同源语料增强的低资源神经机器翻译

王琳[1], 刘伍颖[2]

[1](上海外国语大学 贤达经济人文学院，上海 200083)

[2](广东外语外贸大学 语言工程与计算实验室，广东 广州 510420)

摘要：缺少平行句库的低资源机器翻译面临跨语言语义转述科学问题。围绕具体的低资源印尼语-汉语机器翻译问题，探索了基于同源语料的语言资源扩建方法，并混合同源语料训练出神经机器翻译模型。这种混合语料模型在印尼语-汉语机器翻译实验中取得了20.30的BLEU4评分。对实验结果的人工简单随机抽样分析发现混合语料模型的机器翻译效果与同时期的谷歌翻译效果相当。对于某些句子混合语料模型的译文质量甚至更优。实验结果证明同源语料能够有效增强低资源神经机器翻译，而这种有效性主要是源于同源语言之间的形态相似性和语义等价性。

关键词：同源语料, 低资源机器翻译, 神经机器翻译, 印尼语, 马来语

中图法分类号：TP 391.1　　**文献标识码**：A

Cognate-Corpus-Enhanced Low-Resource Neural Machine Translation

Lin Wang[1], **Wuying Liu**[2]

[1](Xianda College of Economics and Humanities, Shanghai International Studies University, Shanghai 200083, China)

[2](Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510420, China)

Abstract. Low-resource machine translation lacking parallel sentence corpus faces a scientific problem of cross-language semantic paraphrasing. We address the specific low-resource machine translation issue from Indonesian to Chinese, explore a language resource extension method based on a cognate corpus, and train a neural machine translation (NMT) model by mixing a cognate corpus. This model from mixed corpus achieved 20.30 BLEU4 score in the experiment of Indonesian-Chinese machine translation. The manual analysis after simple random sampling of experimental results finds that the effect of the mixed corpus NMT is comparable to that of the contemporary Google translation. For some sentences, the translation quality of mixed corpus model is even better. The experimental results prove that the cognate corpus can improve the low-resource NMT effectively, which mainly depends on the morphological similarity and semantic equivalence between the cognate languages.

Keywords: Cognate Corpus, Low-Resource Machine Translation, Neural Machine Translation, Indonesian, Malay

1. 引 言

在超大规模外语与汉语平行句库的支持下，基于深度学习的神经机器翻译(neural machine translation, NMT)的效果比较理想[1]，已经能够满足浏览和粗译等有限机器翻译需求[2]。但对于缺少超大规模语料库的低资源机器翻译问题尚在探索之中。而全世界现存的7 000多种语言当中，绝大部分的非通用语言都不同程度上存在与汉语平行句库稀缺的问题。因此，低资源机器翻译研究成为当前极具挑战的研究热点[3]。

印度尼西亚是东南亚最重要的国家之一，在中国“一带一路”倡议与印尼“全球海洋支点”战略的遥相呼应之下，中国同印尼的经贸、文教、科技等交流合作不断深化。印尼语是印度尼西亚的官方语言，属于南岛语系马来波利尼西亚语族。目前母语为印尼语的人数为4 500万，而全球印尼语使用者高达1.6亿。但印尼语与汉语的平行句库相对于英语等通用语言而言仍然规模不够大。为了能够更高效地服务中国和印尼的交流合作，我们从语言资源建设入手增强印尼语-汉语神经机器翻译模型。

同源语言之间的形态相似性有望缓解语言资源稀缺问题。已有研究表明不仅英国英语和美国英语之间具有极强的形态相似性，在英语、德语、法语以及其他欧洲语言之间也具有较多的共有词汇。已有的研究还发现汉语、日本语、韩国语和越南语中有60%以上的共有词汇[4]。受语言同源性和跨语言交流的影响，相似的形态往往蕴含等价的语义，这在同源语言之间表现得更为明显。基于这些认识，我们希望利

基金项目：教育部人文社会科学研究规划基金项目(20YJAZH069)；教育部人文社会科学研究青年基金项目(20YJC740062)；上海市哲学社会科学“十三五”规划课题(2019BYY028)；广州市科技计划项目(202201010061)

作者信息：1. 王琳(1983-)，副教授，硕士，硕士生导师，研究方向为计算语言学和语料库语言学，E-mail: lwang@xdsisu.edu.cn；2. 刘伍颖(1980-)，专职研究员，云山学者，博士，硕士生导师，研究方向为计算语言学和自然语言处理，E-mail: wyliu@gdufs.edu.cn

用与印尼语同源的马来语-汉语句库扩建印尼语-汉语句库，并混合同源语料改进印尼语-汉语神经机器翻译模型。

2. 相关研究

国外较早开展了相关研究。早在1998年，美国国防高级研究计划署(DARPA)集成推出跨语言信息发现、提取和文摘(TIDES)项目[5]，旨在对多语言信息进行自动发现、提取、摘要和翻译。接着2006年底，DARPA又启动全球自动化语言开发(GALE)项目[6]，从语音转录、翻译和过滤入手研发计算机软硬件，收集、分析和解释巨量多语言文本和语音信息，提高语言翻译和文本分析效率，及时为决策者提供过滤后的高价值信息。此后，DARPA还相继启动了针对光学图像文本分析和翻译的MADCAT项目、针对语音识别和文本转录的RATS项目、面向口语交流和翻译的TRANSTAC实用系统研发项目等。

这期间的国外研究主要围绕阿拉伯语、普什图语、汉语等当时几种特定的语言，采用系统的顶层架构和标准的数据链接口，研究项目之间能够互相衔接，支持快速搭建非英语到英语的机器翻译系统，服务军事情报和国家安全领域。而面向特定语言的机器翻译研究从最初采用规则方法逐渐过渡到采用统计方法[7]，最终确立了统计机器翻译(statistical machine translation, SMT)的主导地位[8]。同时带动了语言资源工程的研发，促使美国宾夕法尼亚大学(UPenn)的语言资源联盟(LDC)壮大成为国际语言资源组织。最初的规则机器翻译主要是根据形式语言理论，围绕上下文无关文法进行转换生成翻译。而主流的统计机器翻译是根据噪声信道理论，采用Bayes条件概率公式将机器翻译间接转化为翻译模型和语言模型的学习和概率计算，产生了经典的基于词、基于短语和基于句法树的统计机器翻译算法。

随着大数据时代的到来，2012年初，美国政府投入2亿美元发起大数据研发倡议，推动在人工智能领域利用语言大数据。紧接着2012年底，DARPA和美国空军研究实验室启动文本深度勘探和过滤(DEFT)项目[9]，旨在推断语义、探求关系、发现文本异常，辅助分析师高效处理多语言文本大数据。接着2014年，DARPA启动紧急事件低资源语言(LORELEI)项目[10]，旨在研发处理全语言的系统，让分析人员在短时间内掌握话题、名字、事件、情绪、关系等关键信息。2017年4月，DARPA又启动差异备选方案的主动译释(AIDA)项目，旨在融合多来源多模态且潜在矛盾和欺骗的语种混杂信息，生成真实世界事件、现状和趋势的方案译释。该项目将语言资源工程升级为语义知识工程，通过整合语境碎片实现事件和趋势的全息认知以辅助分析决策。此外，2012年，美国情报高级研究计划署(IARPA)正式开始低资源语言语音识别(BABEL)项目的研究[11]，旨在有限语料和极短时间条件下，实现对任意新语言语音识别的跨语言关键词搜索。接着2017年初，IARPA又启动了对全语言信息进行英语检索的机器翻译(MATERIAL)项目，旨在实现对全球超过7 000种语言的文档进行跨语言信息检索。

国外近来的机器翻译研究对象已不再满足于几种特定语言，而是试图利用互联网语言大数据实现全语言覆盖。同时，深度学习方法[12]已取代统计方法的主导地位，在机器翻译上取得了重大突破[13]。深度学习方法也是一类统计学习方法，其根据神经网络理论，利用向量计算部件和大规模语料，直接训练翻译模型。迄今提出了“Sequence to Sequence神经机器翻译”[14]、“纯Attention神经机器翻译”[15]、“无监督神经机器翻译”[16]等一系列优秀算法。促使谷歌、百度、必应等机器翻译应用换装神经机器翻译引擎，并在英汉翻译效果上取得显著提升。但深度学习神经机器翻译算法的有效性很大程度上依赖于平行句库的规模与质量[17]。尽管深度学习方法在富资源语言上取得了很大成功，但对于低资源语言却效果有限[18]。

国内相关研究起步也不晚，几乎与国外研究同步发展。早在1998年，国内就已开展了汉英机器翻译评测[19]。从2003年开始，在国家863项目“中文信息处理及智能人机接口技术评测”[20]的支持下，连续几年开展了机器翻译评测，涉及英语、日本语、汉语等语言。在这些评测的带动下，国内也相继成立了中文语言资源联盟(CLDC)、国家语言资源监测与研究中心、中国语言资源开发应用中心、语言资源高精尖创新中心等组织，构建了一些英语、汉语和国内民族语资源。2007年，国家863重点项目“面向跨语言搜索的机器翻译关键技术研究”[21]和“面向网络海量信息处理多策略机器翻译系统研究”[22]启动。后来2011年，又启动国家863重大项目“互联网语言翻译系统研制”[23]。上述项目从关键技术探索到实用系统

研制也基本遵循从规则方法到统计方法[24]，再到混合方法的研究路径。近来，国内机器翻译研究主要围绕神经机器翻译方法展开[25][26]，而针对低资源机器翻译的探索也形成“大规模语言资源构建”和“高级机器学习算法开发”两种思路。例如有研究利用低资源语言形态特性抽取词典[27]、有研究从低资源语言概念层次思考资源体系[28]，有尝试基于中间语的低资源机器翻译[29]。还有研究围绕俄语[30]、日本语[31]、韩国语[32]、越南语[33]以及南岛语系的印尼语、马来语等进行了探索[34]，并取得了显著的前期成果[35]。与国外研究态势相似，一定程度上受限于相关语言专家稀缺的困扰，当前国内低资源机器翻译研究方兴未艾，印尼语-汉语机器翻译成果亟待提升。

3. 语言资源扩建方法

3.1. 总体框架

为了训练出高效的印尼语-汉语神经机器翻译模型，我们设计了图1所示的语言资源扩建总体框架，主要包括同源语言相似性分析、同源句库构建、神经机器翻译模型训练三部分。在同源语言相似性分析部分，我们定量分析了马来语到印尼语的共有形态率和语料迁移率。在同源句库构建部分，我们提出了基于可比语料的平行句库构建方法和基于中间语桥接的平行句库构建方法。除了采用简单并集的方法，简单合成同源句库构建部分生成的印尼语-汉语平行句库和马来语-汉语平行句库之外，我们还根据同源语言相似性分析结果，挑选部分有用的马来语-汉语句对迁移合成到印尼语-汉语平行句库中。

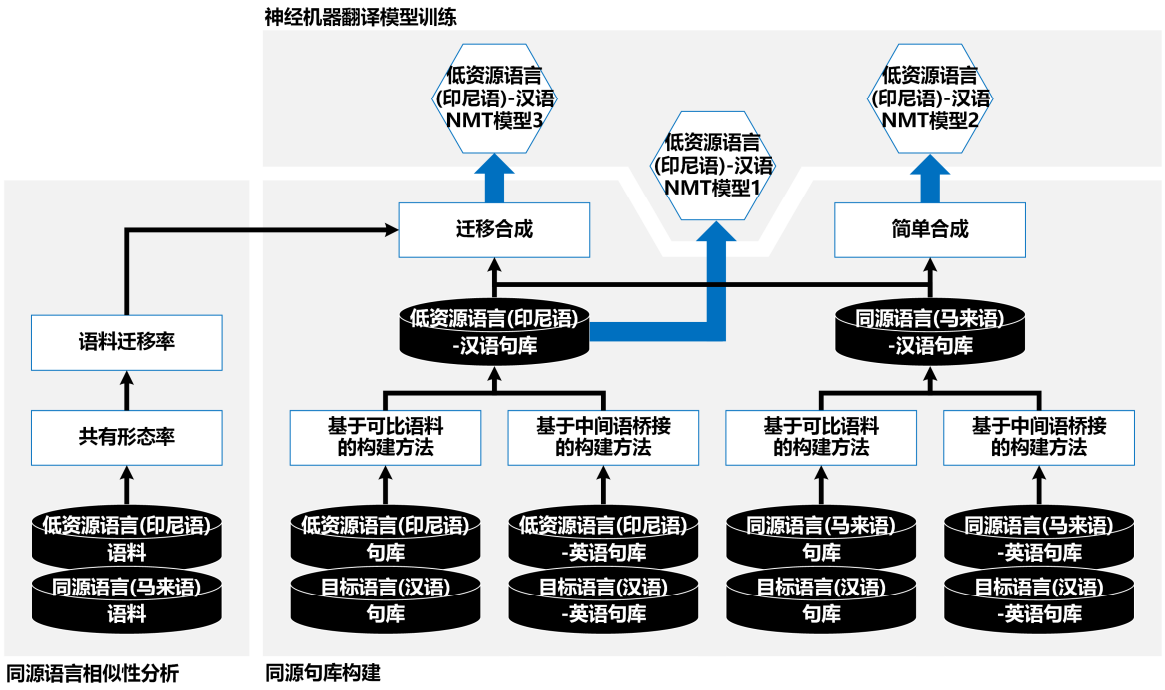

图1. 总体框架

Figure 1. General Framework

在神经机器翻译模型训练部分，我们利用同源句库构建部分生成的印尼语-汉语平行句库直接训练得到印尼语-汉语NMT1模型，利用简单合成的混合语料训练得到印尼语-汉语NMT2模型，利用迁移合成的混合语料训练得到印尼语-汉语NMT3模型。

3.2. 同源句库构建

为了分别构建印尼语-汉语平行句库、马来语-汉语平行句库，我们设计了两种工程级规则方法。

一是基于可比语料的平行句库构建方法。我们认为人类大脑的认知机能决定了自然语言语义的等价性。当存在印尼语(马来语)与汉语的可比语料，可以借助双语词典并根据可比语料中的标点符号、阿拉伯

数字等语义不动点抽取语义等价的双语句子。具体实现时，我们收集了印尼语(马来语)教材、网络双语新闻、Wikipedia等可比语料。

```
PSB-CC算法
Input: Df, Dc, threshold, {<Wf, Wc>}, {<Pf, Pc>}
Output: {<Sf, Sc>}
Begin
    <Sf>←SentenceSpliter(Df);
    <Sc>←SentenceSpliter(Dc);
    For each sentence Sf in <Sf>
        For each sentence Sc in <Sc>
            score1←WordScorer({<Wf, Wc>}, Sf, Sc);
            score2←PunctuationScorer({<Pf, Pc>}, Sf, Sc);
            score3←NumberScorer(Sf, Sc);
            max←Updater((score1+score2+score3)/3);
        End For
        If (max>threshold) Then {<Sf, Sc>}←Appender(<Sf, Sc>);
    End For
End
```

图2. 基于可比语料的平行句库构建算法

Figure 2. Comparable-Corpus-based Parallel Sentence Bank Construction Algorithm

具体的基于可比语料的平行句库构建算法(PSB-CC)如图2所示。输入可比文本对D_f和D_c，以及置信度分数阈值***threshold***、双语词典{<W_f, W_c>}、双语标点符号表{<P_f, P_c>}。输出平行句库{<S_f, S_c>}。主要过程是双层循环。其中SentenceSpliter()函数根据语种将文本切分成句子列表。三种评分器WordScorer、PunctuationScorer、NumberScorer采用类似的规则，分别从词命中、标点命中、数字命中给出句子对的平行对译概率。以NumberScorer规则为例，如果句子S_f包含n个数字，句子S_c包含m个数字，两个句子中相同的数字有k个，则数字置信度分数$score_3=k^2/(nm)$。再直接经过算术平均得到最终置信度分数，内循环时暂存最大的置信度分数***max***。内层循环结束时，如果最大置信度分数***max***大于事先设定的阈值***threshold***，则将句对<S_f, S_c>加入输出平行句库{<S_f, S_c>}。

二是基于中间语桥接的平行句库构建方法。我们认为自然语言不仅存在语义等价性，而且这种等价关系在不同形态表示的自然语言之间还具备传递性。当同时存在印尼语(马来语)与中间语大规模平行句库、中间语与汉语的大规模平行句库，我们利用形态传递性，计算中间语句子的相似度桥接构建印尼语(马来语)与汉语的平行句库。具体工程实现时，我们以英语为中间语。

```
PSB-IB算法
Input: {<Sf, Si1>}, {<Si2, Sc>}, threshold
Output: {<Sf, Sc>}
Begin
    For each pair of sentences <Sf, Si1> in {<Sf, Si1>}
        For each pair of sentences <Si2, Sc> in {<Si2, Sc>}
            distance←Levenshtein(Si1, Si2);
            min←Updater(distance);
        End For
        If (min<threshold) Then {<Sf, Sc>}←Appender(<Sf, Sc>);
    End For
End
```

图3. 基于中间语桥接的平行句库构建算法

Figure 3. Interlanguage-Bridged Parallel Sentence Bank Construction Algorithm

具体的中间语桥接的平行句库构建算法(PSB-IB)如图3所示。输入平行句库{<S_f, S_{i1}>}和{<S_{i2}, S_c>}，以及Levenshtein距离阈值***threshold***。输出平行句库{<S_f, S_c>}。主要过程也是双层循环。其中相同语种的句子对S_{i1}和S_{i2}，通过计算得到Levenshtein距离，内循环时暂存最小的Levenshtein距离***min***。内层循环结束时，如果最小的Levenshtein距离***min***小于事先设定的阈值***threshold***，则将句对<S_f, S_c>加入输出平行句库{<S_f, S_c>}。

采用上述两种算法，我们最终构建出的印尼语-汉语平行句库包含2 489 442句对，马来语-汉语平行句库包含761 373句对。其中印尼语-汉语平行句库样例如表1所示。

表1. 印尼语-汉语平行句库样例

Table 1. Examples of Indonesian-Chinese Parallel Sentence Corpus

印尼语	汉语
Acara ini sampai jam berapa?	这个表演到几点呢？
Aku berharap suatu hari aku bisa melakukan hal yang benar.	我只希望有一天我能处理好一切。
Anak perempuan ini bekerja sebagai pembantusebelum datang ke sekolah.	这个女孩在来上学之前是做女佣的。
Apakah kita sendirian di alam semesta?	我们是宇宙中唯一的文明吗？
Bali Masih Aman Untuk Dikunjungi!	巴厘岛对游客来说还是安全的！
Berapa yang kau butuhkan?	拍摄这部影片需要多少钱？
Dan permainan dimulai.	游戏就此开始。
Frekuensi serangan semakin meningkat.	频繁的袭击不断增加。
Indonesia adalah komunitas ekonomi yang terbesar di Asia Tenggara.	印尼是东南亚最大经济体。
Jika modul itu sudah mendarat, kau tak akan mau memindahkannya.	等飞船着陆后，就不方便移动了。
Kau harus kembali ke tempat dudukmu.	请你回到座位去。
Kerja keras dan dukungan para orang tua merupakan kunci kesuksesan mereka saat ini.	他们的成功离不开家长的辛勤培育和支持。
Kita berevolusi bersama peralatan, dan peralatan berevolusi bersama kita.	人类发明了工具，工具也影响着人类。
Lain kali ingin berbicara, lakukan itu sebelumnya.	下次你演讲之前，提前做到这几项。
Lebih baik tulis nama anda di bawah, suoaya penerima tahu Fax nya dari siapa.	最好在下面写上您的名字，收取人好知道传真是谁发的。
Lingkaran biru itu lebih kecil daripada lingkaran merah itu.	那个蓝色的圆比那个红色的圆小。
MasyarakatBali terkenal ramah dan penuh senyum.	巴厘人以友好善良和微笑而著名。
Saya bangun di dalam sel penjara, diborgol, dan mata disekap.	我发现自己在监狱里，带着手铐，被蒙住眼睛。
Tapi para murid kalah dari simpanse.	这些优等生们却做不到。
Tiongkok mendesak negara tetangga untuk meningkatkan upaya dalam Memerangi Ekstremisme.	中国敦促邻国加大极端主义打击力度。

3.3. 同源语言相似性分析

据统计当前母语为马来语的人数超过8 000万，主要分布在文莱、印度尼西亚、马来西亚、新加坡等地，而全球马来语使用者则超过3亿。广义而言印尼语也是一种马来语，类似于汉语与新加坡华语、英国英语与美国英语之间的关系，马来语与印尼语同属南岛语系马来波利尼西亚语族，具有极强的同源性，使得马来语与印尼语在发音、词汇、句法等方面十分相似。二者的拉丁形态字母表完全相同，就是英语字母表。因此同源语言之间的形态相似性可以被用于语言资源扩建。然而印尼语与马来语之间的形态相似性究竟有多大？我们通过共有形态统计给出具体的定量分析。

表2. 词汇级*N*-gram串数

Table 2. Number of Word-Level *N*-gram Token

***N*-gram**	**Indonesian** (MOR)	**Malay** (MOR)	**Overlap**
1-gram	**657 409** (32.16%)	**395 365** (53.48%)	**211 453**
2-gram	**8 035 771** (15.36%)	**3 994 868** (30.89%)	**1 233 931**
3-gram	**19 614 857** (6.51%)	**8 756 914** (14.58%)	**1 276 456**
4-gram	**23 038 089** (2.84%)	**9 708 373** (6.73%)	**653 356**
5-gram	**20 797 228** (1.66%)	**8 658 728** (4.00%)	**345 955**

我们以20180501版Wikipedia为统计分析源，该版Wikipedia包含印尼语文档bz2压缩包439.3MB、马来语文档bz2压缩包173.8MB。详细的去重后词汇级*N*-gram串数如表2所示。其中1-gram数据表明全部印尼语文档是由657 409个不同的印尼语词汇组成，而全部马来语文档只由395 365个不同的马来语词汇组成，两种语言的共有词汇数量多达211 453。

两种语言之间共有形态串数对每种语言的效力是不一样的。我们把这种具有方向性的A语言到B语言的*N*-gram形态共有率(Morphological Overlap Ratio，简记为$MOR_{N-gram}^{A\to B}$)定义为A语言和B语言共有*N*-gram串数除以B语言的*N*-gram串总数得到的百分比，具体$MOR_{N-gram}^{A\to B}$数值反映了A语言语料的*N*-gram形态学习结果用于B语言的有效程度。经过计算可知马来语(M)到印尼语(I)的1-gram形态共有率$MOR_{1-gram}^{M\to I}$=32.16%。

我们根据形态共有率计算公式，把A语言到B语言的语料迁移率(Corpus Transfer Ratio，简记为$CTR_N^{A\to B}$)定义为$\sum_{i=1}^{N}\alpha_i MOR_{i-gram}^{A\to B}$，其中系数$\alpha_i$表示A语言语料集中未去重*i*-gram串数除以全部未去重*N*-gram串数总和的百分比。由于语言的长程相关性相对较弱，通常只需要简化计算*N*=5时的$CTR_N^{A\to B}$用于表

示A语言语料到B语言语料的增强程度。基于马来语印尼语共有形态的语料迁移率能够支撑我们用马来语-汉语平行句库扩建印尼语-汉语平行句库的思路。

4. 机器翻译实验

为了验证马来语-汉语平行句库对印尼语-汉语神经机器翻译的增强作用，我们设计了印尼语-汉语神经机器翻译实验。

实验语料分成训练、验证、测试三个集合，其中验证集和测试集各包含20 000印尼语-汉语句对，都是从2 489 442句对印尼语-汉语平行句库中采用不放回采样技术随机抽取而成。我们把2 489 442句对的印尼语-汉语平行句库简记为IdChSens，把除去验证集和测试集后剩下的2 449 442句对命名为印尼语-汉语部分句库IdChSens-P，把761 373句对的马来语-汉语平行句库简记为MsChSens。接着我们还统计出MsChSens和IdChSens中马来语和印尼语的1-gram共有形态集OverlapWordSet。然后依次检测MsChSens中的每个马来语句子，如果该句子中的每个词都属于OverlapWordSet，则将包含该句子的马来语-汉语句对加入马来语-汉语部分句库MsChSens-P。由此可见MsChSens-P是MsChSens的一个子集，实际检测结果MsChSens-P包含696 840句对，也可以看成是来自马来语语料的印尼语语料。

实验中三个NMT模型是采用开源TensorFlow NMT Tutorial[1]分别在各自数据集上训练得到的Sequence to Sequence模型[36]。主要配置参数：

"attention": normed_bahdanau,
"attention_architecture": gnmt_v2,
"batch_size": 128,
"beam_width": 10,
"encoder_type": gnmt,
"epoch": 10,
"num_decoder_layers": 4,
"num_encoder_layers": 4,
"num_units": 512

三个神经机器翻译模型的实验结果如表3所示。我们发现：混合马来语语料，使得印尼语-汉语训练句对总数增强到300万以上时，神经机器翻译的BLEU4评分从17.13增长到20.30，效果显著提升；从NMT2和NMT3的BLEU4评分相同判断马来语语料中对印尼语-汉语神经机器翻译有增强作用的主要是共有形态语料。

表3. 神经机器翻译实验结果

Table 3. Experimental Results of Neural Machine Translation

Model	Train Set	Validation Set	Test Set	BLEU4
NMT1	IdChSens-P **2 449 442**	**20 000**	**20 000**	**17.13**
NMT2	IdChSens-P+MsChSens **3 210 815**	**20 000**	**20 000**	**20.30**
NMT3	IdChSens-P+MsChSens-P **3 146 282**	**20 000**	**20 000**	**20.30**

为进一步评价NMT3模型效果，我们从上述测试集中随机挑选了20句对，并将其中的印尼语句子提交给同时期的谷歌和必应机器翻译系统。印尼语专家结合印尼语源文对汉语参考译文、我们的神经机器翻译NMT3模型译文、谷歌和必应的机器翻译译文进行了量化评分。根据译文符合源文的程度依次评为4、3、2、1四档分数，分数越大表示译文质量越高。具体的源文、译文以及得分如表4所示，我们还利用底色从深到浅依次表示四档分数。表4分数显示我们的NMT3模型译文总共获得56分，总体译文质量最优，接下来依次是谷歌译文54分、必应译文47分、汉语参考译文45分。据此发现：一方面，我们改进的神经机器翻译效果与谷歌机器翻译效果相当，明显优于必应机器翻译效果，而马来语语料能够有效改进印尼语-汉

[1] https://github.com/tensorflow/nmt/

语神经机器翻译主要是源于同源语言之间的形态相似性和语义等价性；另一方面，印尼语-汉语平行句库中的汉语参考译文质量不是很高，但神经机器翻译能够从低质量语料中学出超过参考译文的翻译模型，这主要是归功于深度学习方法充分细化了学习特征，并充分利用了巨量训练数据。

5. 结 论

低资源机器翻译是挑战与价值俱高的研究问题。针对具体的低资源印尼语-汉语机器翻译问题，我们从马来语到印尼语形态共有率和语料迁移率的定量分析出发，探索了基于同源平行语料的资源扩建方法，并混合同源句库训练出改进的神经机器翻译模型，最终在印尼语-汉语机器翻译实验中取得了最优的翻译效果。同源句库能够有效改进低资源神经机器翻译主要是源于同源语言之间的形态相似性和语义等价性。

迄今效果优良的神经机器翻译实验系统大多需要1 000万句对以上规模的训练语料，而效果更优的神经机器翻译商用系统训练语料规模往往超过5 000万句对。未来平行句库构建研究中，我们希望在双语资源之外发挥单语资源的规模优势，通过高效机器翻译算法生成句对。此外针对神经机器翻译超参考质量的现象，需要探索平行句对的质量评价算法，优化训练语料质量和译文评价方法。最后将关注更多的其他语族，通过推广基于同源句库的资源扩建方法实现更多的非通用语言到汉语的有效机器翻译。

参考文献

[1] 刘洋. 神经机器翻译前沿进展. 计算机研究与发展, 54(6):1114–1149, 2017.

[2] Wuying Liu, Lin Wang. Fast-Syntax-Matching-based Japanese-Chinese Limited Machine Translation. The 5th Conference on Natural Language Processing and Chinese Computing (NLPCC) Proceedings, LNAI, 10102:621–630, 2016.

[3] Benjamin Philip King. Practical Natural Language Processing for Low-Resource Languages. Doctoral Dissertation, University of Michigan, 2015.

[4] Wuying Liu. Supervised Ensemble Learning for Vietnamese Tokenization. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 25(2):285–299, 2017.

[5] Christopher Cieri, Mark Liberman. TIDES Language Resources: A Resource Map for Translingual Information Access. The 3rd International Conference on Language Resources and Evaluation (LREC) Proceedings, ELRA, 1334–1339, 2002.

[6] Joseph Olive, Caitlin Christianson, John McCary. Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation. Springer, ISBN: 9781441977120, 2011.

[7] Nakov Preslav, Ng Hwee Tou. Improving Statistical Machine Translation for a Resource-Poor Language Using Related Resource-Rich Languages. Journal of Artificial Intelligence Research, 44:179–222, 2012.

[8] Philipp Koehn. Statistical Machine Translation. Cambridge University Press, ISBN: 9780521874151, 2009.

[9] Ann Bies, Zhiyi Song, Jeremy Getman, Joe Ellis, Justin Mott, Stephanie Strassel, Martha Palmer, Teruko Mitamura, Marjorie Freedman, Heng Ji, Tim O'Gorman. A Comparison of Event Representations in DEFT. The 4th Workshop on Events: Definition, Detection, Coreference, and Representation Proceedings, ACL, 27–36, 2016.

[10] Christopher Cieri, Mike Maxwell, Stephanie Strassel, Jennifer Tracey. Selection Criteria for Low Resource Language Programs. The 10th International Conference on Language Resources and Evaluation (LREC) Proceedings, ELRA, 4543–4549, 2016.

[11] K. M. Knill, M. J. F. Gales, A. Ragni, S. P. Rath. Language Independent and Unsupervised Acoustic Models for Speech Recognition and Keyword Spotting. The 15th Annual Conference of the International Speech Communication Association (INTERSPEECH) Proceedings, ISCA, 16–20, 2014.

[12] Yann LeCun, Yoshua Bengio, Geoffrey Hinton. Deep Learning. Nature, 521:436–444, 2015.

[13] Julia Hirschberg, Christopher D. Manning. Advances in Natural Language Processing. Science, 349(6245):261–266, 2015.

[14] Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Sequence to Sequence Learning with Neural Networks. The 28th Annual Conference on Neural Information Processing Systems (NIPS) Proceedings, Curran Associates, 3104–3112, 2014.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia

Polosukhin. Attention Is All You Need. arXiv:1706.03762v5, 2017.
[16] Mikel Artetxe, Gorka Labaka, Eneko Agirre, Kyunghyun Cho. Unsupervised Neural Machine Translation. arXiv:1710.11041v1, 2017.
[17] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144v2, 2016.
[18] Christopher D. Manning. Last Words: Computational Linguistics and Deep Learning. Computational Linguistics, 41(4):701–707, 2015.
[19] 刘洋, 刘群, 林守勋. 机器翻译评测中的模糊匹配. 中文信息学报, 19(3):45–53, 2005.
[20] 崔世起, 刘群, 孟遥, 于浩, 西野文人. 基于大规模语料库的新词检测. 计算机研究与发展, 43(5):927–932, 2006.
[21] 刘群. 机器翻译研究新进展. 当代语言学, 11(2):147–158, 2009.
[22] 李业刚, 黄河燕, 史树敏, 冯冲, 苏超. 多策略机器翻译研究综述. 中文信息学报, 29(2):1–9, 2015.
[23] Jiajun Zhang, Feifei Zhai, Chengqing Zong. A Substitution-Translation-Restoration Framework for Handling Unknown Words in Statistical Machine Translation. Journal of Computer Science and Technology, 28(5):907–918, 2013.
[24] 苏劲松, 董槐林, 陈毅东, 史晓东, 吴清强. 引入基于主题复述知识的统计机器翻译模型. 浙江大学学报(工学版), 48(10):1843–1849, 2014.
[25] Jiajun Zhang, Chengqing Zong. Deep Neural Networks in Machine Translation: An Overview. IEEE Intelligent Systems, 30(5):16–25, 2015.
[26] Yanzhuo Ding, Yang Liu, Huanbo Luan, Maosong Sun. Visualizing and Understanding Neural Machine Translation. The 55th Annual Meeting of the Association for Computational Linguistics (ACL) Proceedings, ACL, 1150–1159, 2017.
[27] Wuying Liu, Lin Wang. Vietnamese Multisyllabic-Word Extraction for Word Segmentation. International Journal of Asian Language Processing, 27(1):61–77, 2017.
[28] 于施洋, 杨道玲, 王璟璇, 傅娟. "一带一路"数据资源归集体系建设. 电子政务, 2017(1):8–14, 2017.
[29] 李强, 王强, 肖桐, 朱靖波. 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. 计算机学报, 40(4):925–938, 2017.
[30] Wenjun Du, Wuying Liu, Junting Yu, Mianzhu Yi. Russian-Chinese Sentence-level Aligned News Corpus. The 18th Annual Conference of the European Association for Machine Translation (EAMT) Proceedings, EAMT, 213, 2015.
[31] 刘伍颖, 张兴. 基于自然句法标记的日汉机器翻译架构. 山西大学学报(自然科学版), 41(1):1–9, 2018.
[32] Wuying Liu, Lin Wang. POS-Tagging Enhanced Korean Text Summarization. The 13th International Conference on Intelligent Computing (ICIC) Proceedings, LNCS, 10363:425–435, 2017.
[33] Wuying Liu, Li Lin. Probabilistic Ensemble Learning for Vietnamese Word Segmentation. The 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) Proceedings, ACM, 931–934, 2014.
[34] Wuying Liu, Lin Wang. Natural-Annotation-based Malay Multiword Expressions Extraction and Clustering. The 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), 2018.
[35] Junting Yu, Wuying Liu, Hongye He, Mianzhu Yi. BLEUS-syn: Cilin-Based Smoothed BLEU. The 12th China Workshop on Machine Translation (CWMT) Proceedings, CCIS, 668:102–112, 2016.
[36] Graham Neubig. Neural Machine Translation and Sequence-to-sequence Models: A Tutorial. arXiv:1703.01619v1, 2017.

表4. 印尼语-汉语机器翻译人工评测

Table 4. Manual Evaluation of Indonesian-Chinese Machine Translation

印尼语源文	汉语参考译文	45	我们的神经机器翻译译文	56	谷歌机器翻译译文	54	必应机器翻译译文	47
dengan kata lain, untuk efesiensi dan mengelola diri sendiri.	也就是自足和自理。	1	也就是自足和管理自我。	1	换句话说，就效率和自我管理而言。	3	换言之, 为了有效地使用和管理自己。	4
dude, anda tampak konyol.	哥们，你穿得太荒唐了。	2	伙计，你看起来很傻。	4	老兄，你看起来很傻。	4	伙计你看起来很可笑	3
haruskah kita masuk?	那我进去了？	1	我们要进去吗？	4	我们应该进入吗？	2	我们应该进去吗？	3
ini adalah pertama kali saya mendengar kabar darinya dalam setahun.	这是一年来我第一次有他的消息。	4	这是我一年来第一次听到他的消息。	4	这是我一年中第一次听到他的消息。	3	这是我一年来第一次收到他的来信。	2
istri polisi mungkin akan berguna.	一个警察的妻子也许有点用。	2	警察老婆可能有用。	3	警察的妻子可能会有用。	3	警察的妻子也许有用。	4
itu seperti rakun mabuk mendapat seluruh berlian hope.	就像一只醉酒的浣熊，手上拿着蓝宝石。	2	这就像一只醉醺醺的浣熊，得到了希望的钻石。	3	就像醉酒的浣熊得到了整个钻石的希望。	1	就像喝醉了的浣熊到处都是希望钻石	1
jangan ungkit lagi.	不要去想它。	3	别再提了。	4	不要备份。	2	不要再 ungkit 了。	2
kau punya kebenarannya? berkendara dari peru ke polinesia dengan rakit dari pohon [illegible]	搭轻木木筏从秘鲁漂到玻里尼西亚！	3	从秘鲁开车到玻里尼西亚，从轻木木岛上开木筏！	2	你明白了吗？ 从一个巴尔萨树驱车从秘鲁到波利尼西亚！	2	你知道真相了吗？从秘鲁到波利尼西亚, 用一筏轻的木筏开车!	2
lakukan apa yang anda anggap terbaik, sayang.	做你认为最好的事，亲爱的。	3	尽你所能，亲爱的。	1	做你认为最好的，亲爱的。	4	做你认为最好的, 爱。	2
pintu belakang ke neraka.	进入地狱的后门。	4	后门进入地狱。	1	后门到地狱。	1	回到地狱的大门。	3
setelah bertengkar dengan dua wanita itu, aku menelepon penyelamatku.	但是那两妞都跑走了，弄的我浑身骚动。	1	和两个女人吵架之后，我就呼唤了我的救世主。	3	在和这两个女人打架之后，我打电话给我救援。	3	和两个女人吵架后, 我给 penyelamatku 打了电话。	2
stasiun listrik siap, pak!	通电站准备好了！	2	电力站好了，长官！	3	电站准备好了，先生！	4	电车站准备好了, 先生!	1
sudah lama sejak yang terakhir menung-ganginya selama 8 detik.	很少有人可以骑超过八秒。	4	已经很久没有人骑过它 8 秒了。	4	自从最后一次骑了 8 秒钟以来已经很久了。	3	从上一次坐了 8 秒后已经很久了。	3
tidak, tidak, sayang, tidak ada jalan.	不，不，不，没门！	1	不，亲爱的，没有路。	1	不，不，亲爱的，没办法。	4	不, 不, 亲爱的, 没有办法。	4
sekarang, giliranmu membereskan semuan-ya.	现在，轮你来处理后事了。	1	现在轮到你收拾残局了。	4	现在，轮到你照顾好一切。	1	现在轮到你了, 保重一切。	1
katakan, kapan?	说吧，什么时间？	3	告诉我，什么时候？	4	说什么时候？	1	说什么时候？	1
berpotensi, punya potensi untuk mengubah dunia.	能力，有能力就能改变天下。	2	潜在的潜力，能改变世界。	3	潜在的，有可能改变世界。	3	有可能改变世界。	3
apakah aku pernah membuat kita tersesat?	我有没有害我们迷路过？	4	我有没有让我们迷路？	3	我曾让我们误入歧途吗？	4	如果我让我们迷路了呢？	1
pak, kita sudah tiba dikoordinat yang telah kita hitung. tidak ada apa2 di sini, apa [illegible]	长官，我们已经到达了你预定的坐标，这里什么都没有？	1	长官，我们已经到达预定座标了，这里什么都没有？	2	先生，我们已经到达了我们计算的坐标。什么都没有，你的命令是什么？	3	我们已经计算过 dikoordinat 了这里没有 apa2, 是什么吩咐你的？	1
aku akan menemukannya, dan kau akan membantuku menemukannya.	我要找到他，你要帮我找到他。	1	我会找到她，你会帮我找到她。	2	我会找到的，你会帮助我找到它。	3	我会找到的, 你会帮我找到的。	4