南京大学 CWMT2013 机器翻译评测技术报告

The NJU machine translation system for CWMT2013

黄书剑 陈华栋 孙辉丰 朱长峰 周浩 奚宁 赵迎功 戴新宇 陈家骏
软件新技术国家重点实验室（南京大学）
南京大学计算机科学与技术系
{huangsj, chenhd, sunhf, zhucf, zhouh, xin, zhaoyg, daixy, chenjj}
@nlp.nju.edu.cn

摘 要

本文描述了南京大学自然语言处理研究组参加 CWMT2013 机器翻译评测的系统情况。本次评测中，我们参加了汉英新闻、英汉新闻、英汉科技三个项目的评测。每个评测项目提交了一个主系统结果。该系统的基本框架是我们自行实现的基于层次短语的统计机器翻译模型。该系统采用受限方式进行了模型学习和训练，即训练数据完全来自于评测方提供的训练数据。相比于基本的基于层次短语的系统，我们在以下几个方面对系统进行了改进：词对齐融合、翻译表的平滑和增强、语言模型的融合和命名实体翻译。系统性能获得了显著的提升，在评测中表现优秀。

Abstract

This document describes the participation of Natural Language Processing Research Group of Nanjing University in the machine translation evaluation of CWMT2013. We participated three translation tasks: Chinese-to-English News Translation (CE), English-to-Chinese News Translation (EC), English-to-Chinese Sci-Tech Translation (ES). For each task, we submitted one single system result as the primary system, respectively. The submitted systems were trained under constrained training condition, i.e. all the training data come from the organizer. Compared with a standard hierarchical-phrase based translation system, our systems were enhanced in the following aspects: combination of word alignment, smoothing and enhancing of translation table, combination of language models, named entity translation. With all these enhancements, our system produced significantly improved translation results and achieved outstanding result in the evaluation.

1 引言

南京大学自然语言处理研究组从事自然语言处理和机器翻译的研究工作已有二十多年的历史。本研究组设计研发的基于规则的和多策略的日汉机器翻译系统曾得到多轮国家 863 高科技研究发展计划的连续支持，并获得多项省部级科技表彰[1]。本研究组自 2008 年底开始进行统计机器翻译的研究与系统开发。2009 年起，连续参加了三届 CWMT 机器翻译评测，从对开源的 Moses 系统 (Koehn et al., 2003) 的改进到自行研发翻译引擎并进行改进，本研究组的机器翻译系统研究取得了显著的进展，系统效果也逐步进入于评测前列。

本次评测中，我们参加了汉英新闻、英汉新闻、英汉科技三个项目的评测。每个评测项目都提交了一个主系统结果。该系统的基本翻译模型是我们自行实现的基于层次短语的统计机器翻译模型。该系统采用受限方式进行了模型学习和训练，即训练数据完全来自于评测方提供的训练数据。相比于基本的基于层次短语的系统，我们在以下几个方面对系统进行了改进：词对齐融合、翻译表的平滑和增强、语言模型的融合和命名实体翻译。系统性能获得了显著的提升，在评测中表现优秀。

本文的后续部分的安排如下：第二部分概述系统的总体框架以及本次参评过程中我们对基本系统进行的改进，第三部分介绍数据使用的情况，第四部分是实验结果及分析。

2 系统描述

本节按照数据预处理、词对齐、翻译模型训练、语言模型训练、命名实体翻译、解码器、后处理的顺序介绍参评系统的方法和实现情况：

[1]http://nlp.nju.edu.cn/homepage/research.html

2.1 预处理

英文数据（包括训练语料的英文部分、英文语言模型、英文的源语言或目标语言文件）的处理首先进行全角到半角的转化，然后经由一个自行实现的 tokenizer 进行单词切分。与基本的 tokenizer 程序相比，自行实现的 tokenizer 在"-" 和 "'" 周围单词的切分上做了进一步的划分。

中文数据（包括训练语料的中文部分、中文语言模型、中文的源语言或目标语言文件）有两种处理方案，一是使用 ICTCLAS[2]进行中文分词；二是按照 unicode 字符进行逐字分割。

中文数据中的英文部分采用与英文数据相同的处理策略。

2.2 词对齐

经过预处理的平行语料采用 Giza++[3]工具包进行词对齐。受 Xi et al. (2012) 工作的启发，我们思考利用多种粒度的中文单元进行词对齐融合的方法。由于本次评测受限项目不允许使用额外的双语资源，因此，无法使用判别式的方法对不同的词对齐结果进行融合。为此，我们设计了一种基于启发式规则进行词对齐融合的方法。

通过使用中文部分的不同处理方法和调整 Giza++ 的不同参数设置，我们按照如下三组不同的运行配置对训练数据进行了词对齐：

- $1^5h^54^3$ 中文分词 -英文单词
- $1^5h^53^54^5$ 中文分词 -英文单词
- $1^5h^54^3$ 中文分字 -英文单词

其中上标表示对应模型的迭代次数；中文分词、中文分字、英文单词表示用来训练词对齐的双语数据的预处理方式。

对于每一组运行配置我们都输出 IBM Mode1、HMM、IBM Model4 三个模型的训练结果。共计 9 个模型的训练结果。每一个模型的训练结果包含都双向的词对齐。我们使用 grow-diag-final (Koehn et al., 2003) 启发式规则对每个模型的双向词对齐进行合并，从而为每个模型得到一个合并后的词对齐。共计 9 个合并后的词对齐结果。我们对 9 个合并后的词对齐使用扩展的 grow-diag-final 启发式规则进行合并，从而得到最终使用的唯一的词对齐结果。

由于 grow-diag-final 启发式规则是针对双向词对齐合并而设计的，因此在用于合并多个词对齐时其策略需要做一些调整，调整后的合并过程如下所示：

1. 将所有待合并的词对齐的交集作为起始词对齐 I，并集与交集的差作为待增长的集合 C。将 C 中的元素（即词对齐连接）按照模型的投票排序。

2. 循环进行如下步骤直至没有新的词对齐连接可以加入 I 或者 C 为空：

   1）按顺序序遍历 C，判断 C 中的每一个词对齐连接 a 是否符合 grow-diag 的条件，若满足，则将 a 加入 I，并将 a 从 C 中删除。

3. 循环进行如下步骤直至没有新的词对齐连接可以加入 I 或者 C 为空：

   1）按顺序序遍历 C，判断 C 中的每一个词对齐连接 a 是否符合 final 的条件，若满足，则将 a 加入 I，并将 a 从 C 中删除。

2.3 翻译模型训练

翻译模型的训练包括翻译规则的抽取和评分。层次短语翻译系统 (Chiang, 2005) 中使用的翻译规则包括短语和层次短语两种。本次参评系统使用的短语的最大长度是 10，层次短语的中文端最大长度为 3（以分词后的结果为单位），英文端最大长度为 5。

考虑到大量低频翻译规则出现的次数少，直接通过其出现次数进行概率估计可靠性较低，我们采用了 Good Turing(GT) 和 Modified Keneser-Ney(MKN) 两种方法分别对翻译模型的概率估计进行了平滑 (Foster et al., 2006)。

除了正向、反向的短语翻译概率和词汇化翻译概率之外，我们在训练翻译模型时还记录了每个短语的出现次数，用以控制低频短语的概率（Enhanced Low-Frequency，ELF） (Chen et al., 2011)。

本次评测中的翻译规则的抽取和评分是在各项目的全部平行训练数据上进行的。评分完成后，每个源语言端短语或层次短语保留前 100 个候选翻译结果用于解码。为了处理百万句规模的数据，我们设计了一套利用分布式计算模式 Hadoop[4]进行翻译模型训练的方法，在本单位内部的一个分布式计算平台上进行执行。

2.4 语言模型训练

我们按照数据来源、处理方法和训练方向对语言模型进行了不同的训练。每个训练得到的语言模型都作为一个独立的特征加入到统计机器翻译系统中。不同的语言模型训练包括如下方面：

[2]http://ictclas.org/

[3]http://code.google.com/p/giza-pp/

[4]http://hadoop.apache.org/

- 数据来源 对平行训练语料中的目标语言部分和单语语言模型语料分别训练语言模型
- 处理方法 在英汉翻译项目中，对中文分词和中文分字处理的语料分别训练语言模型
- 训练方向 按照自左向右和自右向左的顺序分别训练语言模型，也称为正向、反向语言模型 (Backward N-grams (Xiong et al., 2011))

语言模型的训练使用了自行开发的训练工具，所有的语言模型都采用 5 元语言模型和 MKN 平滑方法。

2.5 命名实体翻译

命名实体属于较为稀疏的知识资源，在训练翻译模型的时候极易与周围的单字连在一起，导致翻译规则的边界发生错误。为了更好的获得命名实体的翻译，我们利用汉语和英语的单语命名实体识别工具来约束命名实体对的边界，用双语词对齐作为命名实体对应关系的知识来源，对命名实体单独训练了一个翻译表。并将命名实体翻译的结果融合到其他翻译规则的翻译过程中。

2.5.1 命名实体翻译的获取

命名实体翻译的获取过程如下：

1. 利用 Stanford 的命名实体识别工具[5]对平行训练语料的中文和英文部分分别进行命名实体识别。

2. 利用 Giza++ 对未进行命名实体处理的平行训练语料进行词对齐，训练过程采用2.2节中描述的第一种词对齐配置。

3. 为平行训练语料中出现的每一个中文命名实体短语 c 抽取可能的英文命名实体短语 e 作为翻译。如果 e 中不存在与 c 以外的词进行词对齐的词，则将翻译对 (c,e) 加入表 T；否则，将 (c,e) 加入表 T'。

4. 为翻译表 T 和 T' 分别计算翻译概率 $P(e|c)$，为每个中文命名实体短语 c 保留概率最高的 5 个翻译结果。

因为对词对齐的要求较为严格，所以 T 表中获得翻译对数量较少，而质量较高；相对而言 T' 表中的翻译对数量较多，但包含一定的噪音。

考虑到受限项目的限制，我们的命名实体翻译没有使用任何的翻译词典，完全由平行训练语料训练得到。对于在上述命名实体翻译表中未能获取翻译的中文命名实体，我们使用汉字-拼音转换表进行翻译，以提高翻译效果；对于其中命名实体的末字为省、市、县、镇、乡、村、区、路、街的实体，分别翻译末字为对应的英文单词。

考虑到英文命名实体翻译为中文时的译法、用字变化和歧义较大，我们并未在英汉评测项目中使用命名实体翻译。

2.5.2 命名实体翻译的应用

我们对每一个待翻译的句子进行命名实体识别，并通过上节所述的命名实体翻译获取方法得到一系列可能的翻译。

将命名实体翻译合理的融入翻译模型的使用之中是一个较难解决的问题。为了简便起见，我们将命名实体翻译分为高、中、低三个置信度等级，并人工设置他们的翻译概率值。对于高置信度的翻译，我们期望系统一定要采用，因而概率设为 0.99；对于低置信度的翻译，我们期望系统将之作为备选项使用，因而概率值设为相对比较低的 0.001；对于置信度中等的翻译，我们设定其概率值为 0.01。命名实体翻译相关的语言模型和其他特征的计算保持不变。

对于翻译置信度的设定采用以下规则：

1. 翻译表 T 中概率大于 0.1 且出现次数大于 5 次的翻译设为高置信度；概率小于 0.01 或出现次数仅为 1 次的翻译设为低置信度；其余为中置信度。

2. T' 中翻译的置信度采用 1 中的计算方法，但依次降一级。

3. 通过汉字 -拼音转换表获取的翻译设为低置信度。

2.6 解码器

本次评测中所有项目使用的是一个自行实现的基于 CKY 算法的解码器 (Chiang, 2005)。解码器的搜索基于一个对数线性模型，该模型包含的特征如下：

- 翻译模型特征 包括正向、反向的词汇化翻译概率和短语翻译概率
- 规则频率特征 短语或层次短语在平行训练语料中出现频次的负倒数 (Chen et al., 2011)
- 语言模型特征 包括每个语言模型的得分（每个语言模型对应于一个单独的特征）
- 计数类型特征 包括翻译中使用的短语数、层次短语数、单词数等

[5]http://www-nlp.stanford.edu/software/CRF-NER.shtml

各特征的权重由最小错误率训练（Minimum Error Rate Training，MERT）(Och, 2003) 训练得到。

2.7 后处理

英文的翻译结果在提交前进行了大小写还原和 detokenize。大小写还原使用的是一个短语长度为 2 且只包含 glue rule 的层次短语翻译系统。detokenizer 为 tokenizer 的反向处理过程。

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。

3 数据使用

本节依次说明我们的系统在汉英新闻、英汉新闻、英汉科技项目使用的训练和测试数据集。

3.1 汉英新闻项目

表 1中给出了汉英新闻项目中使用的平行训练数据，去除在预处理、分词过程中出现异常的句对后，实际使用的训练语料共有句对 4,690,575 条。

数据集名称	句对数
CLDC-LAC-2003-004	252,329
CLDC-LAC-2003-006	200,082
XMU-EC-Movie	176,148
HIT-IR	100,000
HIT-MT	52,227
Datum	1,000,004
ICTWebCEVersion2013	1,909,794
NEUCE	1,000,000
合计：	4,690,582
处理后：	4,690,575

Table 1: 汉英、英汉新闻项目平行训练数据

表 2中给出了汉英新闻项目中使用的语言模型数据。除此之外，我们还使用表 1中数据的英文部分进行了语言模型的训练。

数据集名称	句子数	单词数
reuters	12.3M	199.6M

Table 2: 汉英新闻项目语言模型训练数据

表 3中给出了汉英新闻项目中使用的开发和内部测试数据。

3.1.1 英汉新闻项目

英汉新闻项目的平行训练数据与汉英新闻相同，在表 1中给出。

数据集名称	用途	句子数
cenews-dev	开发	1006
SSMT2007	测试	1002
863-2005zewrit	测试	489

Table 3: 汉英新闻项目开发和内部测试数据

表 4中给出了英汉新闻项目中使用的语言模型数据情况，在英汉新闻项目中，由于发布时间较迟，我们并没有使用 sogou2013 的数据。除此之外，我们还使用表 1中数据的中文部分进行了语言模型的训练。

数据集名称	句子数	单词数
sogou 分词	32.5M	579.8M
sogou 分字	32.5M	879.0M
sogou2013 分词	14.0M	433.4M
sogou2013 分字	14.0M	623.0M

Table 4: 英汉新闻、英汉科技项目语言模型训练数据

表 5中给出了英汉新闻项目中使用的开发和内部测试数据。

数据集名称	用途	句子数
ecnews-dev	开发	1000
SSMT2007	测试	995
863-2005ezwrit	测试	494
ectech-dev	开发	1116

Table 5: 英汉新闻、英汉科技项目开发和内部测试数据

3.1.2 英汉科技项目

表 6中给出了英汉科技项目中使用的平行训练数据，经处理后共计句对 911,527 条。

数据集名称	句对数
ISTIC-EC-Tech	911,527
合计：	911,527
处理后：	911,527

Table 6: 英汉科技项目平行训练数据

表 4中给出了英汉科技项目中使用的语言模型数据情况。除此之外，我们还使用表 6中数据的中文部分进行了语言模型的训练。

表 5中给出了英汉新闻项目中使用的开发数据，由于该领域相关资源较少，所以我们并没有进行内部测验。

4 实验

本节将按照汉英新闻、英汉新闻、英汉科技项目的顺序分别介绍相关的实验情况。由于MERT 的结果有一定的随机性，我们在每组实验中都进行了三次独立的 MERT 运行，在此汇报三次运行的平均结果。本节中的评分都是采用 NIST-BLEU，其中，汉英项目的得分是大小写不敏感的 4gram 得分，英汉项目的得分是基于字的 5gram 得分。

4.1 汉英新闻项目

在汉英新闻项目中，我们主要实施了如下几种改进方式：词对齐融合、加入平行语料的英文部分作为独立语言模型、加入平行语料的英文部分训练反向语言模型、翻译规则平滑、短语频率特征、命名实体翻译等。

表 7中给出了前文提出的诸项改进对基线系统得分的贡献。其中，baseline 系统是一个基本的基于层次短语的翻译系统，+align 表示采用了词对齐融合技术的系统，+pLM 表示加入了平行语料的英文部分作为语言模型的系统，+pLM+rpLM 表示同时加入了平行语料的英文部分的正向、反向两种语言模型的系统，+mkn 表示翻译规则采用 modified Keneser-Ney 方法进行平滑，+elf 表示采用 Enhanced Low-Frequency 即短语频次特征。

从结果中可以看出，单独使用上述诸项改进都在一定程度上提高了系统的翻译性能，其中，使用语言模型的效果最为明显（+2.23%）；翻译模型的平滑和改进带来的提高也很稳定（+0.42%）。综合词对齐、语言模型、翻译模型的改进可以提高系统性能（+2.65%）。在此基础上，加入命名实体的翻译结果，可以进一步提高翻译性能（+3.14%）。本项目最终提交的系统为 all+ner 系统。

4.2 英汉新闻项目

在英汉新闻项目中，我们主要实施了如下几种改进方式：词对齐融合、加入平行语料的中文部分作为独立语言模型、同时采用基于字和基于词的语言模型、翻译规则平滑等。

由于时间关系，我们并未能在英汉项目上完成所有的实验。下面仅列出已完成的实验项目（表 8）。其中，baseline 系统是一个基本的基于层次短语的翻译系统，cLM 表示采用了基于字的 LM 替换了原有基于词的 LM 的系统，+cLM+pLM+pcLM 表示加入了平行语料的中文部分作为语言模型的训练数据，并且同时训练了基于词和基于字的语言模型（共使用 4 个语言模型）的系统，all 表示采用了上述 4 个语言模型，并且加入了词对齐融合和 MKN 平滑的系统。本项目最终提交的系统为 all 系统。

从上述结果中可以看出，语言模型的实验仍是提高系统性能的最主要原因 (+1.59%)。相比之下，采用词对齐融合、平滑等方法带来的提升相对较小 (+0.3%)。

4.3 英汉科技项目

在英汉新闻项目中，我们主要实施了如下几种改进方式：词对齐融合、加入平行语料的中文部分作为独立语言模型、同时采用基于字和基于词的语言模型、翻译规则平滑、短语频率特征等。

基于英汉新闻领域的实验结果，我们将 sogou2013 的中文数据单独作为训练语言模型的数据来源，并训练了基于分词和基于分字两个语言模型。表 9中给出了加入 sogou2013 训练数据前后的翻译性能对比。其中，4LM 表示使用了原 sogou 数据和平行语料的汉语部分进行语言模型训练，并分别训练基于分词和基于分字的语言模型的系统；+sogou2013 表示在上述系统中继续加入 sogou2013 数据上训练的基于分词和基于分字的语言模型的系统。实验结果表明，加入的新的语言模型数据仍能进一步的提高机器翻译的质量（+0.23%）。本项目最终提交的系统为 +sogou2013 系统。

5 小结

本文描述了南京大学自然语言处理研究组参加 CWMT2013 机器翻译评测的情况。在评测过程中，我们在预处理、词对齐、翻译模型、语言模型、解码训练等多个阶段尝试了多种不同的技术以提高系统的性能。综合利用本文中提到的诸项技术，我们成功的将各个项目的 BLEU 性能提高了 1 个点以上，在各个项目上都显著的超越了评测组织方提供的基线系统，并在参评各系统中表现优秀，这对我们系统建设来讲是一个很大的肯定。

在评测实验中我们发现，部分技术带来的提高虽小，但却较为稳定。一些技术虽然在部分数据集上能提高较多，但在另外的数据集上的得分往往发生了下降。综合而讲，需要得到稳定而显著的提高并不容易。这可能与不同数据的领域属性有关。在下一阶段，我们准备在领域适应性方面投入更多的精力进行研究。

致谢

本次评测受到了国家自然科学基金项目（61170181，61003112）、江苏省自然科学基金项目（SBK201341112）的资助。感谢黄宜华教授研究组提供基于 Hadoop 平台的分布式计

系统名称	dev	ssmt07	05zewrit	average	+/-
baseline	27.97%	28.14%	25.58%	27.23%	–
+align	27.84%	27.97%	25.99%	27.27%	0.04%
+pLM	30.83%	29.81%	26.56%	29.07%	1.83%
+pLM+rpLM	31.11%	30.09%	27.20%	29.47%	2.23%
+mkn	28.14%	28.28%	25.56%	27.33%	0.10%
+elf	28.31%	28.41%	25.95%	27.55%	0.32%
+mkn+elf	28.48%	28.62%	25.86%	27.65%	0.42%
all	31.55%	30.64%	27.46%	29.88%	2.65%
all+ner	**32.15%**	**31.11%**	**27.85%**	**30.37%**	**3.14%**

Table 7: 汉英新闻项目上的实验结果

系统名称	dev	ssmt07	05zewrit	average	+/-
baseline	34.61%	36.82%	37.98%	36.47%	–
cLM	34.63%	36.20%	38.71%	36.51%	0.04%
+cLM+pLM+pcLM	36.35%	38.31%	39.53%	38.06%	1.59%
all	36.88%	38.46%	39.75%	38.36%	**1.89%**

Table 8: 英汉新闻项目上的实验结果

系统名称	dev	+/-
4LM	52.49%	–
+sogou2013	52.72%	0.23%

Table 9: 英汉科技项目上的实验结果

算支持，感谢李斌副教授在评测进行中给予的帮助和建议。

References

Boxing Chen, Roland Kuhn, George Foster, and Howard Johnson. 2011. Unpacking and transforming feature functions: New ways to smooth phrase tables. In *Proceedings of Machine Translation Summit*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *annual meeting of the Association for Computational Linguistics*.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable smoothing for statistical machine translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Ning Xi, Guangchao Tang, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2012. Enhancing statistical machine translation with character alignment. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 285–290, Stroudsburg, PA, USA. Association for Computational Linguistics.

Deyi Xiong, Min Zhang, and Haizhou Li. 2011. Enhancing language models in statistical machine translation with backward n-grams and mutual information triggers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1288–1297, Portland, Oregon, USA, June. Association for Computational Linguistics.