻译训练语料选取

姚树杰[1,2] 肖桐[1,2] 朱靖波[1,2]

1. 东北大学自然语言处理实验室，辽宁沈阳 110004

2. 医学影像计算教育部重点实验室（东北大学），辽宁沈阳，110819

E-mail: yaosj@ics.neu.edu.cn, xiaotong@mail.neu.edu.cn, zhujingbo@mail.neu.edu.cn

摘要：本文研究的目的是在待翻译文本未知的情况下，从已有的大规模平行语料中选取一个高质量的子集作为统计机器翻译系统的训练语料，以降低训练和解码代价。本文综合覆盖度和句对翻译质量两方面因素，提出一种从已有平行语料中获取高质量小规模训练子集的方法。在 CWMT2008 汉英翻译任务上的实验结果表明，利用本文的方法能够从现有大规模语料中选取高质量的子集，在减少 80%训练语料的情况下达到与 baseline 系统（使用全部训练语料）相当的翻译性能（BLEU 值）。

关键词：句对质量评价，覆盖度，统计机器翻译，线性句对质量评价模型，训练语料选取

Selection of SMT training data based on Sentence Pair Quality and Coverage

Yao Shujie[1,2], Xiao Tong[1,2], Zhu Jingbo[1,2]

1.Natural Language Processing Lab, Northeastern University Shenyang,Liaoning, P.R.China, 110004

2. Key Laboratory of Medical Image Computing (Northeastern University), Ministry of Education, Shenyang, Liaoning, P.R.China, 110819

E-mail: yaosj@ics.neu.edu.cn, xiaotong@mail.neu.edu.cn, zhujingbo@mail.neu.edu.cn

Abstract: In Statistical Machine Translation, effective selecting training data can generally reduce the burden of system training and decoding. Addressing this issue, considering both coverage and sentence pair quality, we proposed a framework to select a small portion from the whole training data set for SMT training. In this framework, two factors - coverage and sentence pair quality - are considered to model the data selection problem. Experimental results on CWMT2008 Chinese-to-English MT task show that our framework is effective to select a subset from the large training data set. Even trained on the 20% data selected by our framework, the SMT system can achieve comparable performance with the baseline system (using all the training data)

Key word: sentence pair quality evaluation, coverage, statistical machine translation, linear sentence pair quality evaluation model, training data selection.

1. 引言

在统计机器翻译(Statistical Machine Translation, 简写为 SMT)领域[5][6]，系统的训练需要有大规模的高质量双语句对语料库的支持。一般来说增加训练语料规模有助于获得稳定的模型参数和 SMT 系统翻译性能的提高。但是训练语料越多，训练和解码需要的时间越长，并且平行语料中存在的一些噪声数据，也会影响到训练的可靠性。

吕雅娟[1][5]等人曾提出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，她

们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，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好的机器翻译结果。

但是，在实际应用中，待翻译文本往往是未知的，Eck 等[2]对不依赖于待翻译文本的训练语料选取技术进行了研究。他们提出一种基于 n-gram 的覆盖度的方法来构建一个较小规模的训练语料子集，并且用这个子集来达到一个和原始全部语料相比可观的翻译性能。

此外，多数平行语料库包含着错误或噪音，它们也会对统计机器翻译系统的性能产生影响。如果能对双语语料（句对）进行有效地评价，也会有助于除去噪声，选择更加优质的数据来训练统计机器翻译系统。针对双语语料的质量评价的问题，陈毅东，史晓东[4]等曾研究了一种面向处理平行语料库的筛选的排序模型。这个模型利用预先设定的特征将已有的平行语料进行打分排序，之后选取分数靠前的部分组成训练语料。

为了更有效地对统计机器翻译语料进行筛选来降低 SMT 系统训练和解码的代价，本文提出了一种从大规模训练语料中选取小规模高质量子集的方法。该方法同时考虑了语料本身的质量和整体的覆盖度因素来选取训练语料。实验结果表明本文的方法在近百万规模训练语料上取得了明显的效果，使用选取的小规模(原始语料的 20%)数据即达到了与使用全部数据时相接近的翻译性能。

2. 训练语料子集选取框架

本文提出方法的基本框架为：输入原始大规模训练语料；首先对每一句对的质量进行自动评价并给出一个分数；然后，按质量评价分数的高低对句对排序；在句对按质量排序的基础上考虑覆盖度的因素，动态选取一个子集；输出从原始语料中选取的子集作为 SMT 系统的训练语料。

图 1 基于句对质量评价和覆盖度的训练语料子集选取框架

整个框架大致分为两个部分：句对质量的评价和基于覆盖度的训练语料选取。利用②整合不同的特征来综合评价句对质量(见第 3 节)。③整个语料的候选句对按质量评价分数的高低排序；④考虑覆盖度选取语料的一个子集作为训练数据(基于覆盖度选取训练语料的流程在第 4 节做详细描述)。下面对句对质量的评价和基于覆盖度的训练语料选取技术进行讨论。

3. 句对质量评价

从现有语料选取一个高质量的相对规模较小的训练子集，就单个句对来讲，我们希望优先考虑的是那种质量较好的个体。假定质量高的句对需满足以下条件：①构成句对的源语句和目标语句都是比较流畅的句子。②源语句和目标语句的互译比较准确。基于这样的考虑，本文提出一种线性模型整合不同特征来综合评价句对的质量，后面将详细介绍。

3.1 句对质量评价方法

为描述双语句对的好坏本文引入三类特征：基于双语词典的翻译质量，语言模型，翻译模型概率。最后，在 3.1.4 中本文提出一种线性模型整合这些特征来综合评价句对质量。

3.1.1 特征一：基于双语翻译词典的翻译质量

利用现有双语翻译词典，本文给出下式来评价句对翻译质量：

$$P_{dic}(s,t)=\sqrt{\frac{\sum_{w_s} Translate\ (w_s)}{length(s)}\times\frac{\sum_{w_t} Translate\ (w_t)}{length(t)}} \tag{1}$$

其中，s是表示源语言句子，t表示目标语句子；w_s和w_t分别表示双语句对源语句中的词和目标句的词；$length(s)$和$length(t)$分别表示源语句和目标语句的长度（即包含的词的个数）；$\sum_{w_s} Translate(w_s)$表示源语句中所有在目标句能找到译文的词的总数；$\sum_{w_t} Translate(w_t)$表示目标句中所有在源语句能找到译文的词的总数。对于$Translate(w)$，如果词w在它对应的目标与句子中存在翻译项则为 1，否则为 0。

3.1.2 特征二：语言模型

引入语言模型的目的是考察每一句对的单语部分是否流畅。本文把候选训练语料句对的源语句语言模型和目标语句的语言模型作为两个特征加入到句对质量评价线性模型中。假设句子中单词w_i的出现概率仅与其前面的 N-1 个单词有关，句长为 n 的句子w用语言模型概率来考察候选句的流畅度表示如下：

$$P_{LM}(w)\approx\sqrt[n]{\prod_{i=1}^{n}P(w_i|w_{i-N+1}w_{i-N+2}w_{i-N+3}\cdots w_{i-1})} \tag{2}$$

其中$P_{LM}(w)$的下标 LM 是 Language Model 的简写。语言模型参数在大规模双语训练语料上训练得到。实验中对句对的中文句和英文句分别计算其五元语言模型(N=5)，每个句子的语言模型按句子长度进行了归一化处理。

这里举一个 N=2 的例子，用语言模型概率来衡量句子“我是个学生”的流畅度。从训练语料估计到 P(我|<s>)=0.05, P(是|我)=0.01, P(个|是)=0.2, P(学生|个)=0.03，句子长 n=4 那么 $P_{LM}(\text{我 是 个 学生})=\sqrt[4]{0.05\times0.01\times0.2\times0.03}=0.0416$。

3.1.3 特征三：翻译模型概率

本文对 IBM model 1 翻译模型在假设基础上进行了进一步简化，并计算句对源语言到目标语和目标语到源语句翻译概率作为衡量一个句对翻译质量的特征。

对于句对(f,e)，假定源语句 f 有 m 个词，目标语句 e 有 l 个词。假设所有源语句的词至多有一个目标语词对齐，对齐概率只依赖于$t(f_j|e_i)$，对于每一个源语单词我们在目标语中寻找一个最能解释它的目标语词；每个源语句中的词仅由和它对齐的那个目标语词生成；忠诚度不依赖于目标语和源语句的长度。在此基础上，用下面的式子表示每一句对目标语对源语翻译的忠诚度。

$$P_{TM}(f|e)=\sqrt[m]{\prod_{j=1}^{m}max(t(f_j|e_i))} \tag{3}$$

其中，$t(f_j|e_i)$表示句子 e 的第 i 个词到句 f 第 j 个词的翻译概率。$P_{TM}(w)$的下标 TM 是 Translate Model 的简写。源语句对目标语句的忠诚度也类似表示。

3.1.4 句对质量评价线性模型

怎样考虑前述的特征来综合评价句对质量的好坏？用Q(f,e)来表示句对(f，e)的质量，本文通过下面的表达形式整合以上提到的特征：

$$log(Q(f,e)) = \sum_{i=1}^{k} w_i \log(P_i) \quad (4)$$

k表示该模型整合的特征的个数。e 与 f 分别表示句对的源语句和目标语句；这里w_i分别表示每个对应特征的权重，每个权重可在人工构造的少量训练集上通过自动或人工的方法得到。为实现的方便，本文暂时采用了人工的方法。

本文相关实验 k=5，P_1到P_5依次指$P_{dic}(f,e)$, $P_{LM}(e)$, $P_{LM}(f)$, $P_{TM}(f|e)$, $P_{TM}(e|f)$。

4. 基于覆盖度的训练语料选取

4.1 考虑覆盖度的动机

从原始语料中选取一个子集作为训练语料，是要用有限的语料覆盖尽可能多的语言现象，句对之间也不应该存在太多冗余。假如说句对质量评价是考虑这种语言现象的可靠性，那么覆盖度就是要保证要包含广泛的语言现象。本文认为一个较好的训练子集要有足够的覆盖度，并且本文的有关实验也表明，相同规模的数据，高的冗余会导致不好的训练效果，这也是本文在选取训练子集时考虑覆盖度的一个原因。

4.2 基于覆盖度的训练语料子集选取

本文比较了包括 n-gram 在内的三种不同覆盖度，采用一种动态的考虑覆盖度的方法来重新分布训练语料，最后从重新分布的语料中取前 N 个句对构成一个子集作为训练语料。

覆盖度大小的衡量分别比较三个参考指标：①词的覆盖；②n-gram（包括 Unigram Bigram Trigram）的覆盖；③短语翻译对的覆盖。

参照覆盖度选取训练语料子集：用候选训练语料的第一个句对作为所选出的训练语料子集的第一个元素，然后依次向后扫描候选语料，如果当前的句对对增加已选训练语料子集覆盖度有贡献(比如包含新的短语翻译对)，则优先选择这个句对添加到训练语料子集。

5. 结合句对质量和覆盖度的训练语料选取框架

本文的平行语料选取框架综合考虑了句对质量和覆盖度，利用句对质量评价线性模型将候选语料的全部句对按质量打分排序，之后按 4.2 所述的考虑覆盖度选取训练子集的方法从按句对质量排序的语料中选出一个子集作为训练语料，具体如算法 1 所示。

6. 相关实验与分析

6.1 Baseline 系统

Baseline 系统描述：本文实验所使用的统计机器翻译系统为东北大学自然语言处理实验室开发的基于短语的统计机器翻译系统[7][8]，系统实现采用对数线性模型。分词采用东北大学自然语言处理实验室分词系统；词对齐使用 GIZA++工具。实验数据使用 CWMT2008 语料预处理后的 70 万，将句对的分布先后顺序随机排列，从首句对起顺次分别取 1%，5%，10%，20%，(30%)，40%，60%，80%，和 100%作为 Baseline 训练语料，利用 SMT 系统的 BLEU 值来估计这些不同规模训练数据的质量。另外的一些实验相关信息如表 1 所示。

算法 1 基于句对质量和覆盖度的训练语料选取

输入：候选平行语料 $D=\{(s_1,t_1),(s_2,t_2),...\}$

输出：选出的小规模的训练语料

算法：

step1. 循环 For i from 1 to n //i 表示句对编号
用句对质量评价线性模型给句对(s_i,t_i)打分；

step2. 所有句对按 step1 得到的句对按质量分数高低排序，得到重排序的训练语料集D'_q；

step3. 从前往后扫描D'_q，按优先考虑覆盖度的方法选出一个子集，输出这个子集作为所选训练语料；

表 1 一些实验相关信息

评价方法	利用 SMT 系统的 BLEU 值[1]来估计所用训练数据的质量
训练数据	CWMT08（原始 85 万句对，预处理后为 70 万）
开发集	863-2005-writ（489 句+4ref）
测试集	SSMT07 汉英翻译测试数据（1002 句）
SMT 系统	a)BTG（Bracket transduction grammar）框架
	b)最大熵调序模型
	c)Beam search + cube pruning (beam size = 20)

6.2 不同覆盖度指标的比较实验

只考虑覆盖度，用 4.2 提到的方法分别以词，n-gram 和短语翻译对（Phrase pair）为覆盖度指标，从原始未经句对质量评价的语料中选取不同规模子集作为训练语料，其效果与 Baseline 做了比较。需要注意的是：词是指源语言出现的词（Unigram 除去禁用词）；短语翻译对从候选的平行句对获得，参考了[3]中的方法；n-gram 实验中 n=1,2,3，同时包含 Unigram Bigram 和 Trigram。

实验结果如图 2，纵坐标表示选取不同规模语料作为训练数据所达到的机器翻译性能（用 BLEU 值表示），横坐标表示所用数据占整个原始语料的百分比。不难发现在选取的语料规模比较小时，优先考虑语料的覆盖度，能够很大程度上影响 SMT 系统的训练效果，并且相同规模上用短语翻译对（Phrase pair）作为覆盖度指标选取的语料训练效果要好于基于词（unigram）和基于 n-gram（unigram~trigram），三个指标中使用短语翻译对达到的效果最明显。

Baseline 不考虑覆盖度随机选取数据作为训练语料，至少用 60%训练语料达到 BLEU 值（0.2398）与用全部语料时的 BLEU 值（0.2424）相接近。而考虑覆盖度来选取，基于词（Unigram）选取 40%语料达到 0.2411，n-gram（unigram~trigram）选取 40%达到 0.2396，而以短语翻译对覆盖度选取仅占全部候选语料 20%的数据就达到了 0.2404，与使用全部语料的水平（0.2424）相接近。而 Baseline 用 20%的数据达到的性能仅为 0.2277。实验结果表明训练语料的覆盖度对训练效果有很大影响，尤其当要选取的语料规模较小时覆盖度就显得更加重要。

通过这个实验的结果，也不难看到考虑覆盖度来选取语料子集要比随机选取的相同规模的语料的训练效果好；另外分析所用的三个覆盖度指标，词或 n-gram 作为覆盖度指标仅考虑了单语，而短语翻译对覆盖度指标是在词对齐基础上同时考虑双语信息，相比其他两个对选取高质量 SMT 平行训练语料的影响更大。

6.3 综合考虑句对翻译质量和覆盖度来选取语料实验

用 3.1.4 提到的句对质量评价线性模型来评价候选句对的质量。实验中本文暂时采用了人工的方法来设定各个特征的权重：权值开始设置为 1，然后人工观察在较小训练集合上的自动句对

[1] 实验中的 BLEU 值为大小写不敏感的 NIST BLEU.

质量评价结果，之后再根据这个结果的合理性，对权值进行调整，如此反复多次，最后每一个特征的权重由人工给定一个认为合理的经验值。

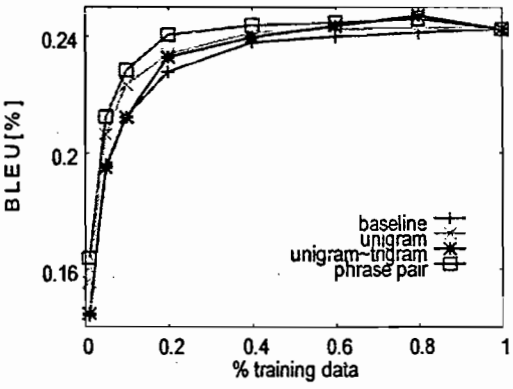

图 2 依不同覆盖度指标选取的语料的训练效果比较

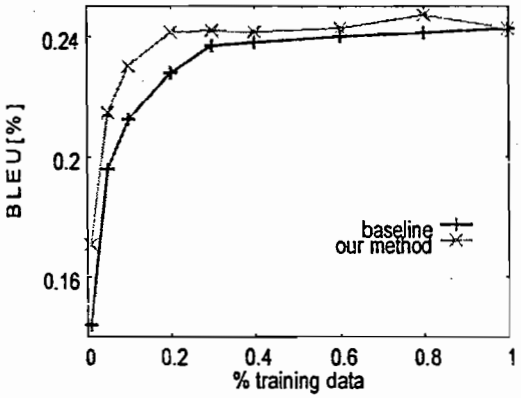

图 3 本文方法选取的不同规模训练语料训练与 Baseline 的比较

实验中$w_1 - w_5$分别为 0.1，0.5，0.5，0.5，0.5。另外，实验中选用短语翻译对作为覆盖度指标。综合考虑句对质量和覆盖度，按照图 1 所示整个框架流程来选取训练语料。选取的训练语料子集分别为全部原始语料规模的 1%，5%，10%，20%，30%，40%，60%，80%，100%，并与 Baseline 作对比。图 3 中 our method 曲线表示利用本文提到的框架，综合考虑句对质量和覆盖度选取的训练语料所达到的翻译性能。可以看出，利用本文的方法从较大规模平行语料中选取较小的子集作为训练语料能使机器翻译性能明显高于 Baseline，甚至用 20%的句对就到达了与用全部训练语料时相接近的性能。实验表明本文所提出的方法用在高质量训练语料子集的选取上有效。

6.4 引入句对质量评价的影响

评价本文句对质量评价模型不是件很容易的事，我们通过比较引入句对翻译质量评价前后所选取的相等规模的数据的训练效果来间接考察句对质量评价方法的有效性。

通过比较两组实验的数据可以发现，在句对质量评价基础上考虑覆盖度选取训练语料子集的效果要优于单纯考虑覆盖度；但是这种优势在本文实验所用数据上并不很明显。反映在 BLEU 值上如表 2（这里的覆盖度仅指短语翻译对的覆盖）。可以看出，综合考虑句对质量和覆盖度来选取小规模的训练语料能够比单纯考虑覆盖度更好些，尽管在本文目前所用数据的实验结果上并不是很明显。

表 2 引入句对质量评价前后按覆盖度选取的训练语料的训练效果比较

Data size	Baseline	Phrase pair	our method （SQ+phrase pair）
1%	0.1438	**0.1638**	**0.1706**
5%	0.1958	**0.2123**	**0.2146**
10%	0.2123	**0.2283**	**0.2299**
20%	0.2277	**0.2404**	**0.2412**

7. 讨论与未来的工作

统计机器翻译所用的双语平行语料不同于单语语料，其句对中源语句和目标语句有着紧密的

关系。比较几种不同的覆盖度指标的实验表明在用双语特征（短语翻译对）作为覆盖度指标时选取训练语料子集效果最好，20%的数据规模即达到接近 Baseline 用全部数据时的训练效果。而同时考虑 Unigram Bigram 和 Trigram 在 40%左右达到相当的性能。在实验基础上，本文认为在选取 SMT 双语训练语料时采用双语的特征（比如短语翻译对）来衡量覆盖度这一指标更合理。

同时，句对的质量好坏也是影响训练效果的因素，为评价句对的质量本文考虑多种特征提出一种线性模型，这些特征包括：基于双语词典的句对翻译质量，语言模型，翻译模型概率等。将句对质量评价引入到训练语料子集的选取框架中，发现在选取的语料规模较小的时候有微弱提升。虽然效果不够明显，但这也间接说明句对质量评价起到了一定作用。分析本文实验中单句质量评价对选取的训练子集质量影响微弱的原因，可能是因为候选语料本身规模就比较小，低质量句对的比例也较低。究竟单句对的质量对选取高质量的 SMT 训练语料的影响有多大本文还不能给出定论。

总之，本文提出了一种综合考虑句对质量和覆盖度选取统计机器翻译训练语料的方法，利用该方法从大规模平行语料中选取高质量的小规模的子集作为训练语料以在不明显损失机器翻译性能的前提下降低训练和解码的代价。从 70 万句对中选取其中 20%的语料即达到了与用整个语料相当的机器翻译性能，通过实验验证了本文方法的有效性。

本文当前的实验中句对质量评价线性模型中各个特征的权重是还只是由人工在较小训练集上调整，给出的一个经验值，后面的工作中我们将考虑采用自动的方式来训练得到各特征的权重。

下一步，我们还将进一步完善本文的训练语料选取框架，并在千万级规模的平行语料上进行相关实验以进一步验证句对质量评价方法在过滤噪声数据方面的功能是否显著。

参 考 文 献

[1] Yajuan Lü, Jin Huang and Qun Liu. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007.343~350.

[2] Matthias Eck, Stephan Vogel, Alex Waibel 2005. Low cost portability for statistical machine translation based on n-gram coverage. MT Summit X: 227~234.

[3] Franz Josef Och Hermann Ney. The Alignment Template Approach to Statistical Machine Translation. 2004 Association for Computational Linguistics.

[4] 陈毅东，史晓东，周昌乐. 平行语料处理初探：一种排序模型. 中文信息学报, 2006 增刊：66~70.

[5] 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. 中文信息学报, 2008, Vol. 22, No. 2, 40~46.

[6] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proc. of HLT-NAACL, pages 127~133, May.

[7] Tong Xiao, Rushan Chen, Tianning Li, Muhua Zhu, Jingbo Zhu, Huizhen Wang and Feiliang Ren. 2009. NEUTrans: a Phrase-Based SMT System for CWMT2009. In Proc. of 5th China workshop on Machine Translation (CWMT), Nanjing, China, pages: 40-46.

[8] Deyi Xiong, Qun Liu and Shouxun Lin. 2006. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In Proc. of ACL 2006, Sydney, pages: 521-528.