哈尔滨工业大学 CWMT2015 机器翻译评测报告

陈科海,周鑫鹏,李晓倩,朱晓宁,曹海龙,赵铁军
哈尔滨工业大学计算机科学与技术学院机器智能与翻译研究室
{khchen,xpzhou,xqli,xnzhu,hlcao,tjzhao}@mitlab.hit.edu.cn

摘要：本文详细介绍了哈尔滨工业大学机器智能与翻译研究室(HIT-MITLAB)参加 2015 年全国机器翻译 (CWMT2015)评测任务的情况。本次评测我们参加汉英、英汉、双盲三个测评任务，每个测评项目提交了一个主系统结果和多个对比结果。我们的系统采用基于短语的翻译模型和基于层次短语的翻译模型，在受限方式下进行模型学习和训练，即训练数据完全来自于评测方提供的训练数据。相对于传统的翻译模型，我们有以下几方面的改进：词对齐融合、神经网络语言模型、基于操作序列的调序模型。系统性能获得了显著的提升，在评测中表现优秀。

HIT-MITLAB Technical Report for the 2015 China Workshop on Machine Translation

Kehai Chen,Xinpeng Zhou,Xiaoqian Li,Xiaoning Zhu,Hailong Cao,Tiejun Zhao
Harbin Institute of Technology, Machine Intelligence and translation Laboratory

Abstract: *This paper describes the details of machine translation and evaluation task for HIT-MITLAB's joining in the China workshop on Machine Translation in 2015(CWMT2015).In this task, we take part in 3 translation sub-tasks: Chinese-to-English and English-to-Chinese News, Double Blind Assessment. We take Phrase-Based Translation Model and Hierarchical phrase-based model to train model with restricted corpus which all come from the data provided the sponsor. Compared with traditional translation model, the improvement of our mode follows: combine with alignment, neural network language model, and operation sequence reordering model. With all these enhancements, our system produced significantly improved translation results and achieved outstanding result in the evaluation.*

1 引言

哈尔滨工业大学机器智能与翻译研究室从事自然语言处理与机器翻译的研究工作已有 20 多年的研究历史。在 2015 年中国机器翻译研讨会（CWMT2015）机器翻译评测中，本实验室共参加了三个项目的评测项目，分别是：英汉新闻领域、汉英新闻领域、双盲评测。我们分别在英汉新闻领域和双盲评测项目中取得总排名第一。本文将主要介绍我们所采用的统计机器翻译系统、新采用的技术以及系统在各个人物上的性能表现。最后我们将对结果进行简单分析。

2 系统描述

本次翻译测评中，我们重现了的是基于短语的翻译模型和基于层次短语的翻译模型，将其作为基线系统，然后加入当前的主流新技术和方法作为我们的参评系统。

基于短语的翻译模型(Koehn et al.,2003)，最小翻译单元为连续的词序列所组成的形式化短语。该系统将给定的源语言句子切分为连续词串（形式化短语），然后对每个形式化短语进行翻译，最后进行调序并输出翻译结果。除了传统的翻译特征正反向翻译概率、正反向词汇化概率、词汇化调序特征、词和短语惩罚特征、扭曲特征和语言模型之外，本次评测中我们引入了神经网络语言模型以及基于操作序列的调序模型。

在基于层次短语的翻译模型中(Chiang 2005)，它基于同步上下文无关文法。这些层次规则可以从双语对齐中进行抽取，将规则在句子中所覆盖的单词数限制为10个，且抽取的规则中含非终结符的个数不超过5个。同时抽取出来的规则为二元规则，即每条规则包含不多于两个非终结符。翻译除了基本的层次翻译特征之外，我们引入单语神经网络语言模型和双语神经网络语言模型。系统解码采用cube-pruning的剪枝策略。

3 新技术和方法

相比于CWMT2013中所使用的系统，我们在此次评测中尝试了当前机器翻译中的一些有效的新技术和方法，主要包括词对齐融合、神经网络语言模型和基于操作序列的调序模型等。

3.1 神经网络语言模型

机器翻译系统严重的依赖语言模型来确保目标语言句子输出的流畅性。传统的语言模型是对单词离散描述的ngram模型，这种模型往往容易受到数据稀疏性的影响，因为ngram模型无法利用单词之间的相似性信息。因此根据[Beingio等 2003],我们使用一个如图1所展示的前向神经网络来训练获取单词的分布式表示，进而构建前神经网络语言模型。

在图1中，网络的输入是上下文$\mathbf{u}$中的单词的序列，输出则是每一个单词的概率$p(w|\mathbf{u})$,按照如下过程计算所得：

隐含层是由rectified线性单元[Nair等，2010]组成，其激活函数为$\phi(x)=\max(0,x)$。第一个隐含层h_1的输出为：

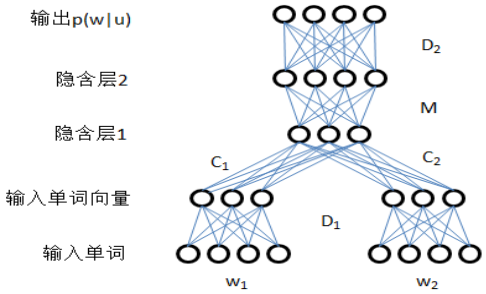

图1 神经网络语言模型(Beingio等,2003)

$$h_1=\phi\left(\sum_{j=1}^{n-1}C_jD_1u_j\right)$$

这里D_1就是输入单词的分布式表示，C_j入单词的则是**u**中的每个单词的上下文矩阵。第二个隐含层h_2的输出为：

$$h_2=\phi\left(Mh_1\right)$$

这里M是两个隐含层之间的关联加权矩阵。最后输出层则为：

$$p(w|u)=\exp(D_2h_2+b)$$

D_2在为输出单词的分布式表示，b是词汇表中每个单词的偏好向量。

3.2 基于操作序列的调序模型

在基于短语的翻译模型中，经常需要学习局部依赖，比如调序、习惯用语搭配、删除和插入等，然而忽视了被翻译或调序的短语经常是相互独立的，同时独立于短语之外的上下文信息。为此基于马尔科夫链的最小单元序列[Durrani等，2011]被提出用于改善这一现象。

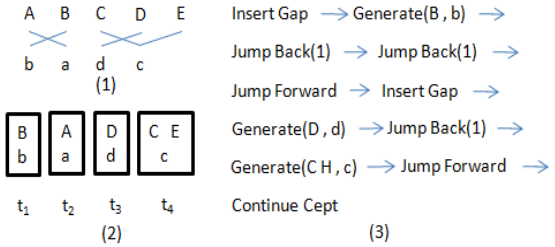

图2:(1)对齐;(2)MTU序列;(3)操作序列(Durrani等,2013)

所谓最小翻译单元就是不能再被分解的翻译单元，例如词、字母等。图2展示的就是一个操作序列过程，给定一个双语句子串(F,E)和对齐A，首先从中识别最小翻译短语。

基于操作序列的调序模型通过一个有被对齐的句子串所产生的操作序列来描述一个双语句对和其对应的对齐。一个操作要么产生对应的源端到目标端的翻译对，要么通过插入间隙并向前或向后跳跃来执行调序操作。这个过程正如图2(1)中所示，被转换为图2(3)中的操作序列。在携带最小翻译单元和调序信息的操作序列上$(o_1,o_2,...,o_J)$的马尔可夫模型形式化为

$$p_{osm}(F,E,A)=\prod_{j=1}^{J}p(o_j\,|\,o_{j-n+1},...,o_{j-1})$$

4 关键步骤

4.1 数据使用

在本次CWMT2015评测中，我们所使用的语料数据完全来自于主办方提供的数据，此外汉语端包含搜狗单语语料以及英语端包含路透社单语语料，分别用于构建对应的目标语言模型，具体数据如表1、2所示。

翻译评测项目	训练集句对数		开发集句对数
	CWMT2013官方句对数	预处理后句对数	
英汉新闻	7278679	6054231	1000
汉英新闻	7278679	6054231	1006
双盲测评	2000000	1967341	1116

表1: 训练翻译规则的语料数据信息

翻译评测项目	语料名称
英汉新闻	目标单语语料+搜狗中文单语语料
汉英新闻	目标单语语料+路透社英文单语语料
双盲测评	目标单语语料

表2: 训练语言模型的语料数据信息

4.2 语料预处理

首先对中文使用stanford分词工具进行分词，然后采用实验室自主开发的语料过滤工具Jessica和基于困惑度的训练语料分析工具mozzie[Liang,2013]对这三个方向的训练语料进行过滤分析。

中文数据（包含训练语料的中文部分、中文语言模型数目）进行全角转半角、特殊标点符号的转义。

英文数据的预处理操作包括大写转小写和标点符号的分词处理，此外就是进行tokenizer切分处理，truecase等。

4.3 语料对齐

我们认为提供更多的候选对齐用于提取翻译规则能够改善翻译质量。我们利用2种开源词对齐工具giza++和bekerley aligner来获取所有方向上的语料对齐文件。我们在评测中分别从两个对齐文件中提取翻译规则，然后融合在一起构成我们系统的翻译规则[Tu等,2012]。为了保证规则的高效性，我们采用相对熵对规则表进行适当的过滤。

4.4 语言模型构建

由于搜狗语言模型训练数据中存在着一些噪音，直接使用对翻译效果的提升不明显，因为调参后该语言模型对应的权重很小甚至为负数。我们适应一种基于困惑度的方法来改变搜狗语言模型数据的分布。其主要思想就是增大和训练数据目标语言相近的句子的分布，同时降低和目标语言端相差较远的句子的分布。

语言模型构建分别采用传统的ngram和神经网络语言模型。一方面，利用srilm[Stolcke，2002]和额外的大规模单语语料（搜狗和路透社语料）构建对应的ngram语言模型，进而兼顾各自数据的特点；另一方面，利用NPLM[Vaswani等，2013]使用目标语料和对应的单语语料构建神经网络语言模型。各个项目使用的语言模型的情况为：汉英新闻和英汉新闻包含两个5-gram语言模型和一个神经网络语言模型，双盲评测包含一个5-gram语言模型和一个神经网络语言模型。

4.5 调序模型训练

在本次评测中，我们在经典基于层次短语的调序模型基础之上，利用操作序列模型[Durrani等，2013]来进一步辅助翻译中的调序操作。基于层次短语的调序模型在机器翻译领域被大家所认可，具有较好的调序能力，但其是在短语上进行调序的，忽视了最小翻译单元之间的依赖以及短语之外的调序约束。为此我们引入基于操作序列的调序模型，其实构建在最小翻译单元上的调序操作，比如最小翻译单元为词，能够有效的辅助基于层次的调序模型，从而使得翻译系统具有更加有效的调序能力。

4.7 重排序 nbest 翻译

我们利用翻译系统所输出的nbest译文，在此基础之上利用语言模型对齐进行重排序操作。这里用于重排序的语言模型，我们选取传统的ngram语言模型和当前流行的递归神经网络语言模型[Mikolov,2012]线性插值来分配不同的权值，对nbest译文进行重排序，进而输出更加符合目标语言语序的译文。

4.6 后处理

在英汉新闻翻译中，我们编写了一个人名、地名的音字转换程序，较好的转换了译文中的未识别的人名和地名。目标语言是中文的译文，去掉词与词之间的空格，目标语言是英文的译文，然后对所有译文中的相关英文缩写等专有名词进行了大小写转换，中文的翻译结果在提交前将半角的冒号、分号、双引号替换成了全角字符。最后清除掉所有译文中的未识别词。

5 实验

本节将按照汉英新闻、英汉新闻、双盲项目的顺序分别介绍相关的实验情况。最终的参数权值通过mert和mira来获得，并对其进行线性插值，来得到最终的参数权重用于解码翻译。本节中的评分皆采用nist-bleu，汉英和英汉皆是在大小写敏感下进行。

5.1 实验环境

硬件：
CPU 品牌：Intel
CPU 型号：Xeon E5-2670
CPU 主频：2.60Ghz
CPU 数量：24
内存容量：256GB
软件：
操作系统类型：Linux
操作系统版本：Centos Release6.0

5.2 实验结果及分析

5.2.1 汉英/汉英新闻项目

汉英和英汉两个子项目的开发集和测试集皆采用官方提供的数据：

评测项目	开发集	测试集
汉英	1006	1100
英汉	1000	1100

表3开发集与测试集数据

英汉新闻和汉英新闻项目采用mert[Och,2003]和mira[Cherry等,2012]进行参数调优。

本次汉英、英汉新闻评测项目上，我们按照CWMT2015官方的受限领域评测规则进行训练，在训练过程中使用最大短语长度为10，并对短语分数计算进行了Kneser-Ney平滑，汉英评测项目上使用了训练集目标语言数据和路透社单语数据分别训练了两个5元语言模型，而在英汉评测项目上使用训练集目标语言数据和搜狗单

语数据分别训练了两个5元语言模型。其次分别使用训练集目标端语言训练汉英和英汉的神经网络语言模型。最后利用双语训练集训练基于操作序列的调序模型来辅助基于层次短语的调序模型。ce\ec-primary和ce\ce-contrast的不同在于前者基于短语的翻译模型，而后者是基于层次短语的翻译模型。汉英评测结果如表4所示，英汉评测结果如表5所示，我们的英汉评测名列测评第一。

5.2.2 双盲测评项目

在CWMT2015测评项目中，新引入了双盲评测项目，也即对双语段进行了特定形式的编码，使其变成一种密文，然后对齐进行模型构建、开发集调优和测试集评测。

在双盲测评中，我们所使用的改进方式主要包括：基于相对熵的语料过滤、词对齐融合、翻译规则平滑、单语神经网络语言模型、双语神经网络语言模型、基于操作序列的调序模型。

在训练过程中，设置最大短语长度为10，采用Kneser-Ney方式对短语进行平滑打分，参数调优过程中将kbest设置为200。在对比系统中，我们通过设置添加设置基于双语的神经网络语言模型以及改变参数调优方法来获得译文结果。主系统db-primary与对比系统db-contrast-b都是基于短语的翻译模型，其区别在于后者为添加基于双语的神经网络语言模型。其次db-contrast-a和db-contrast-c是基于层次短语的翻译模型，二者的区别在于前者使用mira进行参数调优，而后者使用mert进行参数调优。双盲测评结果如表6所示，并且db-contrast-b的评测结果名列测评总体排名第一，这也充分说明了我们的系统具有良好的模型性能。

5.2.3

我们主要阐述在评测中引入神经网络语言模型和机遇操作序列的调序模型在单位。如表7所示，baseline的基本配置为：基于短语的翻译模型、有5-gram语言模型、基于层次短语的调序模型。OSRM表示基于操作序列的调序模型；NNLM表示神经网络语言模型；NNJLM表示双端神经网络语言模型。表7中所列举的是英汉新闻领域在开发集上和cwmt2013测试集上对应的BLEU-4的测试结果。从表7中可以看出，基于神经网络的语言模型和基于操作序列的调序模型皆明显的提升了系统性能，同事当期叠加使用时，起对应的性能改进也一定程度上得到了叠加。同时当我们叠加使用单语神经网络语言模型和双语联合神经语言模型时，性能提升不明显，起原因应该是这两种语言模型都是基于神经网络来训练的，起性能的改进是相互叠加的。

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
ce-primary	0.1953	0.2058	6.8739	0.6738	0.7559	0.5539	0.2510	0.1854	0.7360
ce-contrast	0.1904	0.2005	6.7534	0.6656	0.7597	0.5663	0.2452	0.1831	0.7474

表4: 汉英新闻项目评测结果

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
ec-primary	0.3773	0.3801	0.3148	10.4170	10.4322	0.8565	0.6625	0.3373	0.3505
ec-contrast	0.3706	0.3734	0.3076	10.3363	10.3505	0.8554	0.6673	0.3400	0.3452

表5: 英汉新闻项目评测结果

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
db-primary	0.1934	0.2029	6.6491	0.6611	0.7648	0.5699	0.2286	0.2251	0.7042
db-contrast-a	0.1965	0.2035	6.5375	0.6675	0.7539	0.5621	0.2183	0.2276	0.7210
db-contrast-b	0.1977	0.2051	6.6193	0.6656	0.7549	0.5604	0.2214	0.2281	0.7150
db-contrast-c	0.1943	0.2049	6.6926	0.6634	0.7699	0.5693	0.2331	0.2252	0.7010

表6: 双盲项目评测结果

总结

在本次评测中我们使用了基于短语的翻译模型和基于层次短语的翻译模型。在 3 个评测项目中，我们的系统性能表现稳定，尤其在上次测评的基础之上，引入了神经网络语言模型和基于操作序列的调序模型。其次，我们自主开发的语料预处理模块 Jessica 和 mozzie 能够较好过滤掉训练语料中的噪音，从而提高的训练语料的质量，为后续模型的建立奠定了良好的基础。本次评测与前次评测相比，我们的翻译结果有了较大的提升。最后通过此次 CWMT2015 评测，提供了一个与国内外同行的进行系统对比的机会，为今后开展统计机器翻译的研究工作极大的开拓了思路。

	英汉开发集	英汉测试集（cwmt2013）
Baseline	0.2950	0.2545
Baseline+OSRM	0.3087	0.2660
Baseline+NNLM	0.3034	0.2622
Baseline+OSRM+NNLM	0.3214	0.2746
Baseline+OSRM+NNLM +NNJLM	0.3254	0.2766

表 7: 神经网络语言模型和操作序列模型在英汉方向上的测试结果

References

P.Koehn,F.Och,and D.Marcu. Statistical Phrase-based Translation. In Proc. of NAACL, 2003.

David Chiang. A hierarchical Phrase-based Model for Statistical Machine Translation. In Proc. of ACL, 2005.

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model.Journal of Machine Learning Rsearch, 2003.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted Boltzmann machines. In Proceedings of ICML, 2010.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. A Joint Sequence Translation Model with Integrated Reordering. In Proc. of ACL, 2011.

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? In Proc. of ACL, 2013.

Zhaopeng Tu, Yang Liu, Yi f an He, Jose f van Genabi th,Qun Liu, Shouxun Lin. Combining Multiple Alignments to Improve Machine Translation. In Proc. of Coling, 2012.

A. Stolcke. SRILM - An extensible language modeling toolkit. In Proc. of ICSLP, 2002.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. Decoding with large-scale neural language models improves translation. In Proc. EMNLP, 2013.

Mikolov Tomáš.Statistical Language Models based on Neural Networks. PhD thesis, Brno University of Technology, 2012.

Franz Josef Och. Minimum Error Rate Training In Statistical Machine Translation. In Pro. of ACL, 2003.

Colin Cherry and George Foster. Batch Tuning Strategies for Statistical Machine Translation. In Proc. of NAACL, 2012.

Huashen Liang. Study on Several Key Problems in the Training Process of Phrase-Based Statistical Machine Translation.Statistical Language Models based on Neural Networks. PhD thesis, Harbin Institute of Technology, 2013.