# 基于最长名词短语分治策略的神经机器翻译*

张学强，蔡东风，叶娜，吴闯

（沈阳航空航天大学 人机智能研究中心，辽宁 沈阳 110136）

**摘要：**神经机器翻译自兴起以来，不断给机器翻译领域带来振奋人心的消息。但神经机器翻译没有显式地利用语言学知识对句子结构进行分析，因此对结构复杂的长句翻译效果不佳。本文基于分治法的思想，识别并抽取句子中的最长名词短语，保留特殊标识或核心词与其余部分组成句子框架。通过神经机器翻译系统分别翻译最长名词短语和句子框架，再将译文重新组合的方法，缓解了神经机器翻译对句子长度敏感的问题。实验结果表明，本文提出的方法获得的译文与基线系统相比，BLEU 分值提升了 0.89。

**关键词：**神经机器翻译；最长名词短语；分治策略

**中图分类号：**TP391　　**文献标识码：**A

# Neural Machine Translation Based on the Divide-and-Conquer Strategy of Maximal-length Noun Phrase

ZHANG Xueqiang, CAI Dongfeng, YE Na, WU Chuang

（Human-Computer Intelligence Research Center , Shenyang Aerospace University , Shenyang , Liaoning 110136 , China）

**Abstract:** Neural Machine Translation continues to bring exciting news to the field of Machine Translation since its rise. But Neural Machine Translation did not make explicit use of linguistic knowledge to analyze sentence structure, therefore, the translation effect is not good in long sentences with complex structure. This paper is based on the idea of divide-and-conquer strategy, identifying and extracting the Maximal-length Noun Phrases in a sentence, and retaining special marks or head words and the rest component to form the sentence framework. In order to alleviate the sensitivity of Neural Machine Translation to sentence length, through the way of the Maximal-length Noun Phrases and sentence frames are translated respectively by Neural Machine Translation model, and the translation of Maximal-length Noun Phrases and sentence frames are regrouped to a complete translation. Experimental results shows that, the method proposed yields a BLEU enhancement of 0.89 compared with the baseline system.

**Key words:** Neural Machine Translation；Maximal-length Noun Phrase；Divide-and-Conquer Strategy

## 1 引言

神经机器翻译（Neural Machine Translation，NMT）作为一种全新的机器翻译方法，近年来获得迅速发展。然而，神经机器翻译仅仅使用一个非线性的神经网络实现自然语言之间的转换[1]，相比于统计机器翻译，译文质量对句子长度更为敏感[2]。如何在神经机器翻译中将一个句子在尽量不损失语义信息的前提下，进行长度上的缩减和结构上的简化是一个值得探究的方向。

一般认为，自然语言中语义的基本单位是短语。因此，将句子级别的对齐和翻译进行到亚句子（Sub-sentence）的短语一级显得尤为重要。句子中的实体和概念通常可由名词短语（Noun Phrase，NP）来描述。其捆绑了一个相对完整的语义信息，具有丰富的句法功能，可在句中充当主语和宾语等成分。最长名词短语[3]（Maximal-length Noun Phrase，MNP）指不被其他任何名词短语嵌套的名词短语。与一般名词短语相比，MNP

*基金项目：国家自然科学基金（61403262）；国家自然科学基金（61402299）

作者简介：张学强（1992-），男，硕士研究生，主要研究方向为自然语言处理，机器翻译；蔡东风（1968-），男，博士，教授，主要研究方向为人工智能，自然语言处理，信息检索；叶娜（1981-），女，博士，讲师，主要研究方向为自然语言处理，机器翻译；吴闯（1985-），男，硕士，主要研究方向为自然语言处理，机器翻译。

具有更大的粒度，边界特征较为明显，有利于句子的整体结构分析。采用分治策略处理MNP，既能在亚句子一级上获得更精准的译文，也能在一定程度上将句子缩短为包含主干信息的句子框架。因此，准确地识别和翻译MNP，是利用分治策略提升机器翻译性能的一个有力手段。

针对神经机器翻译在长句翻译任务上的不足，考虑到MNP的处理可以在一定程度上简化句子结构，本文提出一种基于MNP分治策略的神经机器翻译方法。该方法基于一个“抽取-翻译-重组”的MNP处理框架，旨在将MNP独立处理带来更高质量的MNP和句子框架译文的优势，与神经机器翻译学习能力强、译文具有较高准确度和流畅度等优势相结合，以达到提升译文整体质量的目的。

## 2 相关研究

### 2.1 短语知识在机器翻译中的应用

在自然语言中，短语作为语义的基本单位具有重要的意义。将双语短语等语言学知识融入到机器翻译中，一直是研究人员孜孜追求的目标。

针对基于短语的统计机器翻译方法未充分利用语言学知识，长距离调序效果不好的问题，丁鹏[4]等提出一种基于双语句法短语的统计机器翻译方法。首先，采用一种基于期望最大化(Expectation Maximization，EM）的算法来抽取双语句法短语。然后，通过三种方法将短语应用到统计机器翻译系统中：(1)将双语句法短语加入到训练语料中，训练翻译模型；(2)将其加入到短语表中，计算短语的特征值；(3)增加一个句法短语特征到短语表中，表征其是否为句法短语。实验结果表明，这三种方法得到的译文BLEU分值分别在基线系统上提升了0.23、0.41和0.64。丁鹏等人的方法尽管利用到了双语句法短语，但整体框架仍然是基于短语的统计机器翻译方法，仍然面临着长距离调序效果不佳的问题。

针对上述问题，Xiaona Ren等[5]提出一种简化专利句子结构以提高翻译性能和后处理效率的方法。首先，Ren等人采用一种基于统计方法的识别器，对句中的MNP进行识别。在中文树库CTB5.1的专利语料上识别结果的F值达到62.28%。然后，Ren等人对MNP进行分析，在识别正确与错误的MNP中分别有97.92%和38.94%有利于后续的翻译过程。最后，在统计机器翻译方法上分别使用自动方法和人工方法对系统进行评价。与基线系统相比，该系统得到的译文BLEU分值提升了0.62;语义准确度和流畅度分别提升0.18和0.17；翻译效率提升了约100字/小时。Ren等人方法的不足在于，没有使用双语MNP扩展语料，以训练短语表、翻译模型和调序模型。MNP作为句子的一部分，翻译规则却与句子不尽相同。导致训练得到的模型能较好地翻译简化后的句子，却不能准确翻译MNP。

尽管神经机器翻译在英文-法文等语言对翻译任务上都表现出最佳的翻译性能[6][7][8]，但其面临着难以融入词典等语言学知识的挑战。Jiajun Zhang等[9]提出一种能够桥接神经机器翻译与双语词典的框架，将包含训练数据中的低频词与未登录词等信息的双语词典应用到神经机器翻译中，从而提升翻译性能。为此，Zhang等人针对该框架采用了两种模型：(1)混合词/字模型（Mixed Word/Character Model），将语料中频率超过一定阈值的词保留，而低频词和集外词则使用其字符序列进行替换；(2)伪句对合成模型（Pseudo Sentence Pair Synthesis Model），指针对低频词对或未登录词对$(D_{xi},D_{yi})$，构造J组包含$(D_{xi},D_{yi})$的句对$\{(\mathrm{X_{ij}},\mathrm{Y_{ij}})\}_{j-1}^{J}$。实验结果表明，与基线系统相比，混合词/字模型与伪句对合成模型的译文BLEU分值分别提升了1.0和1.62。Zhang等人的方法缓解了神经机器翻译方法的未登录词问题，但未能利用短语知识缓解译文质量对句子长度敏感的问题。

### 2.2 神经机器翻译

统计机器翻译（Statistical Machine Translation，SMT）主要存在三个挑战[10][1]：(1)线性不可分；(2)缺乏合适的语义表示；(3)难以设计特征。由于深度学习可以较好地缓解上述问题，完全基于深度学习的端到端神经机器翻译应运而生，并获得迅速发展。

[1] 实际上，引文中作者给出了六个挑战，这里只列举其三。

研究人员通过将现有的方法和策略引入端到端的神经网络，以实现翻译性能的不断提升。Sutskever 等[11]首次将长短期记忆[12]（Long Short-Term Memory，LSTM）引入到神经机器翻译，以缓解递归神经网络（Recurrent Neural Network，RNN）训练时“梯度消失”的问题，并且在“编码-解码”（Encoder-Decoder）框架两端同时采用 RNN。图 1 给出了 Sutskever 等人提出的神经机器翻译模型。

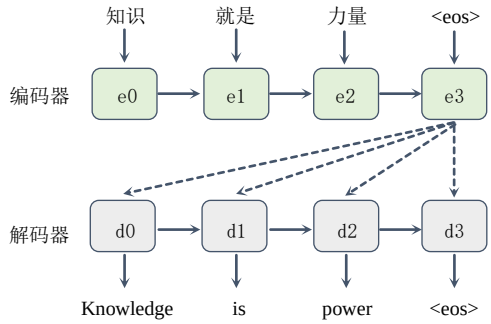


图 1 端到端神经机器翻译，隐状态 e3 作为句子向量 $c$

在源端，对于句子 $X=\{x_0,x_1,x_2,x_3\}$，编码器递归地依据前一时刻隐状态 $e_{t-1}$ 和词 $x_t$ 计算当前时刻隐状态 $e_t$。直到扫描尾词 $x_n$ 随即完成了编码过程，并将最后一个隐状态 $e_n$ 作为表示源语言句子的向量 $c$，指导并约束后续解码过程。$e_t$ 的计算公式如下：

$$e_t = g(e_{t-1}, x_t) \qquad (1)$$

在目标端，解码器递归地依据向量 $c$ 和已生成的目标词 $y_{t-1}$ 以及上一时刻隐状态 $d_{t-1}$ 共同作用于当前时刻隐状态 $d_t$，如下公式所示：

$$d_t = h(d_{t-1}, y_{t-1}, c) \qquad (2)$$

得到解码器隐状态 $d_t$ 后，目标词 $y_t$ 的条件概率分布可由公式（3）得到：

$$p(y_t \mid y_{<t}, X) = soft\,max(f(d_t, y_{t-1}, c)) \qquad (3)$$

其中，$g$、$h$ 和 $f$ 为非线性函数。通过解码器递归地从左至右逐一生成目标词，最终得到完整译文 $Y=\{y_0,y_1,y_2,y_3\}$。尽管引入长短期记忆的神经机器翻译性能上获得大幅提升，却面临着实现准确编码的挑战。因为不论句子长短，编码器都要将其映射为一个固定维度的向量。

针对上述问题，Bengio 等[13]提出了基于注意力（Attention）的神经机器翻译。解码器在生成目标词 $y_i$ 时，动态地注意源语言句中与之相关的上下文 $c_i$，而不再关注整个源语言句子。图 2 给出了引入注意力机制的神经机器翻译模型。

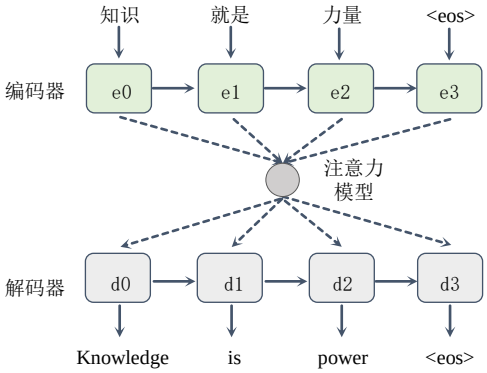


图 2 基于注意力的神经机器翻译，动态生成上下文向量 $c$

引入注意力的神经机器翻译的关键在于基于注意力的上下文向量 $c$ 的生成。当前时刻待生成词 $y_t$ 在源端对应的上下文向量 $c_t$ 由源语言隐状态序列 $e=\{e_0,e_1,e_2,e_3\}$ 和注意力权重 $a_t$ 加权求和得到，而注意力权重 $a_t$ 由上一时刻解码器隐状态 $d_{t-1}$ 和源端隐状态 $e_j$ 共同作用产生。如下公式所示：

$$c_t = \sum_{j=1}^{n} a_{t,j} e_j \qquad (4)$$

$$a_{t,j} = \frac{\exp(b_{t,j})}{\sum_{k=1}^{n} \exp(b_{t,k})} \qquad (5)$$

$$b_{t,j} = m(d_{t-1}, e_j) \qquad (6)$$

其中，$m$ 为非线性函数。得到当前时刻上下文向量 $c_t$ 后，当前时刻解码器隐状态 $d_t$ 与待生成词 $y_i$ 的条件概率分布分别可分别由公式（2）和公式（3）求解。

尽管 LSTM 和 Attention 机制的引入能够更好的处理长距离依赖，从而提升神经机器翻译的性能。然而，自然语言中的句子长短不一、结构复杂，通过单一神经网络学习翻译知识的方法受到限制。如何利用语言学知识结合分治策略对句子的各部分进行分治与整合，是一个值得研究的问题。

## 3 基于 MNP 分治策略的神经机器翻译

尽管神经机器翻译近年来获得了迅速发展，但目前的方法主要是从数据中自动学

习翻译知识，没有充分利用语言学知识显式地指导翻译过程。并且，神经机器翻译使用固定维度的向量表示变化长度的词句，造成结构复杂的长句翻译效果不佳。

针对上述问题，本文提出一种基于 MNP 分治策略的神经机器翻译。该方法主要基于分治法的思想，采用一个“抽取-翻译-重组”的 MNP 处理框架，将单个复杂长句的翻译问题，转化为一个或多个携带子句信息的 MNP 和维系主干信息的句子框架的翻译问题，以实现翻译性能的整体提升。

### 3.1 “抽取-翻译-重组”框架

在分治策略中，通常将单个复杂问题转化为多个相对简单的问题，并分而治之。鉴于 MNP 在句中使用频率高、句法功能丰富以及边界易于识别等事实，本文主要基于“抽取-翻译-重组”的 MNP 处理框架以实现分治策略的神经机器翻译。表 1 给出了该方法的完整示例。

表 1 “抽取-翻译-重组”框架示例

| | | |
|---|---|---|
| 抽取 | 原句 | 儿童 基金会 为 难民营 中 成千上万 个 流离失所 家庭 发放 了 紧急 现金 援助 。 |
| | 短语结构句法分析 | ( (IP (IP (NP (NN 儿童) (NN 基金会)) (VP (PP (P 为) (NP (LCP (NP (NN 难民营)) (LC 中)) (QP (CD 成千上万) (CLP (M 个))) (NP (NN 流离失所) (NN 家庭)))) (VP (VV 发放) (AS 了) (NP (ADJP (JJ 紧急)) (NP (NN 现金) (NN 援助)))))) (PU 。) ) ) |
| | 句子框架与MNP | 主干：MNP1 为 MNP2 发放 了 MNP3 。<br>MNP1：儿童 基金会<br>MNP2：难民营 中 成千上万 个 流离失所 家庭<br>MNP3：紧急 现金 援助 |
| 翻译 | 句子框架与MNP译文 | 主干：MNP1 has provided MNP2 to MNP3 .<br>MNP1：Children's Fund<br>MNP2：thousands of internally displaced families in camps<br>MNP3：emergency cash assistance |
| 重组 | 完整译文 | Children's Fund has provided emergency cash assistance to thousands of internally displaced families in camps . |

在示例中，抽取 MNP 时在句子框架中保留特殊标识“MNPi”(i=1,2,...)。作为对比，本文还使用了在句子框架中保留 MNP 核心词的方法。将在本章 3.3 节、3.4 节和 3.5 节中逐一说明“抽取-翻译-重组”框架的三个步骤，并对抽取 MNP 时保留特殊标识或 MNP 核心词的方法作出详细论述。

### 3.2 双语 MNP 语料库的构建

本文采用神经机器翻译系统分别对 MNP 和句子框架进行翻译，因此，双语 MNP 语料库的构建是其中重要的一个环节。为保证训练和测试过程中 MNP 的抽取规则一致，本文没有使用双语 MNP 对齐算法进行抽取，而是采用一个“抽取+查表”的方法。步骤描述如下：

(1) 使用分析器对源语言句子进行短语结构句法分析，依据标记匹配和括号对齐等规则的方法抽取 MNP；

(2) 训练并查找短语表，匹配其中与源语言 MNP 对齐分值最高的目标语言 MNP。

上述方法的优势在于每一步都可以加入规则条件，以获得较高质量的双语 MNP。本文在抽取源语言 MNP 以及查找短语表匹配其对应的目标语言 MNP 时，过滤掉长度小于 2 或包含符号、标点等特殊字符的 MNP。得到双语 MNP 后，神经机器翻译系统的训练和测试过程如下：

首先，将双语 MNP 分别加入训练数据集和开发数据集中，利用扩展后的数据集训练神经机器翻译模型。这一做法旨在得到能同时翻译句子和 MNP 的神经机器翻译模型；

其次，对测试数据集进行同样的短语结构句法分析，抽取 MNP 的同时在句子框架中保留特殊标识或 MNP 核心词。

最后，分别对句子框架和 MNP 进行翻译，将译文重新组合以得到原句的完整翻译。

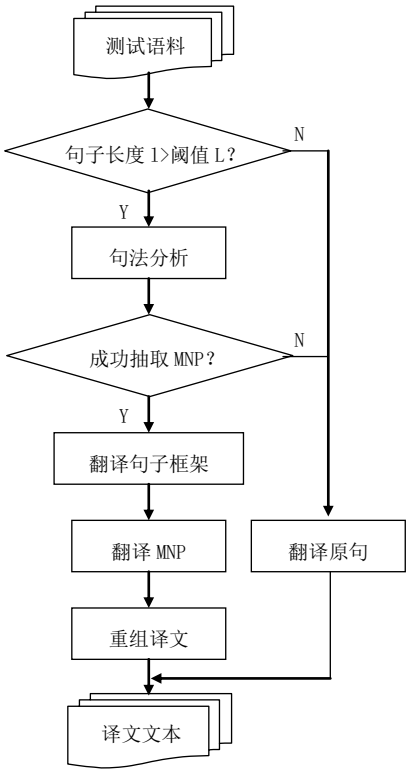


图 3 神经机器翻译系统的“抽取-翻译-重组”过程

图3给出了基于“抽取-翻译-重组”框架的神经机器翻译系统的翻译过程。考虑到短句子译文质量原本较高，本文只对长度超过阈值 $L$ 且可成功抽取 MNP 的句子采用基于“抽取-翻译-重组”框架的分治策略进行处理。

### 3.3 抽取

抽取过程的核心任务是对句子进行短语结构句法分析。考虑到抽取较短的 MNP 对缩减句子长度、降低句子结构复杂度影响较小。因此，本文只对长度不小于 2 的 MNP 进行抽取。

抽取过程的另一个重要问题是，抽取 MNP 时在句子框架中保留何种标记以实现更好的分治效果。本文主要尝试以下两种保留标记的方法：

方法一：采用“MNPi”(i=1,2,...)作为句子框架中的特殊标识，以保留 MNP 与句子框架中标记的对齐关系。

方法二：将 MNP 的核心词保留在句子框架中。通常，MNP 的尾词为其核心词。

两种方法各有其优势和不足：方法一尽管可以保留 MNP 和句子框架译文的对齐关系，为后续的译文重组过程带来积极影响，但是将“MNPi”保留在句子框架中破坏了句子的流畅度，甚至改变了原本含义。相反地，方法二在句子框架中保留核心词，保证了流畅度和语义完整性，从而能够获得较好的句子框架译文。然而，核心词却无法直接对齐到句子框架译文中的相应位置。为此，需额外训练词对齐信息，以在句子框架译文中匹配核心词译文，对其进行替换。

### 3.4 翻译

采用双语MNP扩展后的平行语料可训练得到神经机器翻译模型。图 4 给出了神经机器翻译模型采用分治策略，对句法树中的句子框架和 MNP 进行“分治”翻译的过程。其中，下侧虚线方框表示神经机器翻译模型对 MNP“*流离失所 家庭*”与“*现金 救助*”的翻译，上侧虚线方框给出了对保留特殊标识或核心词的句子框架的翻译。

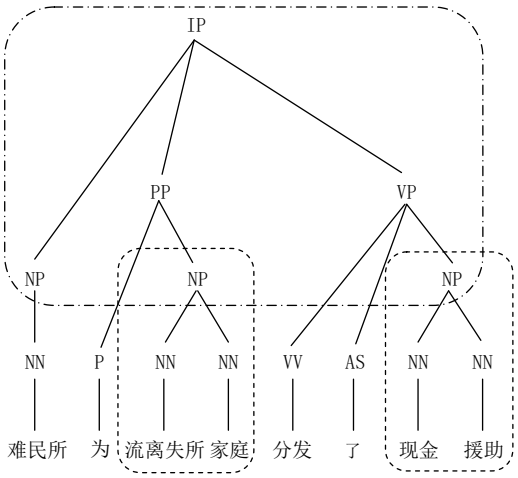


图 4 神经机器翻译模型对 MNP 及句子框架的“分治”翻译

### 3.5 重组

重组过程主要是对句子框架和MNP的译文进行重新组合，即将 MNP 译文替换到句子框架译文中的相应位置，以获得完整译文。根据 MNP 抽取时保留的特殊标识不同，重组过程中也包含以下两种方法：

方法 1：使用第 i 个 MNP 译文替换句子框架译文中的特殊标识“MNPi”；

方法 2：通过预先训练得到的词对齐信息查找 MNP 核心词的可能译文，当译文出现在句子框架译文中时，对其进行替换。

## 4 实验

### 4.1 语料说明

本文实验主要针对中英翻译任务，语料来源于联合国语料库[2]中的中英双语平行语料。其中，训练数据集共 15,886,041 句，实验过程只随机抽取部分语料。官方开发数据集和测试数据集各 4,000 句。

针对双语 MNP 语料库的构建问题，本文随机从训练语料中抽取150,000句中英双语平行句对。首先，采用 Berkeley Parser[3]对长度超过阈值 L=15 的中文句子进行句法分析，采用 NiuTrans[4]开源系统训练短语表。然后，依据3.2节所述抽取方法和过滤规则，抽取中文 MNP，并在短语表中查找其对应英文 MNP，对不符合条件的双语 MNP 进行过滤。最后，使用双语 MNP 扩展训练数据集和开发数据集。表 2 给出了实验数据的相关信息。

[2] https://conferences.unite.un.org/UNCorpus
[3] https://github.com/slavpetrov/berkeleyparser

表 2 训练数据集与开发数据集

| 信息 | Train | Dev |
|---|---|---|
| 双语句对 | 702,490 | 4,000 |
| 双语 MNP | 297,510 | 1,694 |
| 扩展后源语言句子平均长度 | 19.72 | 19.76 |
| 扩展后目标语言句子平均长度 | 22.19 | 22.21 |

针对测试语料，同样采用 Berkeley Parser 对长度超过阈值 L=15 的句子进行句法分析，并使用标记匹配和括号对齐等规则的方法抽取 MNP。表 3 给出了测试语料的相关信息。

表 3 测试数据集

| 信息 | Test |
|---|---|
| 句子数 | 4,000 |
| 成功抽取 MNP 的句子数 | 1,924 |
| 句子框架数 | 1,924 |
| MNP 数 | 5,001 |
| 句子平均长度 | 25.86 |
| 成功抽取 MNP 的句子平均长度 | 39.24 |
| 句子框架平均长度 | 19.60 |
| MNP 平均长度 | 12.14 |

从表 3 可以看出，相比于成功抽取出 MNP 的句子平均长度，MNP 和句子框架的平均长度分别缩短了 19.64 和 27.10。

## 4.2 参数设置

本文主要在深度学习框架 Theano 上采用 DL4MT[5]开源代码，搭建基于注意力机制的神经机器翻译系统。表 4 给出了实验中神经网络的主要参数设置及部分说明。

表 4 网络参数设置

| 参数 | 值 | 说明 |
|---|---|---|
| 网络层数 | 3 | - |
| 隐藏层节点数 | 1024 | - |
| 词向量维度 | 600 | - |
| Mini-batch 大小 | 64 | - |
| 学习率 | 0.5 | 由损失控制衰减 |
| 学习率衰减因子 | 0.99 | - |
| 源语言词表大小 | 80000 | 包含 eos 和 UNK |
| 目标语言词表大小 | 60000 | 包含 eos 和 UNK |
| 结点类型 | GRU | - |

表中，eos 和 UNK 是置于词表首位的特殊词，将 eos 追加在句尾，表示句子结束。当编码器扫描到 eos 时结束编码，同样地，当解码器生成目标词 eos 时终止解码过程。由于网络训练过程中 softmax 函数的计算复杂度与词表规模成正相关，因此词表大小受到限制。考虑到集外词对神经机器翻译系统的性能影响较大[14]，本文将集外词统一替换为特殊词 UNK。

在网络训练过程中，本文采用随机梯度下降（Stochastic Gradient Descent，SGD）算法进行参数更新。模型测试时，采用束搜索（Beam Search）算法生成最终译文，束大小设置为 10。

## 4.3 结果与分析

### 4.3.1 MNP 抽取

本文采用一种基于 MNP 分治策略的神经机器翻译方法，因此，能否准确识别 MNP 直接影响到系统的翻译性能。本文从成功抽取 MNP 的 1924 个句子中随机抽取 200 句，并对句中的 MNP 进行人工标注。通过比对系统的 MNP 抽取结果和人工标注结果，可计算得到系统 MNP 识别的准确率、召回率、F 值，如表 5 所示。

表 5 MNP 识别结果

| 准确率 | 召回率 | F 值 |
|---|---|---|
| 73.59% | 72.42% | 73.01% |

由表 5 可以看出，约 27%的 MNP 识别存在错误。但边界错误的 MNP 并不完全给后续的翻译过程造成消极影响[5]。

### 4.3.2 句长敏感度

为验证句子长度对于译文质量的影响，本文分别在基线系统和 MNP 分治系统上，对测试数据集中的句子按照不同的长度分布进行测试。其中，基线系统指未采用“抽取-翻译-重组”的 MNP 处理框架的神经机器翻译系统。MNP 分治系统包含两种方法，即抽取 MNP 时在句子框架中保留特殊标识“MNPi”与 MNP 核心词的方法。

本文采用 NiuTrans 开源系统中集成的大小写不敏感的 4-gram BLEU 方法对译文质量进行自动评价。如图 5 所示，横坐标表示不同句长分布，纵坐标表示译文 BLEU 分值。

[4] http://www.niutrans.com/niutrans/NiuTrans.html
[5] https://github.com/nyu-dl/dl4mt-tutorial/

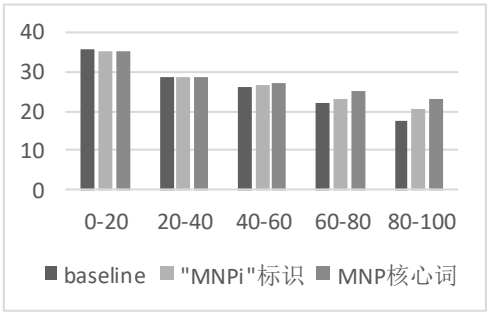


图 5 系统在不同句长分布上的翻译性能

由图5可以看出,随着句子长度的增加,译文质量呈明显下降趋势。特别地，当句子长度超过 20 后译文质量显著下降，基线系统的译文 BLEU 分值下降了 7.23，保留特殊标识“MNPi”方法和 MNP 核心词方法的译文 BLEU 分值分别下降了 6.55 和 6.31。

具体来看，主要有三点结论：⑴当句长小于 20 时，基线系统略优于 MNP 分治系统。原因分析如下：首先，神经机器翻译方法原本在短句上翻译性能较好。其次，MNP 分治系统在“抽取-翻译-重组”框架的三个步骤中都存在一定的损失，当这种损失与分治方法带来的提升持平时，分治系统的优势表现得并不明显。⑵当句长超过 20 后，随着句子长度的增大，MNP 分治系统越来越表现出更优的翻译性能。尤其当句长在 80 和 100 之间时，相比于基线系统，保留特殊标识“MNPi”和保留 MNP 核心词的方法译文 BLEU 分值分别提升了 3.10 和 5.75。⑶保留 MNP 核心词的方法在翻译性能上优于保留特殊标识“MNPi”的方法，且随着句长的增大，优势愈发明显。

4.3.3 翻译性能

本文采用“抽取-翻译-重组”的MNP处理框架，对句子进行短语结构句法分析后抽取 MNP，并保留特殊标识或 MNP 核心词与其他部分组成句子框架。表6给出了基线系统、保留特殊标识“MNPi”以及保留MNP核心词的三种神经机器翻译系统的译文质量。

表 6 译文质量对比

| | BLEU（%） |
|---|---|
| 基线系统 | 28.87 |
| 保留“MNPi” | 29.23 |
| 保留 MNP 核心词 | 29.76 |

由表 6 可以看出，在基于“抽取-翻译-重组”的 MNP 处理框架上，抽取 MNP 时保留特殊标识“MNPi”和保留MNP核心词的方法在基线系统的基础上，都获得一定的提升。相比于基线系统,保留“MNPi”的方法 BLEU 分值提升了 0.36，保留 MNP 核心词的方法 BLEU 分值提升了 0.89。

在分治系统中，由于抽取 MNP 时在句子框架中保留了 MNP 的核心词，在一定程度上提高了句子框架的流畅度和语义完整性，从而相比于保留“MNPi”，表现出更好的性能，译文的 BLEU 分值提升了 0.53。

## 5 总结与展望

本文针对当前神经机器翻译方法的译文质量对句子长度敏感的问题，提出一种基于 MNP 分治策略的神经机器翻译。依据组块分析和分治法的思想，对长句进行 MNP 识别和抽取，进一步对 MNP 和句子框架进行独立翻译，从而在一定程度上缓解了神经机器翻译对句子长度敏感的问题。

实验结果表明，该方法通过对训练数据的扩展、翻译前对 MNP 的识别和抽取、翻译时对 MNP 和句子框架的分而治之、翻译后对译文的重组等策略给神经机器翻译带来积极的影响。相对基线系统的方法，BLEU 分值提升了 0.89。

然而，该方法在 MNP 抽取，句子框架与 MNP 的译文重组等方面都存在一定的损失，并且，诸如目标语言 MNP 的单复数等问题尚待解决。下一步研究工作的重心拟定在以下两个方面：首先，将该方法泛化到其他类型的短语结构，以对目前方法做进一步扩充。其次，因为过程中涉及对句子的拆分与整合，下一步应更多的从语言学角度重新思考“抽取-翻译-重组”的分治策略，以采取更优的方法。

## 参考文献

[1] Zhang J, Zong C. Deep Neural Networks in Machine Translation: An Overview[J]. IEEE Intelligent Systems, 2015, 30(5):16-25.

[2] Cho K, Merrienboer B V, Bahdanau D, et al. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches[J]. Computer Science, 2014.

[3] 蔡东风，赵奇猛，饶齐等. 基于马尔科夫逻辑网的中文专利最大名词短语识别[J]. 中文信息学报, 2016, 30(4):21-28.
[4] 丁鹏. 基于双语句法短语的统计机器翻译研究[D]. 大连:大连理工大学, 2013.
[5] Ren X, Wei Y, Hu R. Simplify Sentence Structure for Improving Human Post-editing Efficiency on Chinese-to-English Patent Machine Translation[J]. Proceedings of 6th Workshp on Patent and Scientific Literature Translation (PSLT6) Miami, 2015: 33-43.
[6] Luong M T, Pham H, Manning C D. Effective Approaches to Attention-based Neural Machine Translation[J]. Computer Science, 2015.
[7] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units[J]. Computer Science, 2015.
[8] Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation[J]. 2016.
[9] Zhang J, Zong C. Bridging Neural Machine Translation and Bilingual Dictionaries[J]. 2016.
[10] 刘洋. 基于深度学习的机器翻译研究进展[J]. 中国人工智能学会通讯, 2015:28-32.
[11] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks[J]. Advances in Neural Information Processing Systems, 2014, 4:3104-3112.
[12] Graves A. Long Short-Term Memory[M]// Supervised Sequence Labelling with Recurrent Neural Networks. Springer Berlin Heidelberg, 2012:1735-1780.
[13] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.
[14] Li X, Zhang J, Zong C. Towards zero unknown word in neural machine translation[C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2016:2852-2858.

作者联系方式：
张学强 辽宁省沈阳市沈北新区道义南大街37号 沈阳航空航天大学 110136
18040032752 724814112@qq.com