稀缺资源机器翻译中改进的语料级和短语级中间语言方法研究

李 强 王 强 肖 桐 朱靖波
(东北大学自然语言处理实验室 沈阳 110819)

摘 要 该文以英语作为中间语言的方式对在没有直接的外国语至汉语平行训练数据条件下构建统计机器翻译系统的问题进行研究.文中将基于中间语言的机器翻译方法分为系统级、语料级以及短语级中间语3种方法.在文中提出的改进的语料级中间语方法中,通过扩大生成训练数据的规模以及优化词对齐质量的方式来提高翻译系统的翻译性能.在传统的短语级中间语方法中,由于存在无法进行融合的中间语短语从而导致很多高质量短语对无法生成的问题,该文提出的改进方法通过解码生成的方式来扩大短语翻译表,继而提高翻译质量.该文系统地比较了3种中间语方法的优缺点,通过人工分析发现,任何一种方法无法在所有的翻译任务上取得最佳的翻译性能,故文中提出了语料级-短语级融合的中间语方法,该方法在所有翻译任务上取得了最优的翻译性能.最终,文中成功构建了孟加拉语、泰米尔语、乌兹别克语、匈牙利语至汉语的机器翻译系统.与基线系统相比,文中提出的方法在4种外国语的测试集上获得了0.8至2.8个BLEU点的上涨.

关键词 自然语言处理;统计机器翻译;外国语翻译;中间语言;语料构建
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2017.00925

Research on Improved Corpus-Level and Phrase-Level Pivot Language Based Methods in Low-Resource Machine Translation

LI Qiang WANG Qiang XIAO Tong ZHU Jing-Bo
(*NLP Lab*, *Northeastern University*, *Shenyang* 110819)

Abstract In this paper, we use English as the pivot language to build statistical machine translation systems as parallel training corpora for foreign languages and Chinese are non-existent. We classify the pivot language based methods into system-level, corpus-level, and phrase-level methods. For the proposed improved corpus-level method, we improve the translation performance through enlarging the size of bilingual training corpora and improving the quality of word alignments. For the typical phrase-level pivot language based method, as many high-quality phrase pairs cannot be generated from source-pivot and pivot-target phrase translation tables, we use decoding-generation method to enlarge the size of phrase pairs in phrase translation table and improve the translation performance. We analyze the strengths and weaknesses for system-level, corpus-level, and phrase-level pivot language based approaches during system construction, and we find that there is no one method can achieve the best translation performance among all the translation tasks through human analysis. Therefore we propose the corpus-phrase combination based pivot method which achieves the highest BLEU scores among all the translation tasks. We translate Bengali, Tamil, Uzbek, and Hungarian into Chinese with our proposed pivot language based methods.

收稿日期:2016-05-07;在线出版日期:2016-09-13.本课题得到中央高校基本科研业务专项资金(N140406003)、国家留学基金、国家自然科学基金(61272376,61300097)资助.**李 强**,男,1988年生,博士研究生,主要研究方向为自然语言处理、机器翻译.E-mail:liqiangneu@gmail.com.**王 强**,男,1990年生,博士研究生,主要研究方向为机器翻译.**肖 桐**,男,1982年生,博士,副教授,主要研究方向为机器翻译.**朱靖波**,1973年生,博士,教授,主要研究领域为自然语言处理、机器翻译.

Finally, we observe significant improvements from 0.8 to 2.8 BLEU points when translating Bengali, Tamil, Uzbek, and Hungarian on the test datasets compared with the baseline translation system.

Keywords natural language processing; statistical machine translation; foreign languages translation; pivot language; corpus construction

1 引 言

近几十年来,随着中国国力的不断增强,中国与世界各国的交流不断加深,外国语至汉语的翻译系统在日常生活中有着非常大的需求.随着基于统计的机器翻译技术在近十多年来取得的进展,基于统计的机器翻译系统的搭建形成一套标准的流程,翻译质量也已达到人们可接受的标准[1-3].基于统计的机器翻译系统是构建在高质量的人工翻译的双语平行训练数据的基础上的.在英汉翻译系统的构建中,平行数据资源比较充分,存在千万行人工翻译的高质量的双语平行数据用于基于统计的英汉翻译系统的搭建.然而,并非所有的语言对间都存在大规模的高质量的平行句对用于机器翻译系统的训练[4-9].例如,很多外国语与汉语间并没有高质量的人工翻译的平行训练数据用于翻译系统的训练,从而对构建外国语至汉语翻译系统造成挑战.

图 1 基于中间语言的机器翻译方法

训练数据的缺失对构建基于统计的机器翻译系统造成很大挑战,如孟加拉语、泰米尔语、乌兹别克语、匈牙利语等一些对国人来说并不常见的语言,这些语言与汉语间的双语平行语料收集工作难度大,平行语料的质量不容易保证,需要相关语言专业人员进行大量的体力劳动来构建双语平行语料.但是,由于英语作为国际通用语言,孟加拉语等外国语与英语间的双语平行数据的收集相对来说更为容易.通过英语作为中间语言,是构建没有直接双语平行数据的外国语至汉语的翻译系统的一个非常重要的途径[4-6,10-12].基于中间语言的方法如图 1 所示,从本质上来说,基于中间语言的方法先将源语言句子或片段翻译至中间语言,继而将中间语言翻译结果翻译至最终的目标语言.Koehn 等人[10]在构建欧盟 22 种官方语言间的 462 套互译系统时,发现在 273 套互译系统上使用英语作为中间语言的翻译性能要明显优于直接使用双语数据构造的系统.所以,本文的主要研究工作是,在没有高质量的人工翻译的外国语至汉语平行数据的基础上,利用英语作为中间语言进行外国语至汉语翻译系统的构建.

外国语与英语的双语平行数据的构建工作目前开展的较为顺利,虽然很多外国语与英语之间的平行数据并未达到英汉双语平行数据的千万级句对规模,但是数据规模足以构建基本的基于统计的外国语至英语的机器翻译系统[4].与此同时,利用现有的大规模的英汉双语数据,以及较为成熟的基于统计的机器翻译技术,可以构建英汉翻译系统.继而可以通过以英语作为中间语言的方式进行外国语至汉语翻译系统的构建工作.基于中间语言的机器翻译方法包括系统级[10]、语料级[6]以及短语级[5]中间语方法,本文对 3 种方法进行了系统的比较与研究.针对语料级和短语级中间语方法在构造翻译系统中存在的翻译质量不佳的问题,本文提出了改进的语料级中间语和改进的短语级中间语的翻译系统构建技术.在改进的语料级中间语方法中,本文通过使用 m 最优翻译结果来扩大训练数据以及优化词对齐的方法来提高语料级方法的翻译性能.在改进的短语级中间语方法中,针对无法进行融合的中间语短语,本文提出了解码生成的方法,来增加短语翻译表中高质量短语对的数量,该方法优化了传统短语级中间语方法中存在的部分中间语短语无法进行融合的问题[5].通过对实验结果的分析,本文发现任何一种中间语方法都无法在所有的翻译任务上都获得最优的翻译性能,故本文提出了语料级-短语级融合的中间语方法,该方法在所有翻译任务上都获得最优翻译结果.此外,本文对外国语汉语翻译系统构建的 3 种方法进行了详细的分析,系统地比较了每一种方法的优缺点,为未来的外国语汉语翻译系统构建工作提供了更多的参考方案.本文通过以英语作为中间语言的方法,成功构建了孟加拉语、泰米尔语、乌兹

别克语、匈牙利语至汉语的统计机器翻译系统.最终,与基线系统相比,本文提出的方法在孟加拉语、泰米尔语、乌兹别克语、匈牙利语这4种语言的测试集上获得了0.8～2.8个BLEU点的提高.

本文第2节介绍和分析关于基于中间语言构建翻译系统的相关研究工作;第3节提出具体翻译任务;第4节对传统的系统级中间语言的方法进行说明;第5节提出改进的语料级中间语言的翻译系统构建方法;第6节提出改进的短语级中间语言翻译系统的构建方法;第7节提出语料级-短语级融合的中间语方法;第8节为相关实验结果;第9节对本文工作进行总结.

2　相关工作

Koehn等人[10]在构造欧盟462种语言对间的翻译系统时,发现在273种语言对上使用以英语作为中间语言的方法比直接使用双语数据构造的翻译系统性能提高2～10个BLEU点,其使用的方法为系统级中间语方法.Wu和Wang[5]提出了短语级中间语方法,即通过中间语言间接生成或优化源语言至目标语言的短语翻译表.在稀缺资源的语言对间通过使用短语级中间语的方法有效地提高了翻译系统的翻译性能.该方法在法语西班牙语翻译系统、法语德语翻译系统以及汉语日语翻译系统上得到验证并获得不错的翻译性能.与此同时,该方法在没有直接双语资源的条件下可以进行源语言目标语言间短语翻译表的生成,但其方法需要使用源语言目标语言间的开发集对翻译系统进行调优.Dabre等人[4]在进行日语到印地语翻译时,通过使用7种额外的语言作为中间语言来加强日语到印地语翻译系统的翻译性能.其方法使用的7种中间语言与日语和印地语训练数据为多语言间的平行数据,其主要的优化过程为通过使用基于中间语言的翻译系统中的短语翻译表来优化直接构造的日语印地语翻译系统中的短语翻译表,本质上是一种短语级中间语方法.Zhu等人[11]通过使用随机漫步(Random Walk)的方法获取潜在的源语言至目标语言的短语路径,从而缓解了部分中间语言短语无法进行融合的问题.而后,Zhu等人[12]在后续工作中提出在融合短语表前直接对源语言-目标语言的短语对共现次数进行估计的方法,避免了在短语推导时由于存在无法进行融合的短语从而导致破坏翻译概率空间的问题.

与上述基于中间语工作不同的是,首先本文定义的问题是:在没有直接的源语言与目标语言间双语平行训练数据、开发集、测试集的条件下进行翻译系统的搭建.其次,与传统的语料级中间语方法相比,本文提出的改进的语料级中间语方法优化了传统方法中的训练数据规模不足以及词对齐质量不高的问题.在传统的短语级中间语方法中短语翻译表的生成过程中,由于存在无法进行融合的中间语短语,导致很大一部分潜在的高质量的外国语汉语短语对并没有生成.针对这一问题,Zhu等人[11]通过使用已存在的源语言-中间语言和中间语言-目标语言的短语翻译表,对无法生成的源语言和目标语言短语对应的可能性进行估计,提出随机漫步的方法来缓解无法融合的问题.本文提出的改进的短语级中间语方法与Zhu等人方法不同的是,本文方法通过解码生成的方式构造一些新的并未出现在原始的中间语言-目标语言短语翻译表中的短语对,继而生成许多不存在于传统方法生成的源语言-目标语言短语翻译表中的高质量的短语对.此外,通过对多种中间语方法结果的分析,本文提出了语料级-短语级融合的中间语方法,该方法在所有的翻译任务上均取得了最佳的翻译性能.Wu等人[6]在其工作中使用了系统融合的方法来选择最佳的翻译结果,与之不同的是,Wu等人的工作使用的是回归学习模型来对最优翻译结果进行选择,而本文的工作则是直接在统计机器翻译的最小错误率训练框架中来进行最优翻译结果的挑选.本文使用的方法与现有的统计机器翻译方法进行了有效的融合,取得不错的效果.最终使用本文提出的方法,成功构建了孟加拉语、泰米尔语、乌兹别克语、匈牙利语至汉语的统计机器翻译系统.

3　外国语至汉语翻译

本文的主要目标是,在没有高质量的人工翻译的外国语至汉语的双语平行数据的情况下,通过基于中间语言的方法,构建外国语至汉语的基于统计的机器翻译系统.在翻译系统的构建过程中,对所有基于中间语的方法进行详细的分析,同时提出相应的改进方法.本文工作主要解决以下几个问题:

(1)训练数据缺失.没有直接用于构建基于统计的外国语至汉语机器翻译系统的双语平行训练数据;

(2)开发集缺失.没有用于参数优化的人工翻

译的外国语至汉语的开发集；

(3)测试集缺失.没有用于机器翻译系统评价的人工翻译的外国语至汉语的测试集；

(4)传统语料级中间语方法中翻译性能不佳的问题；

(5)传统短语级中间语方法中存在无法进行融合的中间语短语问题；

(6)不存在一种方法在所有的翻译任务上均取得最佳翻译性能的问题.

在目前的研究中,已经存在的先决条件是：

(1)外国语英语双语资源.具有直接构建机器翻译系统的外国语至英语的双语平行数据、开发集、测试集,可以构建外国语至英语翻译系统；

(2)英语汉语资源.具有直接构建英汉机器翻译系统的双语平行数据、开发集、测试集；

(3)成熟的基于统计的机器翻译方法.

本文在现有资源和方法的基础上,对外国语汉语机器翻译系统中的基于中间语言构建方法进行研究.由于外国语至汉语双语平行训练数据的缺失对构造基于统计的机器翻译系统造成极大的挑战,所以,本文的最终目的是,以英语作为中间语言,构建孟加拉语、泰米尔语、乌兹别克语、匈牙利语至汉语的机器翻译系统.本文在以英语为中间语言构建外国语至汉语机器翻译系统的过程中,详细地分析了现有的系统级、语料级、短语级这3种基于中间语言方法的优缺点.与此同时,针对其存在的问题,本文提出了改进的语料级、改进的短语级以及语料级-短语级融合的中间语方法,并最终在4种翻译任务上获得了不错的翻译性能.

4 系统级中间语

4.1 系统级中间语方法

本文使用3种以英语作为中间语的方式构建外国语至汉语的机器翻译系统.第1种为常规的系统级中间语方法[10],系统级中间语方法如图2所示,该方法的主要构建过程如下：

(1)使用外国语英语双语平行数据构建基于统计的外国语至英语的翻译系统；

(2)使用英语汉语双语平行数据构建基于统计的英语至汉语的翻译系统；

(3)对于给定的待翻译外国语,首先将待翻译外国语翻译至英语,继而将英语翻译成汉语,最终得到对应待翻译外国语的汉语翻译结果.通过两个翻译过程最终建立外国语至汉语的统计机器翻译系统.

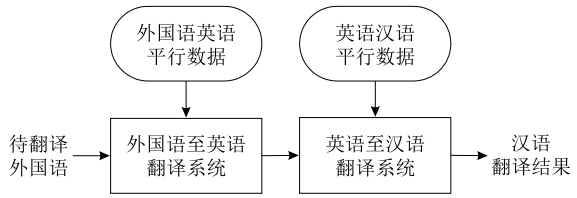


图2 系统级中间语方法构建流程(图中箭头自上至下为翻译系统构建过程,自左至右为翻译过程)

系统级中间语方法为目前最为常用的在没有直接双语平行数据的基础上,构建翻译系统的方法,英语是最为常用的中间语言.

4.2 短语翻译模型

在系统级中间语方案中,外国语英语翻译系统以及英汉翻译系统均使用基于短语的统计机器翻译模型[1],基于短语的翻译模型如图3所示.

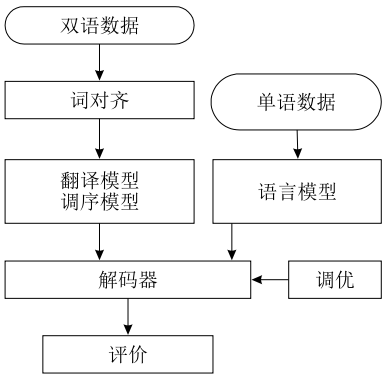


图3 基于短语的统计机器翻译系统构建过程

在搭建一套基于短语的外国语至汉语的统计机器翻译系统的过程中,如孟加拉语至汉语的翻译系统,首先,需要收集一定规模的孟加拉语与汉语的双语平行数据来建立翻译模型与调序模型；其次需要收集汉语的大规模的单语语料来构建高质量的目标语语言模型；此外,需要收集孟加拉语至汉语的人工翻译的高质量的开发集及测试集,在开发集与测试集中,针对每一个源语言句子,至少提供一个人工翻译的汉语结果.开发集用于对机器翻译系统的特征进行参数调优.测试集用于评价当前搭建的机器翻译系统的质量.

所以,基于短语的统计机器翻译模型的主要训练过程如下：

(1)词对齐训练.使用GIZA++等词对齐工具对双语平行数据训练词对齐[13],使用*grow-diag-final-and*方法对双向的词对齐进行对称化处理[14]；

(2)翻译模型训练.使用含有词对齐的双语平行数据训练翻译模型[1];

(3)调序模型训练.使用含有词对齐的双语平行数据训练调序模型[15];

(4)语言模型训练.使用目标语言单语语料,训练 n 元语言模型[16];

(5)参数调优.在解码器中使用翻译模型、调序模型以及语言模型,在开发集上使用最小错误率方式对翻译系统的参数进行参数优化[17];

(6)结果评价.在测试集上对机器翻译系统的质量进行评价[18].

基于短语的统计机器翻译模型为本文实验使用的机器翻译模型.在真实的翻译系统搭建中,可以将基于短语的统计机器翻译模型替换为任何基于统计的机器翻译模型,如基于层次短语[2]和基于句法[19]的翻译模型.

4.3　优点和不足

通过对系统级中间语方法构建过程的分析,该方法的主要优势是:

(1)数据易收集.外国语英语的双语平行数据、开发集、测试集相对来说较容易收集,从而可以方便构建第 1 级的外国语英语翻译系统;英语与汉语间拥有大量的高质量人工翻译的双语平行数据用于搭建第 2 级的英汉翻译系统.

(2)真实的数据上进行训练与评价.外国语英语翻译系统,英语汉语翻译系统都是在真实的开发集上进行优化,在真实的测试集上对翻译系统的质量进行评价.

然而这种方式又有其不可忽视的缺点,主要的不足有:

(1)间接建模.由于该方法分别对外国语英语、英语汉语翻译系统分别建模,翻译系统构建的过程中并没有直接对外国语汉语的翻译系统进行建模,没有办法直接优化外国语至汉语的翻译性能.

(2)错误蔓延.外国语英语翻译系统、英语汉语翻译系统的翻译性能对最终构建的外国语汉语翻译系统有很大影响,通过两次翻译风险增加,容易造成错误蔓延,即第 1 级的翻译错误一定会传递到第 2 级翻译中.

(3)优化困难.外国语英语翻译系统、英语汉语翻译系统中任何一个翻译系统性能的增长不能确保最终外国语汉语翻译性能的增长.较容易收集到的外国语汉语互译词典无法直接加入系统中对系统进行优化.

(4)耗时.每一个待翻译外国语句子需要使用两级的机器翻译系统进行两次解码才可得到最终的汉语翻译结果,而解码是非常耗时的一项工作,故消耗更多的时间.

由于系统级中间语方法存在以上诸多的问题,机器翻译相关研究人员继而提出了语料级中间语方法以及短语级中间语方法.

5　改进的语料级中间语

在这一节中,针对传统语料级中间语方法的不足,本文提出了改进的语料级中间语方法来直接构建外国语至汉语的统计机器翻译系统.通过对语料级中间语方法中不同组件的优化,可以直接优化最终的翻译性能.

5.1　语料级中间语方法

传统的语料级中间语方法如图 4 所示,该方法构建外国语汉语翻译系统的主要过程如下[6]:

(1)使用英汉双语平行数据、开发集、测试集构建英语至汉语的翻译系统;

(2)使用已经构建好的英汉机器翻译系统对外国语英语平行数据、开发集、测试集的英语部分进行翻译,从而间接生成外国语与汉语的双语平行数据、开发集、测试集;

(3)在构建好的外国语汉语训练数据上进行翻译模型、调序模型的训练,使用构建好的开发集进行参数优化,使用构建好的测试集进行评价,从而构建外国语至汉语的基于统计的机器翻译系统.

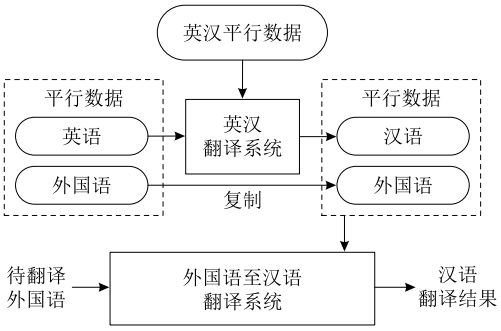


图 4　语料级中间语方法流程(图中箭头自上至下为翻译系统构建过程,箭头自左至右为翻译过程)

在系统级和语料级两种中间语方法中同时用到了英汉翻译系统,在两种方法中,英汉翻译系统的构建方式以及使用数据相同.

在语料级中间语方法中,基于统计的翻译系统中所有的参数的调优直接通过优化外国语至汉语翻

译系统来完成.与系统级中间语方法相比,该方法的优点是直接构建的方式更为灵活,与此同时,很多直接的优化方法,如易收集的外国语汉语词典可以非常容易地加入到此方法中.但是,这种方法的缺点也是显而易见的,本文将在后续章节对该方法的优缺点进行详细论述.

5.2 m 最优翻译结果

在通过英语至汉语的翻译系统构建外国语汉语双语平行数据的过程中,本文采用 m 最优的翻译结果,即针对平行数据中的英文数据,每个句子生成 m 个最优候选翻译,将平行句对里外国语部分复制 m 份,与翻译的汉语构成新的平行数据,具体如图 5 所示.

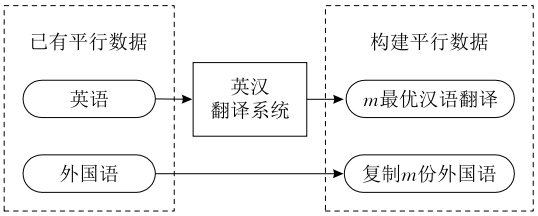


图 5 外国语汉语平行句对生成

在这里,使用的 m 最优结果相当于对机器翻译的结果进行一次平滑处理,即 1 最优结果不能保证所有的翻译结果都是合理的、正确的,但是 m 最优结果可以使翻译的正确性增大.在使用 m 最优翻译结果训练外国语至汉语的词对齐时,词对齐结果将更加准确.在本文的翻译任务中,经过大量实验证实,当 m 取值为 5 时,可以得到最优的结果.

5.3 词对齐优化

在语料级中间语方法中,本节使用一种非常简单的词对齐优化方法直接优化外国语至汉语的翻译系统.通过该优化方法,间接地证明了在系统优化方面,语料级中间语方法与系统级中间语方法相比更加快速、有效.具体的优化方案如图 6 所示,过程如下:

(1)首先通过两种词对齐工具,即 GIZA++和伯克利对齐[20]工具训练两种词对齐;

(2)将两种词对齐结果直接在文件一级进行合并,即得到新的词对齐结果,与此同时,将双语训练数据的外国语与汉语部分分别复制两份,从而与新的词对齐相对应.在合并后的两种词对齐结果上进行翻译模型和调序模型的训练.与文献[21]提出的 3 种合并词对齐的方法不同的是,本文使用的方法为语料级的词对齐合并方法;

(3)最后在新的含有词对齐的双语训练数据上训练翻译模型和调序模型.

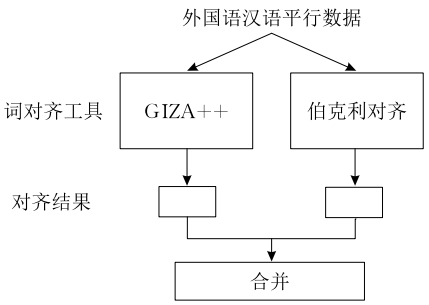


图 6 外国语至汉语翻译系统词对齐优化

在数据规模有限的翻译任务上,该方法是一种简单、有效的优化翻译性能的方法.具体的实验结果见第 8 节的实验部分.本文通过使用此方法证实了语料级中间语方法优化的灵活性与优势,通过对翻译组件的优化可以达到最终优化翻译系统性能的目的.

5.4 优点和不足

语料级中间语方法与系统级中间语方法相比,主要有以下的优缺点.该方案的优点如下:

(1)直接建模.直接对外国语至汉语的机器翻译系统进行建模,可以有针对性地进行翻译系统的调优工作;

(2)易于优化.许多优化方法可直接应用到外国语至汉语的机器翻译系统上,如词对齐优化、外国语汉语词典嵌入,使用 m 最优翻译结果的语料构建方式,翻译性能的增长直接体现在最终的汉语翻译结果上;

(3)稀缺资源生成.可间接构造外国语汉语双语语料,为后续外国语汉语翻译工作打下良好基础;

(4)省时、无错误蔓延.待翻译外国语经过一次解码即可得到汉语翻译结果,比两次解码节省更多时间,同时有效地缓解了两次翻译的错误蔓延问题.

但该方法有不可忽视的缺点,具体如下:

(1)非真实数据建模.外国语汉语双语训练数据、开发集、测试集都是通过英汉翻译系统翻译而来,最终构建的外国语至汉语翻译模型并非构建在真实的、人工翻译的数据上;

(2)对英汉翻译系统要求高.与系统级中间语方法相同,对英汉翻译系统的翻译质量要求较高,即英汉翻译系统翻译的越准确,构造的外国语至汉语双语数据就更为准确.

6 改进的短语级中间语

6.1 短语级中间语方法

在使用基于中间语的方法构建统计机器翻译系

统的过程中,除了系统级中间语和语料级中间语方法外,本文研究了另外一种短语级的中间语方法来构造外国语至汉语的机器翻译系统[5-6]. 文献[5]提出的短语级中间语方法的主要过程如下:

(1) 训练外国语至英语的短语翻译表 $p_{\bar{f}-\bar{e}}$ 以及英语至汉语的短语翻译表 $p_{\bar{e}-\bar{c}}$;

(2) 通过相同的英语短语进行短语表的融合,即通过 $p_{\bar{f}-\bar{e}}$ 与 $p_{\bar{e}-\bar{c}}$ 的相同部分 $\bar{e}$ 进行短语翻译表的融合,最终生成 $p_{\bar{f}-\bar{c}}$.

在短语级中间语方案中最为关键的问题是:如何对推导出的短语翻译规则进行概率估计. 即主要包括双向的短语翻译概率以及双向的词汇化加权的计算.

给定外国语短语 $\bar{f}$,中间语短语 $\bar{e}$,汉语短语 $\bar{c}$,通过使用以下两个公式对双向的短语翻译概率进行建模:

$$\Pr(\bar{c}\mid\bar{f})=\sum_{\bar{e}}\Pr(\bar{c}\mid\bar{e})\Pr(\bar{e}\mid\bar{f}),$$

$$\Pr(\bar{f}\mid\bar{c})=\sum_{\bar{e}}\Pr(\bar{f}\mid\bar{e})\Pr(\bar{e}\mid\bar{c}).$$

在这里:

(1) $\Pr(\bar{c}\mid\bar{f})$ 为外国语汉语短语翻译表中正向的短语翻译概率;

(2) $\Pr(\bar{f}\mid\bar{c})$ 为反向的短语翻译概率.

继而,通过使用以下两个公式对双向的词汇化加权进行建模:

$$p_{\omega}(\bar{c}\mid\bar{f},a)=\prod_{i=1}^{L(\bar{c})}\frac{1}{\{j\mid(i,j)\in a\}}\sum_{\forall(i,j)\in a}\omega(c_i\mid f_j),$$

$$p_{\omega}(\bar{f}\mid\bar{c},a)=\prod_{j=1}^{L(\bar{f})}\frac{1}{\{i\mid(j,i)\in a\}}\sum_{\forall(j,i)\in a}\omega(f_j\mid c_i).$$

在这里:

(1) $p_{\omega}(\bar{c}\mid\bar{f},a)$ 为外国语汉语短语翻译表中正向的词汇化加权;

(2) $p_{\omega}(\bar{f}\mid\bar{c},a)$ 为反向词汇化加权;

(3) $L(\cdot)$ 为短语长度.

其中,词汇化翻译概率 ω 通过如下两个公式进行计算:

$$\omega(c\mid f)=\frac{\mathrm{count}(f,c)}{\sum_{c'}\mathrm{count}(f,c')},$$

$$\omega(f\mid c)=\frac{\mathrm{count}(f,c)}{\sum_{f'}\mathrm{count}(f',c)}.$$

在这里:

$\mathrm{count}(f,c)$ 为外国语词汇 f 与汉语词汇 c 在语料中共现的次数.

在短语级中间语的方法中,使用以下公式对 f 与 c 的共现次数进行建模:

$$\mathrm{count}(f,c)=\sum_{k=1}^{K}\Pr_k(\bar{f}\mid\bar{c})\sum_{i=1}^{n_k}\delta(f,f_i)\delta(c,c_{a_i}).$$

在这里:

(1) $\Pr_k(\bar{f}\mid\bar{c})$ 是短语对 k 的短语翻译概率;

(2) n_k 是短语对 k 中源语言短语 $\bar{f}$ 的长度;

(3) 当 $x=y$ 时,$\delta(x,y)=1$,否则 $\delta(x,y)=0$.

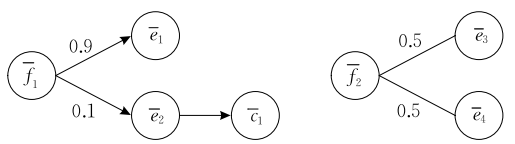
通过以上公式的计算,最终得到外国语汉语完整的短语翻译表. 关于短语级方法的详细描述见文献[5].

6.2　改进的短语级中间语方法

在短语级中间语方法中,虽然该方法能够直接将外国语短语翻译为汉语,避免了系统级中间语方案中多次解码造成的错误蔓延问题,但是该方法也存在一些需要克服的问题,主要如下:

(1) 产生互译性不高的噪声翻译规则. 如图 7(a)所示,外国语短语 $\bar{f}_1$ 翻译为 $\bar{e}_1$ 的概率为 0.9,翻译为 $\bar{e}_2$ 的概率为 0.1. 由于 $\bar{e}_1$ 在英汉短语翻译表 $p_{\bar{e}-\bar{c}}$ 中无法找到对应的翻译结果,只能通过概率不高的 $\bar{e}_2$ 进行推导,继而产生互译性不高的噪声翻译规则.

(2) 源语言短语丢失. 如图 7(b)所示,外国语短语 $\bar{f}_2$ 对应的英语短语 $\bar{e}_3$ 和 $\bar{e}_4$ 在英汉短语翻译表 $p_{\bar{e}-\bar{c}}$ 中均找不到对应的英汉短语对,所以无法生成 $\bar{f}_2$ 对应的汉语翻译结果.



(a) 噪声翻译规则　　(b) 源语言短语丢失

图 7　传统的短语级中间语方法存在的问题

针对这两个问题,本文提出了基于解码-生成的改进后的短语级中间语方法. 改进的短语级中间语方法如图 8 所示,具体的构建过程如下:

(1) 首先,将图 7 中不能进行融合的英语短语 $\bar{e}_1$,$\bar{e}_3$ 和 $\bar{e}_4$ 使用构建好的英汉翻译系统进行翻译,得到其对应的汉语翻译结果. 新生成的英汉短语对本文称之为补充的英汉短语对,将补充的英汉短语对加入到原有的英汉短语翻译表 $p_{\bar{e}-\bar{c}}$ 中,从而生成新的英汉短语翻译表;

(2) 新生成的英汉短语翻译表 $p_{\bar{e}-\bar{c}}$ 与外国语英语短语翻译表 $p_{\bar{f}-\bar{e}}$ 进行融合,最终生成外国语汉语的短语翻译表 $p_{\bar{f}-\bar{c}}$.

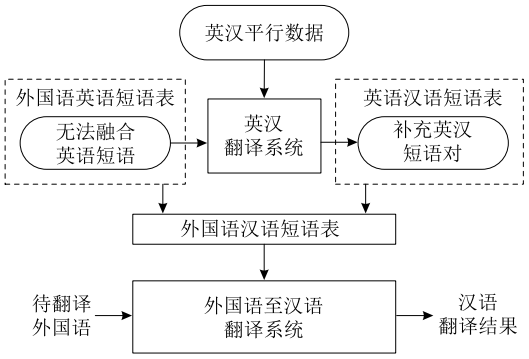


图 8 改进的短语级中间语方法

在此方法中,需要进行系统优化的开发集,进行系统评价的测试集以及调序模型.在这里,使用语料级中间语方法中生成的外国语汉语开发集、测试集、调序模型来进行翻译系统的构建工作.

在新生成的补充的英汉短语对中,双向的短语翻译概率通过如下公式进行计算:

$$\Pr(\bar{c}\mid\bar{e})=\prod_{(\bar{c}',\bar{e}')\in D}\Pr(\bar{c}'\mid\bar{e}'),$$

$$\Pr(\bar{e}\mid\bar{c})=\prod_{(\bar{c}',\bar{e}')\in D}\Pr(\bar{e}'\mid\bar{c}').$$

在这里:

D 为翻译$(\bar{c},\bar{e})$过程中的解码空间,$(\bar{c}',\bar{e}')$为翻译短语对$(\bar{c},\bar{e})$过程中用到的短语对.

双向的词汇化加权计算方法与双向的短语翻译概率相似,通过以下两个公式进行计算:

$$p_\omega(\bar{c}\mid\bar{e},a)=\prod_{(\bar{c}',\bar{e}')}p_\omega(\bar{c}'\mid\bar{e}',a),$$

$$p_\omega(\bar{e}\mid\bar{c},a)=\prod_{(\bar{c}',\bar{e}')}p_\omega(\bar{e}'\mid\bar{c}',a).$$

至此,可以使用本文提出的改进的短语级中间语方法生成外国语至汉语的短语翻译表.在反向的短语翻译表融合过程中,存在相似的问题,但是由于本文翻译任务中使用的英语-外国语翻译系统在小规模平行数据的基础上进行训练,故翻译性能与英汉翻译系统相比有很大差距,所以新构造的源语言端短语在开发集、测试集上解码过程中的匹配度很低,故没有明显地提高翻译系统的翻译性能.与文献[11]提出的方法相比,本文通过解码-生成的方式重新生成了无法融合的中间语-目标语短语,继而进行短语翻译表的融合,文献[11]的方法则是在现有短语翻译表的基础上,通过使用随机漫步的方式进行融合,故这两种方法有着本质的不同.

6.3 优点和不足

改进的短语级中间语方法与前两种方法相比,主要有以下的优缺点.

该方法的优点是:

(1)更多正确的翻译规则.可生成大量的外国语汉语短语翻译对,外国语词汇翻译为正确的汉语词汇的概率增加;

(2)直接建模.直接对外国语汉语翻译系统进行建模,易收集的外国语汉语词典可直接嵌入到此方案中;

(3)解码省时.直接生成外国语汉语短语翻译表,给定待翻译外国语,只需一次解码即可得到汉语翻译结果.

但是,该方法也有一些自身的不足:

(1)数据缺失.在没有开发集与测试集的条件下,无法进行翻译系统的调优和评价工作;

(2)模型缺失.该方法无法同时生成外国语与汉语间的调序模型,造成模块缺失;

(3)噪声规则变多.大量的外国语汉语短语翻译对虽然使得正确翻译的概率增加,同时也引入更多的噪声;

(4)规则翻译耗时.大量的不能融合的英语短语需要使用英汉翻译系统进行解码,消耗更多的时间.

7 语料级-短语级融合的中间语方法

在机器翻译中,机器翻译的目标是,自动将源语言待翻译字符串 s_1^J 转换成目标语言字符串 t_1^I.在统计机器翻译中,该问题被转换为在所有可能的翻译结果 t_1^I 中,找到最优的翻译结果 $\hat{t}_1^I$.具体如以下公式所示:

$$\hat{t}_1^I=\arg\max_{t_1^I}\{\Pr(t_1^I\mid s_1^J)\}.$$

在这里:

$\Pr(t_1^I\mid s_1^J)$为给定 s_1^J 的条件下,翻译结果为 t_1^I 的概率.

为了对概率进行 $\Pr(t_1^I\mid s_1^J)$ 建模,基于统计的机器翻译方法采用对数-线性模型[22],具体如以下公式所示:

$$\hat{t}_1^I=\arg\max_{t_1^I}\Big\{\sum_{n=1}^{N}\lambda_n h_n(t_1^I,s_1^J)\Big\}.$$

在这里:

(1)h_n 为构建机器翻译系统过程中使用的特征,如语言模型,翻译模型,调序模型等;

(2)λ_n 为特征 h_n 对应的特征权重.

基于统计的机器翻译系统是构建在大规模的双

语平行数据的基础上的.在开发集上对系统调优的过程中,使用目前最常用的翻译结果质量评价指标BLEU来定义错误函数[18],使用最小错误率训练(MERT)的方法对参数 λ_n 进行优化[17].

基于以上公式,本文使用的语料级、短语级中间语融合的方法如以下公式所示

$$\hat{t}_1^I = \arg\max_{t_1^I}\{\lambda_{LM}h_{LM} + \lambda_R h_R + \sum_{n=1}^{4}\lambda_n h_n(t_1^I, s_1^J) + \sum_{m=1}^{4}\lambda_m h_m(t_1^I, s_1^J).$$

在这里:

(1) LM 为语言模型;

(2) R 为调序模型;

(3) n 为改进的语料级中间语方法中使用的双向短语翻译概率、双向词汇化翻译概率 4 个特征;

(4) m 为改进的短语级中间语方法中使用的双向短语翻译概率、双向词汇化翻译概率 4 个特征.

通过以上公式,本文实现了语料级-短语级中间语方法的融合,该方法在所有的翻译任务中,都取得了最优的翻译性能.

8　评　价

本文使用 NiuTrans 开源统计机器翻译系统中基于短语的统计机器翻译模型构建实验平台[23],在孟加拉语、泰米尔语、乌兹别克语、匈牙利语至汉语 4 种翻译任务上对本文提出的方案进行评价,4 种语言与英语的平行训练数据、开发集以及测试集的详细信息如表 1 所示,英语汉语间的平行训练数据、开发集、测试集如表 2 所示.

表 1　外国语英语双语平行数据统计情况(词汇数为双语端的词汇数,使用符号'/'进行分割,语料按照训练数据中句数从小到大进行排序,*M*=百万)

源语言	数据	数据规模	
		句数	词汇数
孟加拉语	训练	5679	98809/117920
	开发	1169	19271/22942
	测试	636	9882/11517
泰米尔语	训练	11122	128106/158575
	开发	1577	17535/20682
	测试	720	7777/9669
乌兹别克语	训练	88713	1.49M/1.82M
	开发	1135	19853/22733
	测试	600	10110/11440
匈牙利语	训练	109268	2.21M/2.58M
	开发	1159	20515/23862
	测试	558	10365/12095

表 2　英语与汉语平行数据统计情况

句对	数据	数据规模	
		句数	词汇数
英汉	训练	2.04M	67.1M/56.6M
	开发	995	42407/25305
	测试	1859	77691/44654

所有翻译系统的构建均使用 4.2 节中提到的基于短语的翻译模型.短语翻译表中源语言和目标语言的短语长度限制最长为 5 个单词,调序距离限制为 10.在实验结果中,本文使用系统级中间语方法作为基线翻译系统.基线翻译系统使用GIZA++训练词对齐,使用 *grow-diag-final-and* 方法对双向的词对齐结果进行对称化处理.

8.1　系统级中间语方法结果

在系统级中间语方案中,首先构建外国语至英语的机器翻译系统,外国语英语翻译系统使用数据如表 1 所示,数据来源于 LDC 关于灾害的(disaster)的外国语至英语的双语数据.语言模型使用英文GIZAWORD(LDC2011T07)中 xinhua 数据部分加上外国语英语平行数据的英文部分训练一个五元语言模型.机器翻译系统性能如表 3 所示.由于使用的语料规模有限,语言的复杂程度不同,所以 *BLEU* 值的有所不同,在 4 个语言对的测试集上,*BLEU* 值的范围是 11.7～25.5,在孟加拉语上得到最低的 *BLEU* 值为 11.7,在匈牙利语上得到最高的 *BLEU* 值为 25.5.

表 3　外国语至英语翻译系统性能(目标语言为英语)

源语言	*BLEU* 4/%	
	开发集	测试集
孟加拉语	10.4	11.7
泰米尔语	16.3	19.4
乌兹别克语	18.5	18.3
匈牙利语	29.9	25.5

在外国语至英语的机器翻译系统构建完毕后,构建英语至汉语的机器翻译系统,英汉翻译系统使用数据如表 2 所示.训练数据来源于 NIST MT 2008,开发集为 SSMT-2007 英汉测试集,测试集为 NIST MT 08 英汉测试集.解码器、短语翻译表、调序距离的设定与外国语英语翻译系统相同.语言模型使用中文 GIZAWORD(LDC2009T27)中 xinhua 数据部分加入英汉训练数据中汉语部分训练一个五元语言模型.最终训练的英汉翻译系统性能如表 4 所示,在测试集上的 *BLEU* 值为 21.1.

表 4 英汉翻译系统性能

翻译系统	*BLEU* 4/%	
	开发集	测试集
英汉翻译	22.4	21.1

在这种方案中,本文构建了 4 种外国语至英语的翻译系统,1 种英汉翻译系统.在进行外国语至汉语的翻译时,如翻译孟加拉语,首先将孟加拉语翻译至英语,继而将英语翻译至汉语,最终得到待翻译孟加拉语对应的汉语翻译结果.

为了更好地评价机器翻译系统的质量,并且与语料级中间语以及改进的短语级中间语方案进行比较,本文使用 5.1 节中构建的外国语至汉语的开发集与测试集对系统级中间语方案的翻译结果进行评价,最终的外国语至汉语的翻译性能如表 5 所示.本文使用表 5 所示实验结果作为基线系统,与改进的语料级中间语方法和改进的短语级中间语方法进行比较.在这种评价体系下,所有的翻译系统均对比相同的参考答案,故不同中间语方法间的 *BLEU* 值具有可比性.与此同时,参考答案是由真实的英文语料通过英汉翻译系统进行翻译,故与基于中间语的方法进行两次翻译得到的翻译结果相比,准确度更高,故基本符合测试集的要求.

表 5 系统级外国语至汉语翻译系统性能(目标语言为汉语)

源语言	*BLEU*/%	
	开发集	测试集
孟加拉语	10.4	10.8
泰米尔语	14.5	19.9
乌兹别克语	16.4	17.2
匈牙利语	26.3	23.8

8.2 改进的语料级中间语方法结果

在语料级中间语方案中,首先通过英汉翻译系统将外国语英语双语间的平行数据、开发集以及测试集转换为外国语汉语间的平行数据、开发集以及测试集.此处使用的英汉翻译系统与系统级中间语方案中使用的英汉翻译系统相同.在这里,使用 5.2 节中提出的 m 最优翻译结果,m 取值为 5.最终构建的外国语汉语翻译系统的翻译性能如表 6 所示.通过对比表 5 与表 6 可知,在孟加拉语、泰米尔语、乌兹别克语至汉语的翻译任务上,语料级中间语的方法在开发集上的 *BLEU* 显著优于基于系统级中间语的基线方法,在测试集上的结果两种方案基本相同.然而,在语料规模最大的匈牙利语至中文的翻译任务上,语料级中间语方法与基线系统相比在开发集和测试集上分别高出 1.2 和 1.1 个 *BLEU* 点.

表 6 语料级外国语至汉语翻译系统性能
(目标语言为汉语,m 取值为 5)

源语言	*BLEU* 4/%	
	开发集	测试集
孟加拉语	11.2	10.7
泰米尔语	15.7	20.1
乌兹别克语	17.5	17.1
匈牙利语	27.5	24.9

在表 7 中,本文通过使用 5.3 节中提出的优化词对齐的方法直接优化外国语汉语翻译系统的翻译性能.对比表 7 与表 5 和表 6 可知,简单的词对齐优化方法在 4 种翻译任务上对翻译性能都有大幅度提高.与表 5 中的基线方法的结果相比,本文提出的改进方法在 4 种语言的测试集上分别提高了 1.2、0.7、0.7 及 1.3 个 *BLEU* 点.从这个结果中可以看出,与系统级中间语方案相比,语料级中间语方案的优化更为简单、直接、有效,最终得到较为理想的翻译性能.

表 7 改进的语料级外国语至汉语翻译系统性能

源语言	*BLEU* 4/%	
	开发集	测试集
孟加拉语	12.1	12.0
泰米尔语	16.5	20.6
乌兹别克语	17.6	17.9
匈牙利语	27.9	25.1

8.3 改进的短语级中间语方法结果

在改进的短语级中间语方法中,本文使用语料级中间语方法中生成的开发集和测试集分别进行系统的调优和评价工作.与此同时,调序模型使用语料级中间语方法中生成的调序模型.在直接生成外国语汉语短语翻译表过程中使用的英汉翻译系统与前两种方法相同.

在改进的短语级中间语方法中,本文统计了在 4 种外国语至英语的短语翻译表中,去重后的英文部分短语在英汉短语翻译表中的匹配与未匹配情况,具体匹配率与未匹配率如图 9 所示.从图 9 中可以看出,在 4 种外国语的短语翻译表中,不能匹配的英语短语的比例分别占到了 59.77%,67.26%,75.93%和 80.82%.从这个统计结果可知,在使用传统的短语级中间语方法中,大量的外国语英语短语对无法与英语汉语短语对进行融合.本文提出的改进方法生成了大量的英汉短语翻译表中不存在的短语对,从而有效地进行了外国语英语短语翻译表与英语汉语短语翻译表的整合工作.

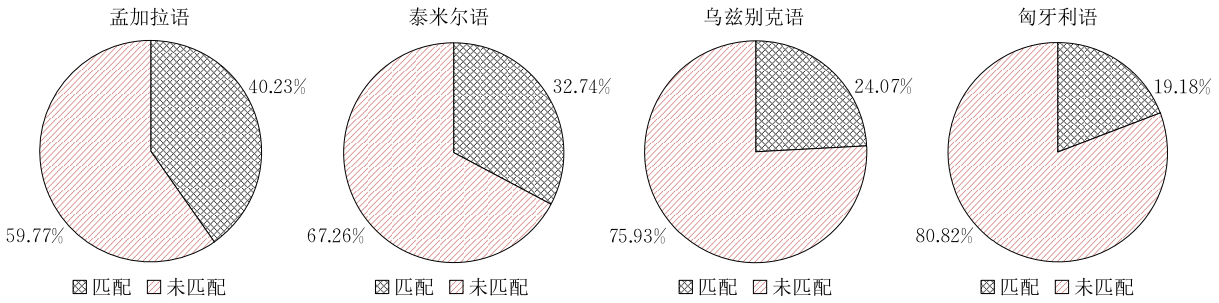


图 9　孟加拉语、泰米尔语、乌兹别克语、匈牙利语至英语的短语翻译表中的英语短语部分在英汉短语翻译表中匹配与未匹配的比例情况

改进的短语级中间语方法翻译性能如表 8 所示.通过与表 5 中的基线方法的结果相比,在孟加拉语与乌兹别克语上,短语级中间语方法在测试集上 *BLEU* 值显著的优于基线系统.然而,在泰米尔语和匈牙利语上,在与孟加拉语、乌兹别克语相同配置的实验条件下,短语级中间语方法却又显著的差于基线方法.通过本组实验可以看出,与系统级和语料级方法相比,短语级中间语方案显示出了一定的不稳定性.通过分析可以看出,由于该方案无法直接生成开发集、测试集、调序模型这 3 个构建机器翻译系统非常重要的部件,所以使用语料级方法中生成的相应模块进行替代,从而导致了翻译性能的不稳定.

表 8　短语级外国语至汉语翻译系统性能(目标语言为汉语)

源语言	*BLEU* 4/%	
	开发集	测试集
孟加拉语	13.4	13.2
泰米尔语	14.7	18.9
乌兹别克语	16.3	17.8
匈牙利语	26.0	23.1

8.4　语料级-短语级融合的中间语方法实验结果

表 9 为语料级-短语级融合的中间语方法的实验性能.从表 9 中可以看出,在所有翻译任务上,融合的方法均取得了最佳的翻译性能.本质上来说,融合的中间语方法是在语料级、短语级的翻译结果中选择最佳的翻译结果进行输出,故而得到优良的翻译性能.

表 9　语料级-短语级融合方法实验性能

源语言	*BLEU* 4/%	
	开发集	测试集
孟加拉语	13.4	13.6
泰米尔语	16.7	20.7
乌兹别克语	17.8	18.0
匈牙利语	28.0	25.4

8.5　性能分析

图 10 为本文提出的方法在孟加拉语、泰米尔语、乌兹别克语、匈牙利语 4 种外国语上的翻译性能的比较结果.从图 10 中可以看出,在 4 种翻译任务上,本文使用的基于最小错误率训练的语料级-短语级融合的中间语方法翻译效果最好、性能最为稳定.与此同时,除最小规模的孟加拉语外,语料级中间语方法在其他 3 种语言对上均表现出良好的翻译性能.系统级中间语方法在所有翻译任务上表现最差.但是,不同的中间语方法在不同的语言对上并没有表现出相同的增长趋势,这也间接说明不存在一种单一的中间语方法在所有语言翻译任务上均取得最优的翻译性能,但是系统融合的方法是一个不错的选择.不同语言上的翻译性能与训练数据质量、参数优化效果以及语言本身的差异性相关.

图 11 为本文提出的方法在孟加拉语翻译任务上测试集的翻译实例.从人工阅读的角度出发,由于使用大规模数据构建的性能优异的英汉翻译系统对孟加拉语-英语双语平行句对的英文部分进行翻译,得到的机翻汉语参考答案基本正确.系统级中间语方法翻译结果性能最差,造成该错误的主要原因是由于第 1 级的孟加拉语-英语翻译系统是构建在小规模训练数据基础上的,故翻译性能较差,继而得到较差的英文翻译结果,最终错误蔓延至第 2 级翻译系统上.语料级-短语级融合的中间语方法则使用最小错误率训练的方式在语料级和短语级中间语方法中选取最佳的翻译结果.

虽然基于统计的机器翻译方法与具体语言无关,但是对于语言的了解则有助于更好地训练基于统计的翻译系统,更好的对原始数据进行预处理操作.孟加拉语、泰米尔语、乌兹别克语、匈牙利语的原始数据与预处理后的数据如表 10 所示.与汉语在书写过程中不添加空格的方式相比,4 种语言在词与词之间显性的加入空格,故 4 种语言的分词操作为对标点符号的切分.在乌兹别克语与匈牙利语中,数据预处理操作还包括对大写字母的小写化操作,以缓解数据稀疏问题.当然,英语预处理中的提取词根操作同时可以应用到这 4 种语言对上,但本方法不在本文的讨论范围,在此不在赘述.

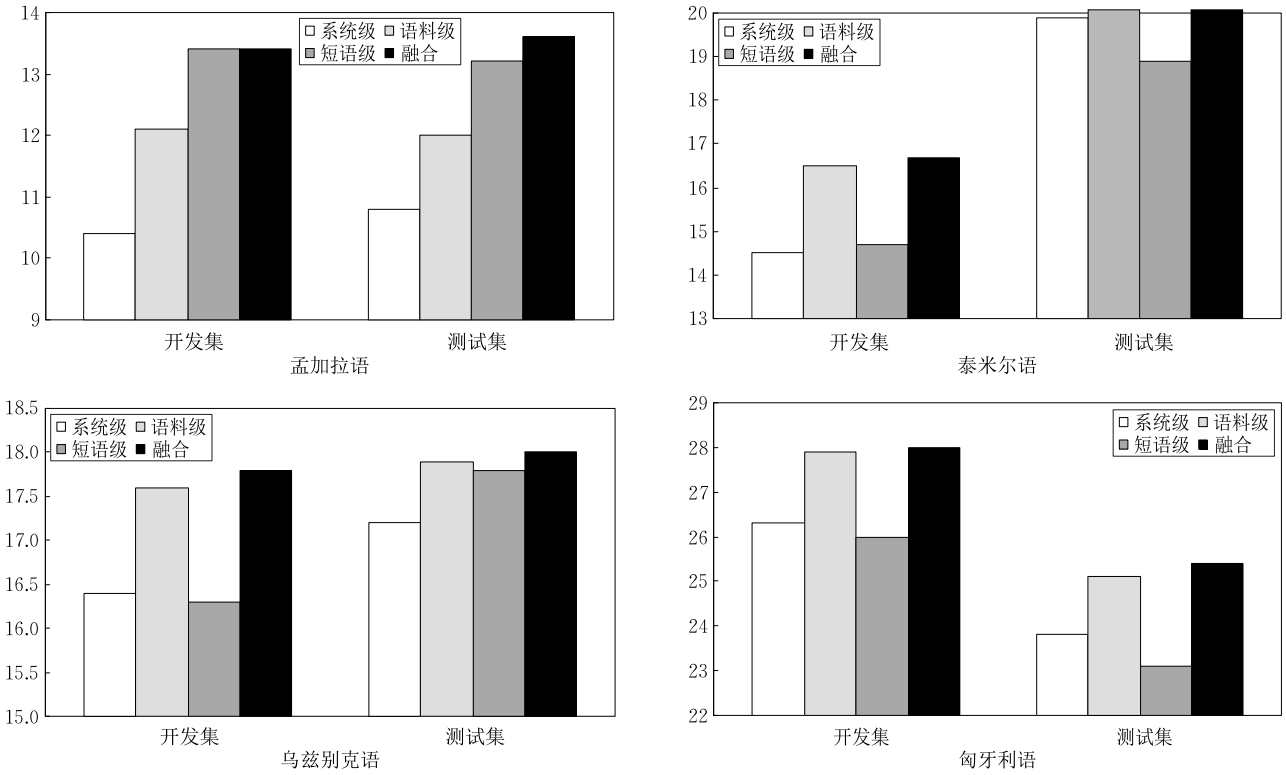


图 10 系统级、语料级、短语级以及融合的中间语方法在 4 种语言上的翻译性能

孟加拉语	তিনি আরো বলেছেন , স্থানীয় সরকার , জনগণ , সামরিক পক্ষ ও আন্তর্জাতিক সংস্থা তাদের কাজের উচ্চ মূল্যায়ন করেছে ।
英语	he also said that the local government, the people, the military and other international organizations have highly praised their efforts.
机翻汉语参考答案	他 还 说，当地 政府，人民，军队 和 其他 国际 机构 高度 评价 他们 的 努力.
系统级翻译结果	他 还 说，当地 政府，人民，军队 已 高度 政党 和 国际 机构 的 工作.
语料级翻译结果	他 还 说，当地 政府，人民，民间 团体 和 国际 机构 高度 评价 了 他们 的 工作.
短语级翻译结果	他 还 说，当地 政府，人民，军队 和 国际组织 高度 评价 了 他们 的 工作.
融合方法翻译结果	他 还 说，当地 政府，人民，军队 和 国际组织 高度 评价 了 他们 的 工作.

图 11 系统级、语料级、短语级以及融合的中间语方法在孟加拉语上翻译实例

表 10 4 种语言原始数据与预处理数据

语言	格式	句子
孟加拉语	原始	তিনি আরো বলেছেন, স্থানীয় সরকার, জনগণ, সামরিক পক্ষ ও আন্তর্জাতিক সংস্থা তাদের কাজের উচ্চ মূল্যায়ন করেছে।
	预处理	তিনি আরো বলেছেন, স্থানীয় সরকার, জনগণ, সামরিক পক্ষ ও আন্তর্জাতিক সংস্থা তাদের কাজের উচ্চ মূল্যায়ন করেছে ।
泰米尔语	原始	கேள்வி:- மத்திய அமைச்சர் அத்வானியுடன் பேசப்பட்டதா?
	预处理	கேள்வி : - மத்திய அமைச்சர் அத்வானியுடன் பேசப்பட்டதா?
乌兹别克	原始	O‘zbekiston Respublikasi Davlat madhiyasi yangradi.
	预处理	o ‘zbekiston respublikasi davlat madhiyasi yangradi .
匈牙利语	原始	A British Telecomnál 1990 óta megszüntetett 100 ezer munkahelyet szinte teljes egészében sikerült ellensúlyozni.
	预处理	a british telecomnál 1990 óta megszüntetett 100 ezer munkahelyet szinte teljes egészében sikerült ellensúlyozni .

9 总结与展望

在没有高质量的人工翻译的双语平行数据的条件下，本文研究了以英语作为中间语言构建外国语至汉语机器翻译系统的技术. 在使用常规的系统级、语料级、短语级中间语方法的基础上，本文提出了改进的语料级中间语方法以及改进的短语级中间语方法用于构建外国语至汉语的机器翻译系统. 通过对翻译结果性能的分析，本文使用基于最小错误率训练的方法对语料级-短语级的翻译结果进行融合. 通过使用本文提出的方法成功构建了孟加拉语、泰米尔语、乌兹别克语、匈牙利语至汉语的机器翻译系统. 与系统级中间语方法相比，改进的语料级和改进的短语级中间语方法可以对外国语汉语翻译系统直接进行建模与优化，明显减少了翻译系统构建过程

中的不确定性,为未来外国语至汉语翻译工作打下了良好的基础.最终,与基线系统相比,本文提出的方法在 4 种外国语言的测试集上分别获得了 0.8～2.8 个 BLEU 点的上涨.

在未来的工作中,我们将在两个方面对外国语至汉语的翻译技术进行研究.(1)我们会使用基于中间语言的方法将更多的外国语翻译成汉语;(2)使用更加优秀的技术来提高机器翻译系统的翻译性能,如基于深度神经网络的翻译方法[24-26].

致　谢　本文主要内容为第一作者作为联合培养博士生在南加州大学信息科学研究所完成(ISI/USC),在完成工作的过程中,得到外导 Kevin Knight 教授的悉心指导,在此表示衷心感谢!

参　考　文　献

[1] Koehn P, Och F J, Marcu D. Statistical phrase-based translation//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Edmonton, Canada, 2003: 48-54

[2] Chiang D. Hierarchical phrase-based translation. Computational Linguistics, 2007, 33(2): 201-228

[3] Devlin J, Zbib R, Huang Z, et al. Fast and robust neural network joint models for statistical machine translation//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 1370-1380

[4] Dabre R, Cromieres F, Kurohashi S, Bhattacharyya P. Leveraging small multilingual corpora for SMT using many pivot languages//Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL. Denver, USA, 2015: 1192-1202

[5] Wu H, Wang H. Pivot language approach for phrase-based statistical machine translation. Machine Translation, 2007, 21(3): 165-181

[6] Wu H, Wang H. Revisiting pivot language approach for machine translation//Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. Suntec, Singapore, 2009: 154-162

[7] Zoph B, Yuret D, May J, Knight K. Transfer learning for low-resource neural machine translation//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: To Appear

[8] Luong M, Manning C D. Stanford neural machine translation systems for spoken language domains//Proceedings of the 12th International Workshop on Spoken Language Translation. Da Nang, Vietnam, 2015: 76-79

[9] Jean S, Firat O, Cho K, et al. Montreal neural machine translation systems for WMT'15//Proceedings of the 10th Workshop on Statistical Machine Translation. Lisboa, Portugal, 2015: 134-140

[10] Koehn P, Birch A, Steinberger R. 462 Machine translation systems for Europe//Proceedings of the MT Summit XII. Ottawa, Canada, 2009: 65-72

[11] Zhu X, He Z, Wu H, et al. Improving pivot-based statistical machine translation using random walk//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 524-534

[12] Zhu X, He Z, Wu H, et al. Improving pivot-based statistical machine translation by pivoting the co-occurrence count of phrase pairs//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1665-1675

[13] Och F J, Ney H. Improved statistical alignment models//Proceedings of the 38th Annual Meeting on Association for Computation Linguistics. Hong Kong, China, 2000: 440-447

[14] Och F J, Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003, 29(1): 19-51

[15] Xiong D, Liu Q, Lin S. Maximum entropy based reordering model for statistical machine translation//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL. Sydney, Australia, 2006: 521-528

[16] Chen S F, Goodman J. An empirical study of smoothing techniques for language modeling. Computer Speech & Language, 1999, 13(4): 359-393

[17] Och F J. Minimum error rate training in statistical machine translation//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Sapporo, Japan, 2003: 160-167

[18] Papineni K, Roukos S, Ward T, Zhu W. BLEU: A method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318

[19] Galley M, Hopkins M, Knight K, Marcu D. What's in a translation rule//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Boston, USA, 2004: 273-280

[20] DeNero J, Klein D. Tarloring word alignments to syntactic machine translation//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic, 2007: 17-24

[21] Tu Z, Liu Y, He Y, et al. Combining multiple alignments to improve machine translation//Proceedings of the COLING 2012. Mumbai, India, 2012: 1249-1260

[22] Och F J, Ney H. Discriminative training and maximum entropy models for statistical machine translation//Proceedings of the 40th Annual Meeting on Association for Computation Linguistics. Philadelphia, USA, 2002: 295-302

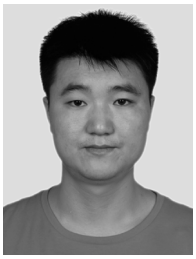
[23] Xiao T, Zhu J, Zhang H, Li Q. NiuTrans: An open source toolkit for phrase-based and syntax-based machine translation//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju, Republic of Korea, 2012:

19-24

[24] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural network//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 3104-3112

[25] Cho K, Gulcehre B V M C, Bahdanau D, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1724-1734

[26] Luong M, Pham H, Manning C D. Effective approaches to attention-based neural machine translation//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal. 2015: 1412-1421



LI Qiang, born in 1988, Ph. D. candidate. His research interests include natural language processing and machine translation.

WANG Qiang, born in 1990, Ph. D. candidate. His research interest is machine translation.

XIAO Tong, born in 1982, Ph. D., associate professor. His research interest is machine translation.

ZHU Jing-Bo, born in 1973, Ph. D., professor. His research interests include natural language processing and machine translation.

Background

Our work belongs to the research of pivot language based machine translation. My paper addresses the problems of foreign languages-Chinese translation as parallel training corpora are non-existent.

Previous work proposed several methods that used pivot language based method to improve translation performance. Wu and Wang (2007) propose a phrase-level method for phrase-based SMT on language pairs with limited bilingual resources by introducing pivot languages. To perform translation between Ls and Lt, they introduce a pivot language Lp, for which there are large $Ls-Lp$ and $Lp-Lt$ corpora which can be used to induce a translation model for $Ls-Lt$. As some useful source-target translations cannot be generated if the corresponding source phrase and target phrase connect to different pivot phrases, Zhu et al. (2013) utilize Markov random walks to connect possible translation phrases between source and target language. Dabre et al. (2015) present their work on leveraging multilingual parallel corpora of small sizes for statistical machine translation between Japanese and Hindi using multiple pivot languages. They focus on a variety of ways to exploit phrase tables generated using multiple pivots to support a direct source-target phrase table. Finally, they obtain improvements of up to 3 *BLEU* points using multiple pivots for Japanese to Hindi translation compared to when only on pivot is used. The method that Dabre et al. used is also a phrase-level method. Koehn et al. (2009) use the system-level pivot language based method and build 462 machine translation systems for all language pairs of the Acquis Communautaire corpus. They compare these systems against pivot translation. When using English as pivot, they find for two thirds of 462 language pairs there are significant gains (2～10 *BLEU* points).

In our work, we use English as pivot language to build foreign languages-Chinese translation systems. First, we use a typical system-level pivot language based approach which chains together the foreign-English and the English-Chinese system. Given a foreign sentence, system-level based method translates it to English with foreign-English system, and then translates it to Chinese with English-Chinese system. Second, we propose the improved corpus-level method which improves the translation performance through enlarging the size of bilingual training corpora and improving the quality of word alignments. The advantage of the corpus-level method lies in the fact that we can perform translation between foreign languages and Chinese even if there are no bilingual corpora available for these language pairs. Third, we propose the improved phrase-level approach which uses decoding-generation method to enlarge the size of phrase translation table and improve the translation performance. For the improved phrase-level approach, it combines two phrase tables from foreign language-English system and English-Chinese system into one foreign languages-Chinese phrase table, and then translation system is built on the foreign language-Chinese phrase table. Also, we propose the corpus-phrase combination method which achieves the best translation performance among all the translation tasks. We analyze the strengths and weaknesses for the typical system-level approach and our improved corpus-level and improved phrase-level approaches during system construction in our paper. Finally, we translate Bengali, Tamil, Uzbek, and Hungarian into Chinese with our proposed methods. With our proposed approaches, optimization methods can be easily integrated into our system to improve translation performance.

This work is mainly supported by the Fundamental Research Funds for the Central Universities under Grant No. N140406003, the National Natural Science Foundation of China under Grant Nos. 61272376 and 61300097.