第47卷第4期 中国科学技术大学学报 Vol.47,No.4
2017年4月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Apr. 2017

文章编号:0253-2778(2017)4-0283-07

基于双语句对覆盖度的维汉机器翻译语料选取技术

朱少林[1,2,3],杨雅婷[1,2],米成刚[1,2],李 晓[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 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 830011;
3. 中国科学院大学,北京 100049)

摘要:在进行语料的选取时,语料中的冗余信息包括词汇和句子层面的冗余.目前的方法主要集中在词汇层次的语料覆盖度进行选取,这种方法可以有效地降低词或者短语的信息冗余,但是没有考虑句子层次的覆盖度.为了从大规模的双语语料中选取较小规模的训练语料,得到与大规模训练相同甚至更优的翻译系统,基于双语句对覆盖度进行平行语料的选取,提出一种将unseen n-grams和编辑距离相结合进行语料的选取的方法.实验结果表明,该方法可以在使用较少训练语料的情况下,得到与原始训练翻译效果相同的翻译系统.

关键词:统计机器翻译;双语句对;语料选取

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2017.04.001

引用格式:朱少林,杨雅婷,米成刚,等.基于双语句对覆盖度的维汉机器翻译语料选取技术[J].中国科学技术大学学报,2017,47(4):283-289.

ZHU Shaolin, YANG Yating, MI Chenggang, et al. Corpus selection for Uyghur-Chinese machine translation based on bilingual sentence coverage[J]. Journal of University of Science and Technology of China, 2017,47(4):283-289.

Corpus selection for Uyghur-Chinese machine translation based on bilingual sentence coverage

ZHU Shaolin[1,2,3], YANG Yating[1,2], MI Chenggang[1,2], LI Xiao[1,2], WANG Lei[1,2]

(1.*The Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi* 830011, *China*;
2.*Key Laboratory of Speech Language Information Processing of Xinjiang, Urumqi* 830011, *China*;
3.*University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: When making the selection of corpora, information includes not only redundancy at the vocabulary level but also redundancy at the sentential level. Present methods for this purpose are mainly focused on selecting corpora at the vocabulary level of coverage. These methods can effectively reduce the redundancy of words and phrases, but does not take into account the level of sentence coverage. Aiming at selecting a smaller training corpus from large-scale bilingual corpus, in order to get a the same or better translation system than the mass training data, the corpus from sentence coverage was mainly selected, by combining unseen n-grams method and edit distance. The experimental results show that the proposed method uses less training corpus, but still achieves almost equivalent performance compared with the

收稿日期:2016-03-01;**修回日期**:2016-09-17

基金项目:国家自然科学基金(61473001,71071045,71131002)资助.

作者简介:朱少林,男,1989年生,博士生.研究方向:机器学习、信息处理. E-mail: zhushaolin003@163.com

通讯作者:杨雅婷,博士/副研究员. E-mail: yangyt@ms.xjb.ac.cn

original training corpus.

Key words: statistical machine translation; sentence pairs; corpus selection

0 引言

平行双语语料被认为是统计机器翻译的必不可少的资源和前提条件,同时大多数研究者都认为更多的双语训练语料,可以获得更好的翻译结果[1-2],但是近年,越来越多的研究发现,当训练语料的数量达到一定的程度,增加语料对机器翻译的翻译效果提高愈加不明显,同时随着训练语料规模的不断扩大,将导致翻译系统占用越来越多的计算资源,在一定程度上反而影响翻译效果[2-5].

由于统计机器翻译,是通过训练双语语料得到系统中的各个参数,在训练过程中,翻译系统不会对语料中句子本身的好坏进行筛选,随着训练语料的不断增加,语料的噪音也逐渐放大[3-4,6],对翻译效果的负面影响愈加明显,因此有必要对训练语料进行筛选,在不降低翻译性能的前提下,使用更少的高质量语料[1-2,6].

机器翻译是从一种语言到另一种语言的翻译,由于语种的不同,在进行语料选取时,语言本身的特性对语料的选取也会产生重要的影响[5-6],首先,维汉机器翻译实质是从黏着语到孤立语之间的翻译,维语是一种黏着性语言,词汇的形态非常丰富,一个维语词汇可以由一个词干接多个词缀表达丰富的语义,汉语中的很多实词在维语中可以通过加词缀来表示[7-8],一对多现象普遍,降低了句对词对齐的准确率,严重影响语料选取;其次,在进行实际的翻译系统训练时,所使用的语料已经是经过人工校对的语料,词与词之间互译的准确性已经达到了翻译标准,但是由于人的记忆力有限,很难剔除具有冗余词汇和句式的大规模双语语料.本文提出采用双语句对的覆盖度方法进行语料选取,避免采用词对覆盖度时由于维语词干词缀特性造成的对齐错误的影响,通过句对中未出现的 n-grams(unseen n-grams)的比例和句对相似度技术进行语料的选取.实验结果表明,该方法可在不降低机器翻译质量的前提下,使用更少的语料,甚至在某些情况下,能提高机器翻译的质量.

1 相关背景

1.1 语料选取相关工作

目前已经有多种关于平行双语语料选取的技术,Eck等提出一种基于 n-gram 覆盖度的方法进行平行语料的选取[3],通过计算句子中未出现的 n-gram的比例对句子覆盖度进行评价,最终出具有较高覆盖度的语料,用相对较小规模的训练语料得到与大规模训练语料近乎相同的翻译效果.

一些研究人员试图通过选取与测试集领域相关的训练语料来提高机器翻译的质量,Hildebrand等采用信息检索中向量空间模型(VSM)方法选取与测试集相似的语料作为训练语料,以适应翻译模型[8-9].吕雅娟等也采用相同的方法,不同之处是将相似高的双语句对进行加权重,使相似度高的句子权重更大[10],在进行翻译训练时,相似度高的句对对翻译的贡献更大[11].

此外,一些学者采用基于图论的方法进行高质量双语训练语料的选取,这种方法兼顾词对齐的准确率和覆盖度两个方面,这种方法不依赖待翻译文本,能确保在各个领域都有一个较好的翻译效果.Chao等按照双语语料中短语对的相似度构建图,再从句对信息量和覆盖度两个方面进行语料的选取[1,12-13].Cui等用基于图的方法构建句对和短语对,在进行语料的选取时采用对语料加权和对短语对打分的方法进行语料选取[2].还有一些学者通过构建分类器及多模型融合的方法进行语料选取[14-15].

1.2 维汉语料选取的特点

目前关于统计机器翻译的双语语料的选取主要集中在英汉、英法、英德、法汉等主流语言,这些都属于屈折语到孤立语,或者屈折语到屈折语,但是维汉双语是从屈折语到孤立语,维语较之英语等语种具有更加丰富的词语形态,在英语等屈折语中,词语的词缀多是表示单复数、时态等,这些词缀一般没有汉语中的实词与之对应,并且词缀较少,有着极强的规则.此外,在进行这种语言的统计机器翻译时,都是以词为基本单元进行翻译系统的训练,在进行这种语言间的语料选取时,现有的技术也是基于词进行语料的选取,但是维语是由一个词干缀着多个词缀来表达丰富的语义,一个维语单词往往能表达汉语中一个短语甚至一个短句的意义,同时汉语中很多动词或名词在维语中是用词缀的形式来表达,如表1所示.

表 1　维语词干缀接多个词缀表达一个汉语句子

Tab.1　Stems compose by multiple affix express a Chinese sentence

维吾尔语词	词干词缀切分	增加词缀	汉语意义
ئۆلچەم	ئۆلچەم	词干	标准
ئۆلچەملەش	لەش+ ئۆلچەم	لەش+（化）	标准化
ئۆلچەملەشتۇر	تۇر+لەش+ئۆلچەم	تۇر +（使）	使标准化
ئۆلچەملەشتۇرەل	ەل+تۇر+لەش+ئۆلچەم	ەل+（能）	能使标准化
ئۆلچەملەشتۇرەلمەم	مەم+ەل+تۇر+لەش+ئۆلچەم	مەم+（否定词缀）	不能使标准化
ئۆلچەملەشتۇرەلمەمسى	سى+ مەم+ەل+تۇر+لەش+ئۆلچەم	سى+（疑问词缀）	不能使标准化吗
ئۆلچەملەشتۇرەلمەمسىلەر	لەر+سى+ مەم+ەل+تۇر+لەش+ئۆلچەم	لەر+（词缀你们）	你们不能使标准化吗

维汉机器翻译是一种从黏着语到孤立语的翻译，维语是属于黏着语与汉语之间存在极大的差异，作为一种黏着语，其具有丰富的词汇形态，在维汉机器翻译中往往面临一个词不同形态导致的数据稀疏问题，因此在进行维汉机器翻译时，研究者往往会根据维语形态学的特点，先对维语进行预处理，再利用统计学的方法，训练翻译系统，主要包括多粒度融合进行翻译系统的训练[5]、根据维汉形态学特点进行语序调整的翻译系统的训练[13]以及基于词干词缀粒度进行翻译系统训练等[7]，可以看到对于维汉机器翻译，参数方法都会对维语词进行词干词缀的预处理，采用多粒度的方法进行翻译系统的各种模型训练，因此在为维汉机器翻译选取训练语料时，不能仅仅采用现有的基于词的方法，应该根据维语形态学的特点，采用一些特殊的改进处理来进行语料的选取.

维语词语通过词干缀接不同的词缀形成丰富的词、短语、甚至句子，而汉语的词汇都是相互独立的，词与词之间不存在缀接的关系，并且汉语词的切分都是根据汉语语义进行的，维语词是用空格进行区分，但不会进行词干词缀的分离，这样在双语语料选取的词对齐互译时，很容易出现多对一的现象，在训练机器翻译系统的词对齐步骤中，会造成严重的对齐错误，图 1 说明了维汉语料不进行词干词缀分离造成对齐混乱.

图 1　维语和汉语词对齐互译

Fig.1　Uyghur and Chinese vocabulary aligned

如果对维语词汇进行词干词缀切分，则可以提高对齐的准确率，但是对维语先切分再进行语料选取时，会丢失一些重要的语言知识，在机器翻译系统训练时，由于没有学习到这些重要的语言知识，导致翻译效果不理想，因此本文在进行语料选取时，采用双语句子覆盖度的方法进行语料的选取，尽量避免对维语进行词干词缀的切分，同时也避免词对齐造成的不合理的词语对.

2　基于双语句对覆盖度的语料选取

对于实际用于统计机器翻译的平行语料，语料的覆盖度要保证源语言和目标语言都有较高的覆盖度，本文的任务是在平行语料的前提下，尽可能用较少的语料覆盖较多的词汇信息和句式、句法等语言现象.本文在具体的语料选取时综合考虑了词汇和句子风格两个语言方面，采用句子 unseen n-grams 和句子相似度相结合的方法进行语料的选取.

2.1　基于 unseen n-grams 的词汇覆盖度语料选取

引入 unseen n-grams 是为了保证选取的语料具有较高的短语对覆盖度，使选取出来的语料覆盖较多的词汇和词语搭配，在进行覆盖度语料选取时考虑这一因素的原因是目前所流行统计机器翻译方法都考虑了词或者短语，如基于短语的统计机器翻译、基于层次短语的统计机器翻译等.

对于一个句子，其词汇覆盖度可以用句子中 unseen n-grams 的权重 W_j 来表示，它的计算方法为

$$W_j = 1 - \frac{\sum_{i=1}^{n} \#(n\text{-gram})}{\text{Length}_j} \tag{1}$$

式中，Length_j 表示语料中第 j 个句子中含有 n-gram 短语的数量，$\sum_{i=1}^{n} \#(n\text{-gram})$ 表示句子中出现在已选取语料中的 n-gram 数量，这里 n 取 1、2、3 表示采用 1 元 gram，2 元 gram，和 3 元 gram 时，句子不同覆盖度的权重.考虑句子含有总的 n-gram 数

的原因是避免翻译代价的影响，句子越长就会含有比短句子有更多的 n-gram 数量，长句子比短句子更加不易翻译.这种方法仅仅适用于单一语种，不适用于双语语种语料的处理.在统计机器翻译的训练中，双语语料是一个整体，不能仅仅考虑一端语种的覆盖度，一种源语言可以有不同的翻译，图 2 可以说明这一问题.

ئۇ ئۆيدە ئولتۇرۇشتىن بەكرەك خىزمەت قىلىشنى خالايدۇ.

他宁愿工作，也不愿在家休息

ئۇ خىزمەت قىلىشقا رازىكى ھەرگىز ئۆيدە ئارام ئالمايدۇ.

他宁愿工作，也不愿在家休息

图 2 不同的维语形式表达同一汉语

Fig.2 Different forms to express the same meaning

由图 2 可以看出，如果在进行语料的选取时，仅仅从一端语种考虑，就会将其中的一种语言对删除，丢失很多的语言现象，在统计机器翻译的训练中，导致多种语言现象不能获取，产生严重的数据稀疏.为了克服这一缺点，本文采用采用线性插值的方法进行双语语料的选取.

$$W_j = \alpha W_{\text{target}} + (1-\alpha) W_{\text{sourse}} \tag{2}$$

式中，W_j 上表示句对 j 的覆盖度，W_{target} 表示双语句对中目标语言的覆盖度，W_{soruse} 表示源语言的覆盖度，α 是线性插值的系数，本文取 $\alpha = 0.5$ 进行计算.

用 unseen n-grams 进行语料选取的步骤如下：

第 1 步 设置两个集合 S_1、S_2，S_1 是用于存贮 n 元短语，S_2 表示满足要求的句对.

第 2 步 扫描整个语料，按照公式(1)计算句子 j 的覆盖度权重，同时扫描 S_1.如果句子 j 中的 n 元短语维护出现在 S_1 中，则将其加入 S_1.

第 3 步 重复循环步骤 2，直到所有的句子都被扫描一遍.

第 4 步 输出覆盖度大于阈值的句对.

2.2 基于编辑距离的覆盖度选取技术

编辑距离首先被用在单词的拼写校对，它是计算两个字符串的相似度，在机器翻译的语料选取中，使用这种方法，可以计算两个句子的相似程度，句式和搭配相近的句子，通常具有较高的相似度.

编辑距离定义为对于两个字符串 S_1 和 S_2，将 S_1 替换为 S_2 的最小编辑距离.通常，这样的操作包括：①将一个字符插入字符串；②从字符串中删除一个字符；③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，也就说明两者更相似.计算两个句子相似度的方法是将得分 FMS 定义如下：

$$\text{FMS} = 1 - \frac{\text{LED}_{\text{word}}(S_G, S_R)}{\text{Max}(|S_G|, |S_R|)} \tag{3}$$

式中，LED_{word} 表示计算 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，也即单词的个数.

为了使用编辑距离进行双语语料的选取，本文使用线性插值的方法来计算源语言和目标语言端的语料覆盖度，计算如下：

$$\text{FMS}_j = \alpha\ \text{FMS}_{\text{target}} + (1-\alpha)\ \text{FMS}_{\text{sourse}} \tag{4}$$

式中，FMS_j 是句对 j 的覆盖度，$\text{FMS}_{\text{target}}$ 是句对 j 的目标端句子覆盖度，$\text{FMS}_{\text{sourse}}$ 是句对 j 源语言端的覆盖度，α 是线性插值的系数，这里取值为 0.5.

采用编辑距离进行语料选取的步骤如下：

第 1 步 设定两个集合 S_1 和 S_2，S_1 存放覆盖度大于阈值的句对集合，S_2 是原始语料中未经过扫描的句对集合.

第 2 步 在计算时，首先将原始语料第一个句对加入 S_1.将剩余的句对都存放于 S_2

第 3 步 从 S_2 中取出一个句对 β，同时在 S_2 中将改句对删除，将 β 同集合 S_1 中的句对利用公式(3)和(4)计算器覆盖度，若其覆盖度大于设定的阈值，则将其加入集合 S_1，否则丢弃.

第 4 步 重复循环步骤 3，直到集合 S_2 为空.

第 5 步 输出集合 S_1.

2.3 结合 unseen n-grams 和编辑距离的双语句对覆盖度语料选取

采用句对中 unseen n-grams 短语的数量可以保证在较高的覆盖度下，具有较低的词汇数据稀疏，但是仅仅使用这种方法不能确保在这种较高覆盖度下，语料的句式等语言现象也有较高的覆盖度，因此本文结合了编辑距离，综合进行语料的选取，综合两种方法进行语料选取的伪代码如下：

第 1 步 设置集合 S_1、S_2 和 S_3，S_1 是具有较高覆盖度的语料集合，S_2 是原始语料，S_3 是 n 元短语集合.

第 2 步 扫描集合 S_2，将集合 S_2 中的第一个句子分解为若干个 n 元短语，扫描集合 S_3，若 n 元短语在集合 S_3 中没有出现，将其标记后加入到集合 S_3.

第 3 步 根据句子中未出现的 n 元短语的数量，按照公式(1)和(2)计算其覆盖度，若覆盖度大于阈值则将其加入集合 S_1，同时在 S_2 中删除.

第 4 步　重复训练步骤 2 和 3，直到集合 S_2 扫描完成.

第 5 步　重新扫描集合 S_2，从 S_2 中取出一个句子 β，采用公式(3)和(4)计算其在集合 S_1 上的覆盖度.若所得覆盖度大于预先设定的阈值，则将其加入集合 S_1，否则丢弃，同时在 S_2 中删除.

第 6 步　重复循环步骤 5，直到集合 S_2 为空，将集合 S_1 输出.

这里先用 unseen n-grams 进行覆盖度的初步选取，然后再使用编辑距离的方法进行双语句对覆盖度的选取.难点是进行计算时，两种方法的阈值的选取，特别是在步骤 3 中阈值大小的选取以及 n-grams 中 n 的选取.对于阈值大小的选取，使用 EM 算法进行设定，n 在试验中取 1、2 和 3 分别进行计算.采用 EM 算法进行阈值估算的步骤如下：

第 1 步　初始化阈值 $\alpha=0.0$.

第 2 步　按照本节前面介绍的步骤进行语料的选取，并将选取的语料进行翻译系统的训练，使用测试集来计算 BLEU 值 β_1.

第 3 步　将 α 更新为 $\alpha=\alpha+0.01$.

第 4 步　重复步骤 2，得到 BLEU 为 β_2.直到 $\beta_2-\beta_1<0$，得到阈值 α.

3　实验与分析

实验分别测试采用 unseen n-grams 方法和编辑距离方法进行语料选取的结果，然后测试综合两种方法进行语料选取的结果.

3.1　实验设置

实验中，本文使用的语料是 CWMT2013 机器翻译评测所使用的 11 万的维汉训练语料，机器翻译系统是现今广泛使用的 Moses 机器翻译系统，使用 BLEU 进行翻译性能的评测，测试集使用的是 CWMT2013 提供的标准测试集.实验中使用的数据如表 2 所示.

表 2　实验所使用数据集

Tab.2　Size of data sets

数据集	维语	汉语
原始训练语料	110 000 句	110 000 句
调参集	1 000 句	1 000 句
测试集	1 500 句	1 500 句

3.2　实验结果

为了分析本文方法的有效性，本文进行了多组实验对比，本文设置的基准系统是直接训练 11 万句的维汉平行语料，用上节给出的测试集进行测试，其 BLEU 为 33.7.

首先是使用 unseen n-grams 进行双语语料的选取，Eck 等的工作已经证明使用 unseen 3-grams 更加有利于语料的选取[3]，但是他的工作主要集中在选取源语言端的语料，没有对平行语料进行深入的分析，本文分别从源语言端、目标语言端及双语语料两端综合使用 unseen 3-grams 进行双语语料的选取，同时使用随机选取的方法作为基本的参照，本文所使用的随机选取是采用顺序选取的方法进行.

整个实验中，分别选取不同比例的双语语料进行翻译系统的训练，通过测试训练的翻译系统的 BLEU(%)来评测选取语料的优劣.实验的结果如表 3 所示.

表 3　不同 unseen 3-grams 的测试结果

Tab.3　Result for different unseen 3-grams

数据	Og	Sg	Tg	Bg
1	8.06	9.55	8.20	9.05
2	12.57	13.57	12.17	14.04
3	15.29	17.12	15.35	17.59
4	19.99	19.45	18.97	21.11
5	22.78	22.31	22.02	23.61
6	25.23	24.60	24.27	25.61
7	27.67	27.14	27.40	28.68
8	29.43	29.33	29.07	30.16
9	30.32	31.05	31.48	31.95
10	32.53	32.59	32.78	33.13

为了便于在表中进行表示，本文将使用源语言端的 unseen 3-grams 标记为 Sg，将目标端的 unseen 3-grams 标记为 Tg，双语端的 unseen 3-grams 标记为 Bg，顺序选取标记为 Og.需要说明的是，实验选取语料的大小单位为万句对.

首先，通过实验可以看出，增加语料可以提高机器翻译的翻译质量，说明统计机器翻译增加训练语料是必要的.随着语料的不断增加，其对机器翻译的质量提升逐渐减弱，如语料从 1 万句对增加到 2 万句对时，是 BLEU 值得提升最高可达 4.99，但是语料从 9 万句对提升到 10 万，对 BLEU 的提升最多只有 2.21，说明当训练机器翻译的语料规模巨大时，不能简单地对语料进行叠加，需要对训练语料进行处理，在大规模降低语料的前提下，得到近乎等同于大规模语料的翻译质量.

其次，虽然增加语料可以提高机器翻译质量，但

是从表3可以看出,从不同端进行语料的选取对机器翻译的效果也各不相同,通过双语端进行语料选取效果好于从单一语种端进行语料选取.

分析表3中随着语料的不断增加,对翻译性能的提升愈加不明显的原因,会发现当选取出来的语料较大时,语料中会存在大量句式及用词风格相近的句子,这种冗余的信息对翻译系统能提升贡献较小.下面通过实验验证编辑距离计算相似度的方法进行覆盖度语料选取.表4是使用编辑距离(ED)方法和unseen n-grams方法以及随机选取语料的实验结果.

表4 对比编辑距离的测试结果

Tab.4 Test results for contrasting edit distance

数据	Og	Bg	ED
1	8.06	9.05	7.46
2	12.57	14.04	11.97
3	15.29	17.59	16.36
4	19.99	21.11	22.09
5	22.78	23.61	25.43
6	25.23	25.61	27.63
7	27.67	28.68	31.61
8	29.43	30.16	32.83
9	30.32	31.95	33.52
10	32.53	33.13	33.53

从表4可以看出,不论采取何种方法,增加训练语料都能提高机器翻译的质量,但是不同的方法对于机器翻译的效果提升明显不同.与随机选取和使用unseen n-grams方法相比,采用编辑距离的方法进行语料的选取,对机器翻译的翻译质量提升更加明显.原因是机器翻译不是简单的词到词的互译,更多地体现在句子的整体,词与词之间的搭配以及句子的主谓宾结构等都会影响翻译效果,在进行语料的选取时要从词语之间的相互影响出发进行语料的选取.

第三,使用编辑距离方法时,由表5中的数据可以看出,在选取语料非常小时(本文实验是在不超过20%时),翻译性能严重不足,甚至不如随机选取语料的方法,同时较之unseen n-grams方法更加明显.观察训练的机器翻译系统翻译的测试结果会发现,此时大部分的单词不能翻译出来,这是由于编辑距离选取语料时没有考虑词汇覆盖度,使得词汇的覆盖度严重不足,造成数据稀疏,使得翻译性能不尽如人意,说明在选取的语料规模极少时,词汇的数据稀疏是选取语料的关键因素之一,同时句式、修辞等因素的也是语料选取的重要因素,因此,本文将两种方法进行结合,进行双语句对的语料选取,为了便于在表格中表示,将其简写为Hybrid,实验的结果如表5所示.

表5 对比Hybrid的测试结果

Tab.5 Test results for contrasting Hybrid

数据	Og	Bg	ED	Hybrid
1	8.06	9.05	7.46	9.05
2	12.57	14.04	11.97	14.04
3	15.29	17.59	16.36	18.59
4	19.99	21.11	22.09	25.09
5	22.78	23.61	25.43	26.43
6	25.23	25.61	27.63	29.63
7	27.67	28.68	31.61	31.82
8	29.43	30.16	32.83	33.01
9	30.32	31.95	33.52	33.62
10	32.53	33.13	33.53	33.73

从表5可以看出,将两种方法结合进行语料的选取,在选取的语料较少时也不至于效果太差,同时采用这种方法,在进行语料的选取时,明显可以提高选取语料的效率,仅采用编辑距离的方法进行集合的扩充时,需要对整个语料进行扫描,但是采用Hybrid方法,在集合进行扩充时,已经选取出来的初始集合,不用再对整个语料进行扫描,从而显著地提高了选取效率.同时在选取较少的语料时,能更快地达到收敛,即使用更少的数据,也能达到与原始语料性能相近的翻译系统.特别是本文中结合unseen n-grams和编辑距离的方法,在语料规模是原始语料的75%时,就能达到接近于原始语料的训练的翻译系统性能,这表明该方法能够最大限度地除去语料中的冗余信息.

4 结论

本文提出使用unseen n-grams和编辑距离相结合的方法进行语料的选取,可以在使用较少的训练语料前提下,得到近乎等同于原始训练的翻译效果.本文所使用的实验数据规模相对较少,下一步首先会考虑使用更大规模的训练语料对所提出的方法进行验证,进一步验证该方法对于英汉、英法等机器翻译的训练语料选取的可行性.

参考文献(References)

[1] CHAO W H, LI Z J. A Graph-based bilingual corpus selection approach for SMT[C]// Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation. Singapore: Waseda University Press, 2011: 120-129.

[2] CUI L, ZHANG D D, LIU S J, et al. Collective corpus weighting and phrase scoring for SMT using graph-based random walk [C]// The 2nd Conference on Natural Language Processing & Chinese Computing. Chongqing, China, 2013: 176-187.

[3] ECK M, VOGEL S, WAIBEL A. Low cost portability for statistical machine translation based on n-gram coverage [C]// International Workshop on Spoken Language Translation. Pittsburgh, USA: IWSLT Press, 2005: 61-67.

[4] MANDAL A, VERGYRI D, WANG W, et al. Efficient data selection for machine translation[C]// Spoken Language Technology Workshop. Goa, India: IEEE Press, 2008: 261-264.

[5] SKADIA I, BRĀLĪTIS E. English-Latvian SMT: knowledge or data? [C]// Proceedings of the 17th NODALIDA Conference Processing, http://beta.visl.sdu.dk/~eckhard/nodalida/paper_57.pdf, 2009: 242-245.

[6] HAN X W, LI H Z, ZHAO T J. Train the machine with what it can learn: Corpus selection for SMT[C]// Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: From Parallel to Non-Parallel Corpora. Suntec, Singapore: ACM Press, 2009: 27-33.

[7] 王志洋,吕雅娟,刘群. 面向形态丰富语言的多粒度翻译融合[J]. 中文信息学报. 2011, 25(4): 75-81.
WANG Z Y, LV Y J, LIU Q. System combination with multiple granularities for morphologically rich language translation[J]. Journal of Chinese Information Processing, 2011, 25(4): 75-81.

[8] 米莉万·雪合来提, 刘凯,吐尔根·依布拉音. 基于维语尔语词干词缀粒度的汉维机器翻译[J]. 中文信息学报, 2015, 29(3): 201-206.
MILIWAN·XUEHELAITI, LIU KAI, TURGUN·IBRAHIM. Chinese-Uyghur machine translation based on smallest translation units of stem and suffixes[J]. Journal of Chinese Information Processing, 2015, 29(3):201-206.

[9] HAN J W, JI H, SUN Y Z. Successful data mining methods for NLP[C]// Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing. Beijing, China: ACL Press, 2015: 1-4.

[10] LIU L, HONG Y, LIU H, et al. Effective selection of translation model training data[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA: IEEE Press, 2014: 569-573.

[11] HILDEBRAND A S, ECK M, VOGEL S, et al. Adaptation of the translation model for statistical machine translation based on information retrieval [C]// Proceedings of the 10th Annual Conference on European Association for Machine Translation. San Diego, USA: ACM Press, 2005: 133-142.

[12] 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, 2008, 22(2): 40-46.
HUANG Jin, LV Yajun, lIU Qun. The statistical translation system based on information retrieval method selection and optimization of training data[J]. Journal of Chinese Information Processing, 2008, 22(2): 40-46.

[13] 姚树杰, 肖桐, 朱靖波. 基于句对质量和覆盖度的统计机器翻译训练语料选取[J]. 中文信息学报, 2011, 25(1): 72-77.
YAO Shujie, XIAO Tong, ZHU Jingbo. Selection of SMT training data based on sentence pair quality and coverage [J]. Journal of Chinese Information Processing, 2011, 25(1): 72-77.

[14] 王星, 涂兆鹏, 谢军, 等. 一种基于分类的平行语料选取方法[J]. 中文信息学报, 2013, 27(6): 144-150.
WANG Xing, TU Zhaopeng, XIE Jun, etal. Selection of parallel corpus based on classification[J]. Journal of Chinese Information Processing, 2013, 27(6): 144-150.

[15] KIRCHHOFF K, BILMES J. Submodularity for data selection in statistical machine translation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: ACL Press, 2014: 131-141.