第十一届机器翻译研讨会厦门大学技术报告

谭知行[1]，胡金铭[1]，史晓东[1]，陈毅东[1]
（1.厦门大学 智能科学与技术系，福建 厦门 361005）

摘要：本文主要介绍了厦门大学自然语言处理研究室（XMU-NLPLAB）参加第十一届全国机器翻译研讨会（CWMT2015）翻译评测任务的情况。在本次评测中，XMU-NLPLAB 参加了 6 个评测项目中的 3 个子项—汉英新闻领域机器翻译、藏汉政府文献机器翻译以及双盲机器翻译。报告内容主要阐述本次参评系统的实现框架、模型以及它们在评测数据上的性能表现，同时对各翻译结果进行了比较和分析。

关键词：机器翻译；层次短语模型；神经网络

XMU-NLPLAB Technical Report for CWMT2015

Zhixing Tan[1], Jinming Hu[1], Xiaodong shi[1], Yidong chen[1]
(1.Cognitive Department, Xiamen University, Xiamen, Fujian 361005, China)

Abstract: This paper present an overall introduction in the translation evaluation task of the 11th China Workshop on Machine Translation (CWMT2015), submitted by Xiamen University Natural Language Processing Lab (XMU-NLPLAB). In this evaluation, XMU-NLPLAB takes part in 3 translation sub-tasks of the 6 evaluation tasks, which involve the domain of Chinese-to-English news translation, Tibetan-to-Chinese government document translation and double-blind translation. The report mainly describes the implementation framework, model and the performance in the evaluation data set of the evaluated translation system. In addition, all translation results are compared and analyzed.

Key words: Machine Translation；Hierarchical Phrase-based Model；Neural Network

1 引言

CWMT2015 一共包含 6 个评测方向：汉英新闻领域机器翻译（CE）、英汉新闻领域机器翻译（EC）、蒙汉日常用语机器翻译（MC）、藏汉政府文献机器翻译（TC）、维汉新闻领域机器翻译（UC）以及双盲评测机器翻译（DB）。厦门大学自然语言处理研究室在本届机器翻译评测中参加了三个项目：英汉新闻领域机器翻译、藏汉政府文献机器翻译以及双盲评测机器翻译。

本次评测所提交的系统使用近年来比较成功的神经网络模型作为核心。系统的解码器以及神经网络模型均由实验室内部开发，另外使用了开源系统 GIZA++、Moses 以及 SRILM。

本文将主要介绍我们采用的统计机器翻译系统框架，主要技术以及在 3 个评测项目中的性能表现。最后对本次评测的结果进行分析。

2 系统

本次机器翻译评测中我们采用实验室内实现的基于层次短语的统计机器翻译系统（Chiang, 2007）作为基本系统，在此基础实现了神经网络模型 NNJM（Neural Network Joint Model）与 NNLTM（Neural Network Lexical Translation Model）以及其他特征。

2.1 系统框架

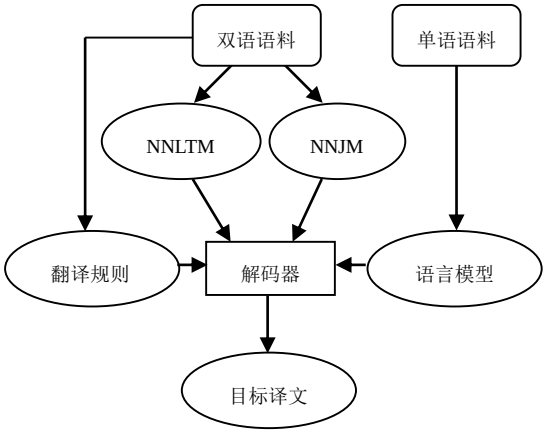

图 1. 系统框架

对于提供的双语语料，我们进行训练翻译规则、NNJM 模型以及 NNLTM 模型。对于单语语料，我们进行训练前向语言模型以

及后向语言模型。所有模型均作为对数线性模型的一个特征融入到解码器中，从而通过翻译源端句子得到目标译文。整个系统的框架如图 1 所示。

2.2 层次短语模型

层次短语模型使用同步上下文无关文法作为翻译的基本单元，形式为：

$$X \rightarrow <\gamma, \alpha, \sim>$$

其中 X 为非终端符，γ 和 α 分别为源端与目标端终端与非终端符组成的串，~为 γ 和 α 中非终端符的对应关系。

本次评测中使用实验室开发的 infdecoder 作为层次短语模型的解码器，训练与调参部分使用开源系统 Moses。

infdecoder 支持多个语言模型与翻译模型的使用。它实现了 CYK+ 算法（J.-C.Chappelier and M. Rajman, 1998），不限制同步上下文文法中非终端符号的个数。但在实验中我们限制非终端符号的数目为 2，与 Chiang 文献中相同。与 Chiang 的文献中不同的是，infdecoder 在 cube pruning 时以翻译候选弹出的个数作为终止条件，而非 Chiang 文献中描述的与最优候选相差的阈值 ε。infdecoder 的性能与开源系统 Moses 中的 moses_chart 解码器性能相当。

2.3 语言模型

本次评测中，我们的语言模型仍使用传统的 n-gram 语言模型。除了普通的前向语言模型，我们也尝试使用了后向语言模型。假设译文为：

$$e = e_1 e_2 \cdots e_{m-1} e_m$$

后向语言模型按如下方式计算概率：

$$P(e) = P(e_m)P(e_{m-1}|e_m)\cdots P(e_{1|}e_n \cdots e_2)$$

后向语言模型通过反转语料中的句子进行实现，同时在解码器中进行特殊的处理。

2.4 神经网络模型

本次评测中，我们的系统在层次短语模型的基础上，添加了神经网络联合模型 NNJM 以及神经网络词汇化翻译模型 NNLTM（Devlin et. al, 2014）。

对于给定的源端句子 S 以及目标端句子 T，假设源端上下文长度为 m，目标端上下文长度为 n，神经网络联合模型根据如下公式计算条件概率 $P(T|S)$：

$$P(T|S) = \prod_{i=1}^{|T|} P(t_i|t_{i-1}, \cdots, t_{i-n}, S_i)$$

S_i 由 t_i 与源端词的附属 a_i 决定，为：

$$S_i = (s_{\frac{m-1}{2}}, \cdots, s_{a_i}, \cdots, s_{\frac{m+1}{2}})$$

神经网络联合模型可以视为一个 m+n+1 元的双语语言模型。

神经网络词汇化翻译模型则计算如下概率：

$$\prod_{i=1}^{|T|} P(t_{s_i}|s_i, s_{i-1}, s_{i+1}, \cdots)$$

其中t_{s_i}是源端词s_i对齐对应的词，若对齐为空则t_{s_i}为 NULL，若对齐一对多则将多个词拼接成为一个新词。

我们按照 Devlin 的论文实现了 NNJM 与 NNLTM 的所有模型，采用 GPU 进行训练，使用 CPU 进行解码。NNJM 有 4 个变种，NNLTM 有 2 个变种，我们的系统中只使用源到目标、自左到右（S2T/L2R）的 NNJM 以及源到目标（S2T）的 NNLTM。

3 数据处理

3.1 语料预处理

我们对所有语料（双语训练语料，开发集，测试集，语言模型训练语料）进行了一定的预处理，具体预处理步骤如下：

汉英：

- 行尾转换；
- 全角转半角
- 英文部分的标点处理
- 转义字符的处理
- 乱码过滤
- 部分已分词语料的还原
- 语言模型训练语料的处理
- 对中文语言模型训练语料（搜狗）进行分句
- 分词
- 领域分类
- truecase

藏汉：

- 行尾转换
- 全角转半角
- 转义字符的处理
- 分词

双盲语料未经过任何处理。

3.2 分词与分句

分句工具使用实验室开发的 MYLINE 系统。

分词方面，在汉语上使用实验室开发的汉语分词工具 segtag；在英语上，使用 moses 中的 token 脚本工具；在藏语上，使用实验室开发的藏语分词工具 ts。

3.3 领域分类

我们使用开源工具 tmsvm 对语料进行新闻类和非新闻类的区分工作。

在模型预测时，我们做了人工判别工作并细心选择阈值，在最终的文本分类结果文件中随机抽样并进行新闻类的人工判断，预测结果的准确率在 90%以上。在不使用 NNJM 模型时，可以发现对语言模型和翻译模型的分类可以有效提升翻译质量。但在使用 NNJM 模型时，在开发集上未能观察到提升，结果在表 4 中列出。

3.4 truecase 与词对齐

我们使用 Moses 中的 truecase 脚本。

词对齐方面，我们采用 mgiza 工具。双语语料完成双向词对齐后，采用启发规则合并两个方向的词对齐结果，作为最终的词对齐结果。

3.5 语言模型

语言模型使用 SRILM(Stolcke et. al. 2002)进行训练。

对于汉英新闻，以英文路透社语料和该项目双语训练语料的目标语言作为一个集合使用了分类工具进行新闻/非新闻分类，新闻类语料训练得到的 5 元模型，记为 CE-LM1，将非新闻语类语料训练得到的模型，记为 CE-LM2。全集语料训练得到的前向语言模型和后向语言模型记为 CE-LM 和 CE-BLM。

对于藏汉政府文献项目，使用中文搜狗语料作为训练集，训练得到了 5 元的前向和后向语言模型，我们称之为 TC-LM 和 TC-BLM。

对于双盲项目，我们同样训练 5 元的前向模型和后向模型，称为 DB-LM 和 DB-BLM。

3.6 语料的使用

在训练集上，我们语料的使用情况如下：

评测项目	训练集(句对)
汉英新闻	7718073
藏汉政府文献	125998
双盲	2000000

表 1: 语料的使用情况

语言模型的大小为：

评测项目	语言模型大小
汉英新闻	4.4GB(LM, BLM) +4.2GB(LM1)+ 356MB(LM2)
藏汉政府文献	9.2GB(LM,BLM)
双盲	2.6GB(LM,BLM)

表 2: 语言模型的大小

4 实验

4.1 运行环境

系统的运行环境如下：

CPU：Intel Xeon 12 核@2.93GHz

内存：96GB

操作系统：Unbuntu Server 12.04

4.2 训练

4.2.1 对齐

使用 mgiza 工具来训练词对齐。

4.2.2 翻译规则

利用 moses 训练获得层次短语规则表，源端短语的长度限制为 5。在部分对比系统中使用源端短语长度为 7 的规则表。

4.2.3 语言模型

我们使用 SRILM 训练 5 元语言模型，将文本反转后训练 5 元后向语言模型。

4.2.4 神经网络模型

我们使用与得到翻译规则相同的语料来训练 NNJM 以及 NNLTM 模型。训练使用 Nvidia GTX 750 GPU 完成。

NNJM 以及 NNLTM 采用前向神经网络进行实现。在所有三个任务中，我们使用同样的网络参数。我们只使用 1 个大小为 512 的隐层，使用 tanh 作为激活函数。使用 1 个隐层可以通过预计算来加快解码中的速度。源端、目标端以及输出端的词表数目分别为 16000，16000 以及 32000，词向量大小为 192。NNJM 的源端历史为 11，目标端历史为 3。

NNLTM 的源端窗口大小为 14。

NNJM 以及 NNLTM 的训练使用后向传播算法。训练目标使用 Self-Normalization，控制因子为 0.1，Batch 大小设为 128，学习率设为 0.01，我们不使用任何正规化。神经网络每次权重的更新设为区间[-0.01, 0.01]。

对于不同的翻译项目，训练的最大轮数如下表所示：

项目	轮数
汉英	3
藏汉	40
双盲	10

表 3. NNJM 与 NNLTM 训练轮数

4.3 调参与解码

提交的主系统在 CE、TC 和 DB 评测中均使用以下特征：

1. 前向语言模型（SRILM）
2. 后向语言模型（SRILM）
3. 神经网络联合模型（S2T/L2R NNJM）
4. 神经网络词汇化翻译模型（S2T NNLTM）
5. 翻译概率
6. 词汇化概率
7. 反向翻译概率
8. 反向词汇化翻译概率
9. 规则惩罚
10. glue 规则惩罚
11. 词惩罚

我们使用 Moses 实现的 mert 进行调参，调参的标准采用 BLEU。在调参中，我们使用的参数均为默认参数。

在解码中，我们将解码器的 cube pruning 的弹出数目由默认的 1000 改为 10000，其余参数不变。

4.4 模型选择与后处理

我们选择在开发集中 BLEU 最高的系统作为主系统。

在所有项目的后处理中，均将未登录词丢弃。在藏汉翻译中，将句末的非句号改为句号。

4.5 评测结果

在下面表格中，开发集中的结果使用 moses 中 multi-bleu 结果进行汇报，测试集中的结果采用 BLEU-SBP 进行汇报。开发集实验结果未经过后处理。

4.5.1 汉英新闻领域翻译结果

汉英新闻中我们的主系统使用前面所述的全部 11 个特征。3 个对比系统中均没有使用 NNLTM 模型。其中对比系统中 contrast-a 系统使用短语长度为 7 的规则表（记为 HPB-7），并且使用了经过领域分类的语言模型。contrast-c 系统使用新闻领域数据训练的规则表（记为 HPB2），和新闻领域的语言模型以及 NNJM 模型。

汉英新闻领域系统的运行结果以及配置情况由表 4 进行描述。

4.5.2 藏汉政府文献翻译结果

藏汉政府文献翻译系统中，我们的所有系统均采用所有的 11 个特征，区别在于使用的参数有所不同。由于观察到结果的未登录词较多，我们的主系统采用 2 种不同分词结果得到的 2 个规则表，将两个规则表进行合并，从而降低未登录词的数目。

藏汉政府文献系统的运行结果以及配置情况由表 5 进行描述。

4.5.3 双盲翻译结果

双盲翻译中，我们所有系统均采用完整的 11 个特征，区别在于使用的参数有所不同。主系统采用训练 6 轮的 NNLTM 模型，其余系统采用训练 10 轮的 NNLTM 模型。

双盲机器翻译系统的运行结果以及配置情况由表 6 进行描述。

4.6 实验结果分析

通过实验结果可以发现，在汉英新闻翻译中，NNJM 以及 NNLTM 效果明显，共带来约 3.0BLEU 的提升。在藏汉政府文献以及双盲翻译中，NNJM 与 NNLTM 效果没有汉英翻译中的明显，但是也获得了约 1.3BLEU 的提升。

我们对实验结果的分析如下：

1. 汉英双语数据较大（约 2 倍于双盲数据，50 倍于藏汉数据），因此 NNJM 和 NNLTM 能够有较好的表现
2. 在不同语言对中，NNJM 以及 NNLTM 带来的提升不同，这个现象在 Devlin 的论文中已经有描述。
3. 在汉英新闻领域的翻译中可以看出，只使用领域分类的语言模型时，在开发集中是有明显提升。但同时加入 NNJM 时

使用领域分类的模型(contrast-a 和 c)和未使用的模型相比(contrast-b)，反而分数降低。这有可能是因为分类的语言模型作用被神经网络模型覆盖掉。

系统	配置	开发	测试
baseline1	基本层次短语模型，使用语言模型 CE-LM	30.34	-
baseline2	使用后向语言模型（CE-LM + BLM）	30.50	-
domain1	加入领域分类信息（CE-LM1）	30.59	
domain2	加入领域分类信息（CE-LM1 + CE-LM2）	31.44	
primary-a	使用全部 11 个特征（CE-LM + NNJM + BLM + NNLTM）	**33.42**	**23.38**
contrast-a	使用领域分类（HPB-7 + CE-LM1 +CE-LM2 + BLM + NNJM）	32.27	22.31
contrast-b	10 个特征，不使用 NNLTM（CE-LM + BLM + NNJM）	32.48	22.66
contrast-c	使用领域分类（HPB2 + CE-LM1 + BLM + NNJM）	32.43	22.64

表 4. 汉英新闻领域系统情况

系统	配置	开发	测试
baseline	基本层次短语模型	56.40	-
primary-a	全部 11 个特征（合并 2 种分词结果）	**57.69**	**88.01**
contrast-a	全部 11 个特征	57.43	87.14
contrast-b	全部 11 个特征	57.43	87.27
contrast-c	全部 11 个特征	57.69	87.48

表 5. 藏汉政府文献领域系统情况

系统	配置	开发	测试
baseline	基本层次短语模型	32.01	-
primary-a	使用全部 11 个特征（NNLTM 只训练 6 轮）	**33.32**	19.73
contrast-a	使用全部 11 个特征	33.31	19.69
contrast-b	使用全部 11 个特征	33.31	**19.75**

表 6. 双盲机器翻译系统情况

5 总结

在本次评测中，我们主要使用神经网络联合模型和神经网络词汇化翻译模型来提升我们的系统。由于时间有限，只使用了 S2T/L2R 的 NNJM 以及 S2T 的 NNLTM，并没有将系统的潜力完全发挥，也没有尝试神经网络语言模型。从评测结果来看，我们取得了不错的成绩。采用神经网络模型能够使翻译质量带来较大的提升，这也与我们的预期相一致。

尽管我们在本次评测中取得了较好的成绩，但我们也意识到我们系统仍有较大的提升空间。近两年神经网络研究进展很大，我们没有进一步采用 RNNLM 等神经网络工具，同时我们的系统在本质上仍使用较为古老的层次短语模型。我们距国内外顶尖研究机构仍然有很大的差距。我们希望能够在本次会议中同国内外同行进行交流和学习，进一步促进我们对机器翻译的理解与研究。

参考文献

[1] David Chiang. 2007. Hierarchical phrase-based translation [J]. *Computational Linguistics*.
[2] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz

and J. Makhoul. 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation[C]. *In Proceedings of Association of Computational Linguistics*.

[3] J.-C. Chappelier and M. Rajman. 1998. A generalized CYK algorithm for parsing stochastic CFG[C]. *In Proceedings of the First Workshop on Tabulation in Parsing and Deduction*, page 133-137.

[4] Stolcke, Andreas and others. 2002. SRILM-an exensible language modeling toolkit [J]. *INTERSPEECH*.

[5] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation[C]. *Annual Meeting of the Association for Computational Linguistics*.

[6] Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation [C]. *Proceedings of the 41st Annual Meeting on Association for Com-putational Linguistics.*