RoleTrans: 中科院自动化研究所多语言文本机器翻译系统

RoleTrans: Multilingual Machine Translation System in CASIA

张家俊 翟飞飞 周玉 汪昆 陈钰枫 涂眉 李小青 宗成庆
中国科学院自动化研究所

江涛 于洪志
西北民族大学

Abstract

本文将主要介绍中科院自动化研究所参加CWMT2013机器翻译系统评测的总体情况。我们参加了所有评测任务，包括汉英新闻领域机器翻译、英汉新闻领域机器翻译、英汉科技领域机器翻译、蒙汉日常用语机器翻译、藏汉政府文献机器翻译、维汉新闻领域机器翻译。文章首先介绍我们主要采用的统计机器翻译系统框架，然后着重介绍在此次评测中采用的新技术，最后给出实验和评测结果并对结果进行简单的分析。

Abstract

This paper describes an overview of CASIA technical report for CWMT2013. We participated all of the evaluation tasks: Chinese-to-English Translation for News, English-to-Chinese Translation for News, English-to-Chinese Translation for Scientific and Technological Text, Mongolian-to-Chinese Translation for Daily Expressions, Tibetan-to-Chinese Translation for Government Documents, Uyghur-to-Chinese Translation for News. We first introduce the translation framework we mainly used in this evaluation, and then present the new technologies which we have applied. Finally, we give the experimental settings, show the evaluation results and make some analysis.

1 引言

CWMT2013一共包括6个评测方向：汉英新闻领域机器翻译、英汉新闻领域机器翻译、英汉科技领域机器翻译、蒙汉日常用语机器翻译、藏汉政府文献机器翻译、维汉新闻领域机器翻译。为了检验机器翻译系统的进展，每个评测方向都设立了进展评测(Progress)。除了汉英新闻领域机器翻译，其他方向都设立了当前评测(Current)。中科院自动化研究所(CASIA)参加了所有的评测任务。本文将介绍我们主要采用的统计机器翻译系统框架、新采用的技术以及系统在各个翻译任务上的性能表现。最后我们将对结果进行简单分析。

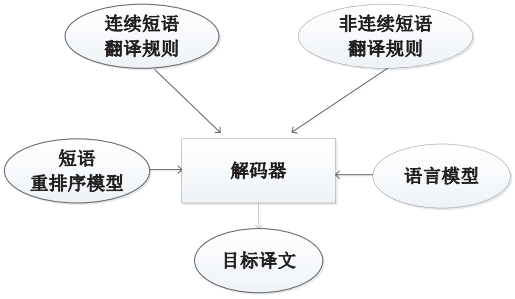

Figure 1: RoleTrans-基于功能的翻译框架

2 RoleTrans: 基于功能的翻译系统

此次评测中，除了使用开源的基于短语的统计机器翻译系统(Moses-BP)以及基于层次短语的统计机器翻译系统(Moses-HP) (Koehn et al. , 2007)，我们主要采用了本单位开发的改进后的基于功能的统计机器翻译系统(RoleTrans)。下面，我们对该系统的框架及模型进行详细介绍。

从翻译功能的角度说来，翻译一个源语言句子需要完成连续短语的翻译、短语的重排序、非连续短语的翻译(分为源端非连续和目标端非连续)以及译文流畅性的度量(语言模型)等等。基于功能的统计机器翻译模型以实现这些功能为目标，并面向不同的语言、不同的翻译方向对各种功能进行重要性优化。Figure 1描述了RoleTrans的基本框架。

不同于2011年我们采用的以括弧转录文法为主要框架的翻译系统 (Zhang and Zong, 2009) ，改进后的RoleTrans以层次短语翻译系统 (Chiang, 2007)为基础，实现连续短语

和非连续短语的翻译。相比于层次短语翻译模型， RoleTrans将连续短语翻译、源语言端非连续短语翻译以及目标语言端非连续短语翻译视作不同的子模型，分别进行权重调节与优化。因为我们在分析不同语言不同方向的翻译现象时发现，非连续短语翻译在某些语言对的翻译中非常重要，例如汉语和英语之间的翻译，而在很多语言对中可有可无，例如西班牙语和英语之间的翻译。而且，即使在同一个语言对中，例如汉语和英语，不同种类的非连续短语翻译对翻译质量的影响差别很大。例如在汉语到英语的翻译中，源端非连续的短语翻译比目标端非连续的短语翻译更加重要，在英语到汉语的翻译中则相反 (Zhang and Zong, 2012)。

众所周知的是，层次短语翻译模型在短语重排序方面的能力较弱。短语之间的调序直接由层次短语翻译规则完成，而层次短语翻译模型仅仅根据词汇或者短语匹配确定规则的使用，因此缺少泛化能力。对此，我们借鉴括弧转录文法的思想，在层次短语翻译模型中融入二元规则$X \rightarrow (X_1 X_2)$。目标译文短语的调序由基于边界词的最大熵分类器进行预测。值得注意的是，在规则抽取过程中，我们并不抽取这类二元规则，解码中非终结符X直接由连续短语的翻译规约而成。

除了翻译规则和短语重排序模型，语言模型是翻译系统中的另一个重要因素。语言模型直接衡量译文是否流畅、是否满足目标语言的文法。最近有不少研究者尝试更有表达能力的语言模型建模方法，例如基于递归神经网络的语言模型等等。我们在系统 RoleTrans中仍然采用传统的基于 $ngram$的语言模型，不同的是，我们不仅采用了传统前向$ngram$模型，同时尝试了后向$ngram$模型。假设目标译文为：

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

那么前向$ngram$语言模型的概率为：

$$\begin{aligned} p_{forward}(e) &= p_{forward}(e_1 e_2 ... e_{m-1} e_m) \\ &= p(e_1) p(e_2|e_1) ... p(e_n|e_{m-n+1} ... e_{m-1}) \end{aligned} \quad (1)$$

对应的，后向$ngram$语言模型的概率为：

$$\begin{aligned} p_{backward}(e) &= p_{backward}(e_1 e_2 ... e_{m-1} e_m) \\ &= p_{forward}(e_m e_{m-1} ... e_2 e_1) \\ &= p(e_m) p(e_{m-1}|e_m) ... p(e_1|e_n ... e_2) \end{aligned} \quad (2)$$

RoleTrans在解码时类似于利用同步文法进行双语分析，目标是搜索一个最优目标译文对应的推导d'：

$$d' = argmax_{d \in D} P(d) \quad (3)$$

其中$P(d)$可由对应各个功能的得分计算：

$$\begin{aligned} P(d) = & P(PhrTrans)^{\lambda_1} P(SrcDis)^{\lambda_2} \\ & P(TgtDis)^{\lambda_3} P(BothDis)^{\lambda_4} \\ & P(BTG)^{\lambda_5} P(fLM)^{\lambda_6} P(bLM)^{\lambda_7} \end{aligned} \quad (4)$$

其中，$PhrTrans$对应了连续短语翻译得分，$SrcDis$、$TgtDis$与$BothDis$分别对应源端非连续短语翻译，目标端非连续短语翻译与两端非连续短语翻译的得分。BTG表示基于BTG的短语调序得分，fLM与bLM分别对应前向与后向$ngram$语言模型的得分。当然，整个推导还要加入译文长度的惩罚特征以及对使用规则数目的惩罚特征。特征权重λ_i由最小错误率训练获得。

3 新技术和方法

相比于CWMT2011中所使用的系统，我们在此次评测中尝试了多种新的技术和方法。除了上述提到的RoleTrans的改进外，我们主要在词语对齐、命名实体的识别和翻译方面进行了针对性的研究。

3.1 基于句法的词语对齐

传统的词语对齐方法基于 IBM翻译模型，未利用任何深层次信息。然而，对于很多语言对，研究者发现依存凝聚性质普遍存在于自然语言之间。依存凝聚性是指源语言端两个不交叠的依存子树所覆盖的短语经翻译得到的目标语言译文也不交叠。可以设想，若能有效利用这种性质指导训练语料双语句对的词语对齐过程，我们将能获得更加准确的对齐结果。为此，我们尝试将依存凝聚性作为软性约束有机地融入基于 Bayesain模型的词语对齐学习模型中 (Wang and Zong, 2013) 。为了得到依存树结果，我们采用了Berkeley Parser[1]。该方法在NIST评测语料中不仅有效提高了词语对齐的质量，而且还显著地改善了统计机器翻译的译文质量。

3.2 受限的命名实体翻译

不同于往年的评测要求，此次评测在进行命名实体的翻译时禁止使用大规模外部资源。为此我们在汉语和英语的翻译任务中进行了针对性的处理。对于人名，我们采用音译的方法。对于其他类别的实体，我们首先从训练语料中抽取汉英实体翻译对作为评测中实体翻译所需的词典和训练语料 (Chen et al. , 2013)。针对汉语命名实体，我们采用自主开发的基于词

[1]https://code.google.com/p/berkeleyparser

典与训练语料的汉语实体识别及翻译系统。特别地，在汉英实体翻译中，我们对地名采用基于词典和音译的方式进行组合翻译，而机构名的翻译则利用基于结构的层次翻译模型。

4 翻译流程中各模块的处理方法

4.1 数据预处理

针对不同的语言，我们采用了不同的预处理方法。对于汉语，我们首先进行分词以及全角变半角的转换，然后我们将所有标点符号统一替换为汉语标点。其中汉语分词利用了我们自主研发的工具Urheen[2]。对于英语，我们首先进行大写转小写操作以及tokenization操作，然后将所有标点符号统一替换为英文标点。对于维吾尔语，我们采用了新疆理化所与新疆大学开发的维吾尔语词法分析工具对维吾尔语进行词干化。对于藏语，我们采用了西北民族大学的藏语分词工具对训练语料、开发集以及测试集的藏语进行分词。对于蒙古语，我们仅仅对标点符号做了分离操作。

4.2 词语对齐

除了自主研发的基于句法的词语对齐方法外，我们同时采用了两种开源对齐工具：GIZA++与Berkeley对齐工具[3]。其中GIZA++的词对齐我们采用grow-diag-final-and(gdfa)的扩展方式来得到多对多的词对齐结果。根据不同的语种，我们评估不同的词对齐对翻译性能的影响，最终选择最好的对齐组合用于翻译模型的获取。

4.3 多语种时间数字识别与翻译

多语种时间数字识别与翻译利用规则方法。考虑到时间数字信息的多样性，我们将各个语种(包括汉语、英语、藏语、蒙语和维语)的时间数字细化为六类：数量(Number)、序数词(Ordinal)、号码(Figure)、月份(Month)、日期(Date)和星期(Week)。其中目标语言的翻译方式主要选择与开发集参考答案相同的形式。该部分工作并将程序与规则严格分离开来（Tu et al. ，2012），从而使之能够迅速进行扩展和移植，成功实现了多语种的时间数字识别及翻译。

4.4 数据后处理

对于汉语，我们没有进行任何后处理，直接使用输出文件。在提交结果时，我们将去掉未登录词的和保留未登录词的作为两个结果提交。对于英文主要是根据tokenization的方式进行了detokenization，并进行了大小写转换。大小写转换中，我们利用原始训练语料的目标语言端训练一个大小写混合的语言模型，然后借助Moses中的recaser工具对翻译结果进行大小写转换。

[2]http://www.openpr.org.cn/index.php/NLP-Toolkit-For-Natural-Language-Processing

[3]http://code.google.com/p/berkeleyaligner

5 实验与结果

在所有的评测任务中，我们只参加了受限的评测。也就是说我们只是使用了主办方提供的所有训练数据，未使用任何外部资源。其中，英语和汉语的语言模型训练用的也是主办方提供的路透社单语语料和搜狗单语语料。在翻译模型的参数调节中，我们也仅使用了主办方发布的所有的开发集。下面，我们首先介绍翻译模型、语言模型和短语重排序模型的一些重要设置和选择。

5.1 短语长度选择

对于所有评测任务，短语长度设置如下：在基于短语的翻译模型中，我们设置短语最大长度为 7；在基于层次短语的翻译模型中，我们设置短语的最大长度为 5。在实验过程中，为了提高层次短语翻译模型的覆盖率，我们也尝试了将短语模型中长度为6和7的短语翻译规则加入层次短语翻译模型。

5.2 词语对齐选择

在翻译模型训练中，我们采用了多种词对齐。我们使用GIZA++，Berkeley对齐工具，以及我们自主开发的基于句法的词对齐工具，分别产生如下的词对齐：

(1) giza-gdfa：使用GIZA++词对齐工具生成两端的词对齐，并使用grow-diag-final-and (gdfa) 策略获取最终的词对齐用于翻译模型的抽取。

(2) berk-align：使用berkeley对齐工具，并利用双语数据生成的词对齐结果。

(3) syn-berk-align：使用berkeley对齐工具，利用双语数据、以及汉语端的句法分析树生成的基于句法的词对齐结果。

(4) iasyn-align: 使用实验室自主开发的基于句法的对齐工具生成两个方向的对齐，并使用gdfa策略获取最终的对齐结果。对于小语种，由于我们只有汉语端的句法分析结果，因此我们只采用了一个方向的对齐结果用于抽取翻译模型。

对于不同的语种，我们分别评估上述不同的词对齐组合对于翻译性能的影响，并最终选择最好的对齐组合用于翻译模型的获取。通过多组实验，各个参评系统最终采用的词对齐方式

Task	Alignments
CENews	giza-gdfa iasyn-align
ECNews	giza-gdfa iasyn-align
ECSci	giza-gdfa berk-align
MC	giza-gdfa berk-align syn-berk-align iasyn-align
UC	giza-gdfa berk-align syn-berk-align iasyn-align
TC	giza-gdfa berk-align syn-berk-align iasyn-align

Table 1: 不同任务的对齐方式选择

如Table 1所示。

5.3 语言模型的选择

我们设置并训练了三种目标语言模型：

1，以主办方发布的双语训练语料的目标语言作为训练集得到5元的语言模型，我们称之为self.lm5；

2，以大规模单语语料（中文搜狗语料或英文路透社语料）作为训练集得5元的语言模型big.lm5；

3，以训练语料的目标语言利用困惑度准则对大规模单语语料进行过滤，用过滤后的部分作为训练集得到的语言模型mid.lm5。

我们在实验中尝试了不同语言模型的结合对开发集和验证集的影响。最终，对于汉英、英汉方向，我们均选择了self.lm5与mid.lm5的组合方式，而对于小语种(蒙汉、藏汉、维汗)，我们选择使用self.lm5与big.lm5的组合方式。

5.4 短语重排序模型的训练

评测发布的训练语料的双语句子的对齐质量参差不齐，影响抽取短语重排序实例的质量，但由于时间与机器资源的因素，我们并没有针对短语重排序模型进行特殊的训练语料选择。在系统RoleTrans中，对于每一个评测任务，我们采用所有训练语料，并利用giza-gdfa的词语对齐方式抽取短语重排序实例。最后，我们以边界词作为特征训练基于最大熵 (Zhang, 2004)的短语重排序模型。

5.5 汉英新闻实验结果

表2 给出了两个翻译系统在开发集和测试集上的实验结果。需要说明的是所有结果都基于BLEU-SBP4打分，其中在开发集上采用的是大小写不敏感打分，而在测试集上采用的是大小写敏感打分。从表中，我们可以看到，无论是开发集还是测试集，汉英翻译中改进后的基于功能的统计机器翻译系统RoleTrans要明显好于基于层次短语的翻译系统Moses-HP，在进展评测的数据上提高了0.87个BLEU-SBP4。结果表明汉英评测中区分非连续短语的类型以及融入基于最大熵的短语重排序模型能够极大地改善统计机器翻译的效果。

Systems	Tune	Progress
Moses-HP	29.84	24.31
RoleTrans	**30.67**	**25.18**

Table 2: 汉英新闻实验结果对比

Systems	Tune	Progress	Current
Moses-HP1		36.21	35.11
Moses-HP2	35.27	36.18	35.13
RoleTrans1		35.94	35.46
RoleTrans2	35.31	35.97	35.47

Table 3: 英汉新闻实验结果对比

遗憾的是，我们的主系统没有选择正确。由于担心系统在开发集上过训练，我们选择了一部分NIST测试集作为验证集来确定主系统。虽然，系统RoleTrans在开发集上表现最好，但在验证集上并没有超越层次短语翻译系统。因此，我们保守地选择了层次短语模型作为我们的主系统。其实，在CWMT评测中，我们还是不应该相信系统在NIST测试集上的表现，因为数据的分布差别很大。在以后的评测中，我们将会针对性地研究主系统的选择方法。

5.6 英汉新闻实验结果

表3 显示的是两个翻译系统在英汉新闻评测任务中开发集和测试集上的实验结果。表中的数字都是基于汉字的BLEU-SBP5打分。Moses-HP1与Moses-HP2的区别在于前者去掉了测试集结果中的未登录词词，RoleTrans1和RoleTrans2的区别类似。由于我们未针对开发集进行对比测试，因此，两个系统在开发集上的得分都是保留未登录词的结果。可以看出，在英汉新闻评测任务中，基于功能的统计机器翻译模型与层次短语翻译模型表现相当。基于功能的翻译系统RoleTrans在Current集合上略优于层次短语翻译系统，而在Progress集合上，层次短语翻译模型则稍好一点。另外，从实验结果中，我们可以发现未登录词的保留与否对翻译结果的影响不大。

Systems	Tune	Progress	Current
Moses-HP1		38.42	36.38
Moses-HP2	41.0	38.36	36.00
RoleTrans1		39.32	35.85
RoleTrans2	41.1	39.38	36.24

Table 4: 英汉科技实验结果对比

Systems	Tune	Progress	Current
Moses-HP	52.65	37.84	13.65
Moses-BP	52.54	37.24	13.38
RoleTrans1		37.25	13.41
RoleTrans2	52.87	36.77	12.89

Table 5: 蒙汉翻译实验结果对比

5.7 汉英科技实验结果

英汉科技领域的句子表达比较固定，非常适合于融入模板匹配等方法。由于时间和资源的限制，我们没有做任何针对性的处理。这可能是我们的结果不是非常理想的原因之一。

类似于英汉新闻翻译，我们同样提交了去掉未登录词与保留未登录词的结果。Moses-HP1和RoleTrans1表示去掉未登录词的结果。

在Progress集合上，基于功能的统计机器翻译系统 RoleTrans要明显好于层次短语翻译系统；而在 Current集合上两个系统具有类似的效果。同样，翻译结果中未登录词的保留与否对译文的自动评价影响不大。

5.8 蒙汉翻译实验结果

在蒙汉评测任务中，除了层次短语翻译系统与基于功能的翻译系统，我们同时尝试了传统的基于短语的翻译系统。从表5 中容易发现，各个系统之间的差别不大。总体而言，层次短语翻译模型在蒙汉翻译中相对更加稳定，在进展评测和当前评测中都取得了最好的翻译性能。

由于我们对蒙古语不甚了解，没有针对蒙古语做任何针对性的预处理工作。这也导致我们的翻译结果不是很理想。这也说明机器翻译研究不能对任何语言之间的翻译直接套用通用的统计模型，而应该针对语言做相应的分析和处理。这次评测让我们更加意识到这一点。

5.9 维汉翻译实验结果

在维汉评测任务中，我们在开发集上的实验结果发现层次短语翻译系统要明显好于基于短语的翻译系统与基于功能的翻译系统。因此，我们在维汉任务中仅提交了层次短语翻译系统的结果，并且所有的译文都是去掉未登录词的结果。

Systems	Tune	Progress	Currentt
Moses-HP1	52.16	52.34	50.21
Moses-HP2	52.21	52.08	49.99
Moses-HP3	52.25	53.10	50.04
Moses-HP4	52.34	52.19	50.47

Table 6: 维汉翻译实验结果对比

Moses-HP1表示在抽取词对齐时，我们使用了giza-gdfa和berkeley合并的词对齐结果。

Moses-HP2表示在Moses-HP1的基础上，我们加入了短语模型中长度为 6和 7的短语翻译规则。

Moses-HP3与Moses-HP1的配置类似，区别在于使用了评测组织方更新后的训练数据。

Moses-HP4表示在抽取词对齐时，我们综合了giza-gdfa、Berkeley和自主研发的基于句法的词语对齐结果。

从表6 给出的实验结果可以看出，各个系统在所有集合上的表现非常相似。这也说明了在维汉翻译任务中，综合各种词语对齐方法对翻译效果的影响不大。

5.10 藏汉翻译实验结果

在藏汉评测任务中，我们同样只提交了层次短语翻译系统的结果，并且所有译文都是去掉未登录词的结果。各个系统的区别如下：

Moses-HP1 表示在抽取词对齐时，我们使用了giza-gdfa 和berkeley 合并的词对齐结果，并且加入了短语模型中长度为6和7的短语翻译规则。

Moses-HP2与Moses-HP1类似，只是没有包括短语模型中长度为6和7的短语翻译规则。

Moses-HP3与Moses-HP2类似，区别是在训练汉语语言模型时，没有使用搜狗提供的2012年数据。

Moses-HP4与Moses-HP2类似，不同之处在于我们进一步综合了自主研发的基于句法的词语对齐结果。

从表7 给出的实验结果，我们可以发现系统Moses-HP4 在进展评测和当前评测的任务中表现最好。对比结果表明，我们自主研发的基于句法的词语对齐方法在藏汉翻译任务中还是起到了积极的正面作用。不过，从各个评测任务综合来看，各种词语对齐方法综合是否有效还是因语言和翻译方向而定。

我们在藏汉翻译任务中使用了多种新技术和方法，例如新的分词工具、不同的词对齐融合策略和不同的语言模型方法。为了检验每一种技术的作用，我们分别对比了不同的技术在开发集上对最终翻译质量的影响。表8 展示了详

Systems	Tune	Progress	Current
Moses-HP1	66.03	63.28	26.47
Moses-HP2	65.85	62.70	26.24
Moses-HP3	65.74	62.39	26.16
Moses-HP4	65.85	63.67	26.69

Table 7: 藏汉翻译实验结果对比

Model	Tech.	Tune
Moses-BP	ictseg	48.49
	nunseg	54.19
	nunseg+biglm	56.99
Moses-HP	nunseg+biglm	61.15
	+berk-align	64.84
	+syn-berk-align	65.85
	+iasyn-align	65.85

Table 8: 不同技术对藏汉翻译结果的影响

细的对比结果。其中，ictseg表示组织方计算所提供的藏语分词工具，nunseg表示西北民族大学提供的藏语分词工具。从实验结果可以看出，相比于ictseg，nunseg藏语分词工具在翻译质量上提高了5.7个BLEU-SBP 值，效果非常显著。加入大规模语言模型也能获得非常明显的提升，提高了近3个 BLEU-SBP。对比短语翻译模型，层次短语翻译模型在相同设置的基础上能够显著提升翻译质量近4个 BLEU-SBP。进一步融合多个词语对齐结果也能够改善最终的翻译质量。可见，在训练数据规模较小的翻译任务中，综合不同的预处理技术、词语对齐方法以及语言模型策略都能有效地改善译文质量。

6 总结

本文主要介绍了中国科学院自动化研究所参加CWMT2013评测的总体情况。在绝大多数的评测任务中，我们都取得了比较理想的成绩。在翻译模型方面，我们发现基于功能的统计机器翻译系统 RoleTrans在汉英新闻任务中表现突出。一方面说明了综合不同的翻译功能并对不同的功能进行针对性优化能够明显改善翻译效果，另一方面，由于我们平时的研究通常都是针对汉英翻译进行实验对比和分析，因此这种方式也基本能够保证该翻译方向的译文质量。相比而言， RoleTrans在其他翻译任务中就没有体现出非常明显的优势。这进一步说明，任何一个翻译模型都不能保证适用于任何语言对和任何翻译方向。在词语对齐方法的结合与语言模型的选择上，实验结果表明词对齐合并技术和语言模型的结合技术在大多数翻译任务上都比较有效，能够提高译文质量；但效果的程度非常依赖于具体的语言。在多语种的时间数字识别和翻译以及命名实体的识别和翻译方面，我们实验结果发现它们非常重要。但是，由于采用的是规则方法，译文的格式（尤其时间数字翻译）就显得格外重要。因此，我们希望以后的评测中译文中时间数字的格式能够有一个统一的标准。

总而言之，从各个任务翻译评测中，我们确实能够发现不少问题。同时，我们也意识到，我们的翻译模型和系统还有很大的提升空间，比如基于功能的翻译系统在很多语言间的翻译中无法体现其模型优势。我们希望在以后的研究中与国内外同行多交流、多学习，不断改善我们现有的模型和系统，也为不断提升我国的机器翻译水平贡献绵薄之力。

7 致谢

此次评测中，中科院自动化研究所模式识别国家重点实验室的很多同学付出了许多艰辛的劳动，给予了很多工作上和精神上的支持。在此对他们表示衷心的感谢！并特别感谢西北民族大学教育部重点实验室中国民族语言文字信息技术实验室的祁坤钰老师和各位同学（青措、努尔、都乐根、英草吉），以及新疆大学的卡尔老师、于斯音和阿哈麦提同学给予的大力帮助和支持。

References

Yufen Chen, Chengqing Zong, and Keh-Yih Su. 2013. A Joint Model to Identify and Align Bilingual Named Entities. *Computational Linguistics*, 39(2):229-266.

David Chiang. 2007. Hierarchical Phrase-based Translation. *Computational Linguistics*, 33(2): 201-228.

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 and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Proc. of ACL 2007*.

Mei Tu, Yu Zhou and Chengqing Zong. 2012. A Universal Approach to Translating Numerical and Time Expressions. *Proc. of International Workshop on Spoken Language Translation 2012*, 209-216.

Zhiguo Wang and Chengqing Zong. 2013. Large-scale Word Alignment Using Soft Dependency Cohesion Constraints. *Transactions of Association for Computational Linguistics*, 1(6):291-300.

Le Zhang. 2004. Maximum Entropy Modeling Toolkit for Python and C++. *http://homepages.inf.ed.ac.uk/s0450736/maxent-toolkit*.

Jiajun Zhang and Chengqing Zong. 2009. A Framework for Effectively Integrating Hard and Soft Syntactic Rules into Phrase-Based Translation. *Proc. of the 23rd Pacific Asia Conference on Language, Information and Computation*, 579-588.

Jiajun Zhang and Chengqing Zong. 2012. A Comparative Study on Discontinuous Phrase Translation. *Proc. of the 1st Conference on Natural Language Processing and Chinese Computing*, 164-175.