• 人工智能及识别技术 • 文章编号：1000—3428(2012)03—0169—03 文献标识码：A 中图分类号：TP391

基于句法调序的汉维统计机器翻译

陈丽娟[1,2]，张 恒[1,2]，董兴华[1,2]，吐尔洪 · 吾司曼[1]，周俊林[3]

(1. 中国科学院新疆理化技术研究所，乌鲁木齐 830011；2. 中国科学院研究生院，北京 100049；3. 中国科学院新疆分院，乌鲁木齐 830011)

摘 要：在汉语到维语的统计机器翻译中，2 种语言在形态学及语序上差异较大，导致未知词较多，且产生的维语译文语序混乱。针对上述问题，在对汉语和维语的语序进行研究的基础上，提出一种汉语句法调序方法，进而对维语进行形态学分析，采用基于因素的统计机器翻译系统进行验证。实验结果证明，该方法在性能上较基线系统有显著改进，BLEU 评分由 15.72 提高到 19.17。

关键词：统计机器翻译；句法调序；形态学；因素模型；翻译模型

Chinese-Uyghur Statistical Machine Translation Based on Syntactical Reordering

CHEN Li-juan[1,2], ZHANG Heng[1,2], DONG Xing-hua[1,2], Turghun Osman[1], ZHOU Jun-lin[3]

(1. Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011, China; 2. Graduate University of Chinese Academy of Science, Beijing 100049, China; 3. Xinjiang Branch of Chinese Academy of Sciences, Urumqi 830011, China)

【Abstract】Chinese and Uyghur are very different in terms of morphological typology and word order, which leads to many unknown words and confusion word order in Uyghur when translate from Chinese to Uyghur using statistical method. On the basis of the word order of Chinese and Uyghur, a Chinese syntactic reordering method is proposed, and an analysis on Uyghur morphological information is made to resolve the difficulties. Experimental results on the factor-based SMT show that the approach achieves a substantial improvement in translation quality over the baseline phrase-based system, and the BLEU score is improved from 15.72 to 19.17.

【Key words】Statistical Machine Translation(SMT); syntactical reordering; morphological; factored model; translation model

DOI: 10.3969/j.issn.1000-3428.2012.03.057

1 概述

目前，在统计机器翻译(Statistical Machine Translation, SMT)中融合句法和形态学信息的研究，受到很多研究者的关注。对于汉维 SMT，这 2 种技术尤其重要，因为汉语和维语在语序和形态学上存在较大差异。在统计机器翻译中，基于短语的统计机器翻译(PSMT)方法的性能较好[1]。在翻译时，主要依靠扭曲模型实现短语间的重排序，这种方法对于语序差异较大的语言来说并不令人满意。因此，本文将句法信息和统计方法相结合进行重排序，在一定程度上弥补了 PSMT 的不足。

2 句法调序

在统计机器翻译中，对于源语言和目标语言的语序差异，句法调序是一种有效的解决方法[2]。本文研究汉语到维语的翻译，由于目前维语的句法分析尚不成熟，因此，选择对源语言进行句法分析。表 1 显示了汉语和维语在语序上的主要差异。

表 1 汉维语序对比

句法结构	汉语	维语
句子结构	[状——修饰全句]+(定)主+[状]谓<动补>+(定)宾<宾补>	[状——修饰全句]＋(定)主＋(定)宾＋[状]谓
介词短语	介词+宾语	无介词
补语成分	中心词+补语	无补语
形态学信息	表示形态学信息的词(本文 5.3 节)+动词	词干+形态学信息词缀
主谓宾	他建议我们去公园。	ئۇ(他) "بىز(我们) باغچىغا(公园) بارايلى(去)' تەكلىپ(建议)

为了能够产生较好的维语译文，对输入的汉语句子做一些重要的调序是必要的。本文在对汉语和维语的语序进行系统研究的基础上，归纳了一系列汉语句法调序规则。为了能够实现汉语的句法调序，对源语言句子进行句法分析，对得到的句法分析树运用调序规则，进行一系列的变换。在汉维 PSMT 中，对训练和测试语料都要进行调序处理，使源语言和目标语言在语序上更加相近。对形态学信息的研究主要集中在形态学切分[3]。

3 汉维语言对比分析

目前，在政治、科技、教育等领域中，只有规模很小的汉维平行语料资源，但是对于以数据为驱动的统计机器翻译方法来说，这是不够的。在有限的汉维平行语料库的基础上，从语言学对汉语和维语进行分析和研究，以期寻找有效的方法提高汉维机器翻译的性能。对汉语和维语进行系统的分析，在语言学方面，影响汉维统计机器翻译主要有以下因素[4]：

(1)汉语和维语的语序差异较大。主要体现在，汉语为主谓宾(SVO)结构，维语为主谓宾(SOV)结构，详情见表 1。汉语和维语中形态学信息描述参见本文 5.3 节。

(2)汉语和维语的形态学差异较大[5]。汉语是孤立语，无

基金项目：中国科学院西部行动计划高新技术基金资助项目(KGCX2-YN-507)

作者简介：陈丽娟(1985－)，女，硕士研究生，主研方向：自然语言处理；张 恒，硕士研究生；董兴华，博士研究生；吐尔洪 · 吾司曼，助理研究员；周俊林，研究员

收稿日期：2011-07-25 **E-mail：**chenlijuanyx@sina.com

形态变化。维语是黏着语，以词干附加多个词缀为派生新词的主要方式，形态变化丰富且复杂。在 PSMT 系统中，生成维语动词时易出现错误。例如，系统产生的维语译文：ئۇ(他) تۈنۈگۈن(昨天) مەكتەپكە(学校) بارمىدىم(没去)。在此句中，主语是第三人称，而谓语动词是第一人称。另外，系统认为拥有相同词干的维语词之间无任何联系，可能出现一部分词被系统翻译，而另一部分词不被识别。

产生上述问题的主要原因在于，PSMT 系统不包含形态学等语言信息。在国外，针对阿拉伯语及印度语等形态变化丰富的语言的主语(S)，学者们提出了很多将形态学知识引入到机器翻译的方法[3]，作为参考，将维语形态学信息引入到汉维统计机器翻译中，研究其对系统翻译性能的影响。

为了获得更好的汉维翻译质量，研究采用如下方法改进：

(1)采用基于因素的模型，对维语词进行形态切分。

(2)对汉语句子进行句法调序，使汉维 2 种语言在语序上更加相似。本文重点阐述汉语句法调序。

系统框架如图 1 所示。

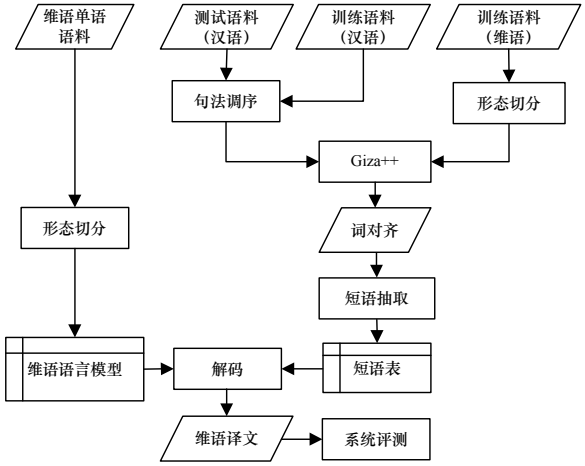

图 1 基于因素的系统框架

4 基于因素的翻译模型

基于因素的模型是在普通的对数线性模型上增加了生成模型。把词看成有词干、词性，形态学标注信息组成的一组向量。这种方法需要在训练前对语料库进行词干、词缀的切分与标注。把形态素作为基本的翻译单位参与语言模型和翻译模型的训练。这样就提高了形态素出现的概率，期望能在源语言和目标语言中找到对应的翻译项。

4.1 形态素

在统计机器翻译中，传统的表面词形的方法没有考虑到语言的形态学信息，即每个词形仅表示该词本身。例如，翻译系统将 apple 与 apples 作为 2 个无任何联系的单词对待。如果训练语料中有 apple 而没有 apples，系统可以识别并翻译 apple，但会将 apples 标记为未识别词。对形态学信息丰富的语言进行机器翻译时，这种现象尤其普遍，很大程度上增加了未识别词的数量，限制了系统翻译性能。为了解决这一问题，本文将维语表面词形切分为词干和词缀，统称为形态素。

4.2 翻译模型

基于因素的翻译模型建立在基于 PSMT 的基础上，是 PSMT 的一种扩展。它将基于短语的翻译分解为一系列的映射步骤。如单词 apples，将其切分为 apple | NN + s | plural，apple 为词干，s 为形态学信息，对词干和形态学信息分别进行翻译，然后产生目标语言表面词形。汉维翻译过程如下：

(1)对于每个维语词，如果存在于训练语料库中，采用方法(1)，如图 2 所示。用基于短语的翻译模型，不对维语进行形态素切分，以词作为最小的翻译单位。

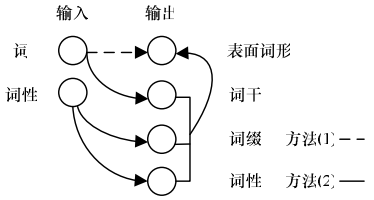

图 2 (汉维)基于因素的翻译模型

(2)对于训练语料中不存在的维语词，采用方法(2)，用基于因素的翻译模型。将语料库中的维语词切分为词干和词缀，以形态素作为最小的翻译单位。翻译时，将汉语词映射为维语词干，汉语词性映射为维语词性和词素信息，最后生成对应的维语表面词形。

5 汉语句法调序规则

本节将描述一系列汉语调序规则。使用斯坦福 PCFG parser 对汉语句子进行句法分析，该分析器使用宾州中文树库做训练语料，可以对任何类型的短语做调序，但是有些类型不需要调序规则。例如，汉、维语中定语的用法和位置相近，因此，调序时不考虑定语。宾州中文树库共使用 23 种短语标记，针对其中 3 种主要的短语结构对汉语句子进行调序：动词短语(VPs)，介词短语(PPs)和形态学信息词。本文主要使用其中 9 种标记，如表 2 所示。

表 2 宾州中文树库部分短语标记

标记	意义
VP	动词短语
PP	介词短语
NP	名词短语
IP	从句
QP	数词短语
VV	动词
AS	动词时态标记
P	介词
LCP	“XP+LC”形式的短语

5.1 动词短语调序

根据其构成方式，汉语动词短语可分为动宾短语和动补短语。汉语和维语在句子结构上存在显著差异，在汉语中，谓语一般在主语之后、宾语之前，位于句中；而在维语中，谓语一般在宾语之后，位于句末。传统的语法学认为，汉语有补语而维语没有补语。汉语中的补语，位于谓语之后，译成维语时，一般作为状语，位于谓语之前。

本文设定以下动词短语调序规则：将动词短语中的宾语及补语移置动词前。由于动词和部分附加成分(如“着”、“了”、“过”等时态标记)一起构成谓语中心词[6]，因此将时态标记和相邻的动词作为整体一起移动。

5.1.1 动宾短语调序

宾语主要由名词和代词构成，下面主要考虑 NP 和 IP 做宾语的情况。

(1)在动词短语中，当名词性短语作宾语直接跟在谓语动词后时，交换两者位置。

[VV][NP]->[NP][VV]

(2)在动词短语中，如果谓语动词后附加了时态标记 AS，将 AS 与谓语动词一起移动到宾语后。

[VV][AS][NP]->[NP][VV][AS]

(3)在动词短语中，如果谓语动词后跟从句，交换两者位置。如果谓语动词和从句间有标点，调序规则相同。IP 从句中调序参考规则 1。

[VV][IP] ->[IP][VV]

5.1.2　动补短语调序

如前文所述，汉语中有补语，而维语没有。补语主要由谓词性词语、数量短语和介词短语充当。这里只对数量短语进行分析，在下一节中，将单独进行介词短语调序规则分析。

数量词短语作补语，设定调序规则：将动词短语中数量短语修饰词移到动词之前。

[VV][QP]->[QP][VV]

5.2　介词短语

介词短语主要是介宾短语。名词性短语做宾语一般位于介词之后；而维语中没有介词的概念，多是后置词与相应实词组合，且后置词置于实词之后。汉语中的介词绝大多数是由动词虚化而来的，继承了动词的一些特征。汉语中介词翻译为维语中的动词、实义词或虚词，只考虑前 2 种情况的排序规则：交换介词及其客体的位置。

(1)在介词短语中，名词性短语作介词宾语时，交换两者位置。

[P][NP]->[NP][P]

(2)在介词短语中，从句作宾语时，交换两者位置。IP 从句中调序参考 5.1 节规则 1。

[P][IP]->[IP][P]

(3)在介词短语中，交换的标记为 P 和 LCP 结构的位置。LCP 是形如“XP+LC”形式的短语。XP 是任意类型短语，LC 是方位词。

[P][LCP]->[LCP][P]

5.3　表示形态学信息的词

维语中一般用动词的词缀表示形态学信息，比如，否定、被动、时态等，而汉语中表示形态学信息的词是分散在句中的。文献[7]描述了汉语中表示形态学信息的特征。因为在调序中只考虑词汇特征，所以构建调序规则时，只考虑能愿动词、时间副词和否定词。汉语中这些词一般位于动词性谓语之前，而在维语中，它们所对应的词一般用在谓语之后，形成合成谓语。如下例，将这些词放置到动词后面：将，要，已近，不(要)，别，没(有)。

现将所有调序规则应用于句子：他去过图书馆两次。

调序前：　他　去　过　图书馆　两次　。

调序后：　他　图书馆　两　次　去　过　。

维语译：ئۇ(他) پىخانىسىغاۋتۇك(图书馆) ئىككىلا(两) قېتىم(次) بارىدى(去：过去式)。

调序前后的实例如图 3 所示。观察表中例句，调序后的汉语句子与其维语译文语序一致。

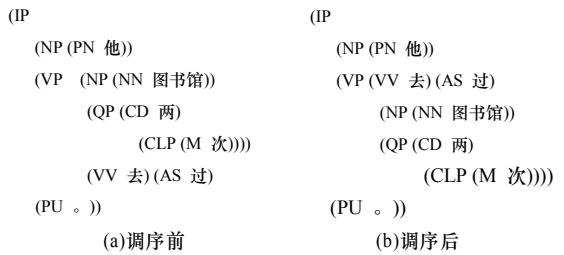

图 3　汉语句法调序实例

6　实验与分析

在收集的汉维平行语料上，如表 3 所示，本文用开源的 Moses 系统训练一个基于短语的汉维 SMT 系统，作为基线系统。使用 3-gram SRLM 语言模型，采用最小错误率训练调参，采用 BLEU 和 NIST 作为评测标准。

表 3　汉维平行语料库

维语单语	训练集	调参集	测试集
98 000 句子	50 100 句对	500 句对	450 句对

6.1　调序系统

在使用了汉语句法信息的系统中，首先要对训练集合测试集中的汉语句子进行句法分析，然后对分析结果应用调序规则。为了获得各类短语结构调序对翻译性能的影响。做了 6 组对比实验，对汉维平行语料运用本文给出的调序规则的子集，在 PSMT 系统上进行训练和测试。实验结果如表 4 所示，给出了应用不同类型的调序规则的 BLUE 评分。

表 4　不同的调序规则对翻译性能的影响

方法	BLUE 评分	NIST 评分
基线实验	0.157 2	4.173 8
基线+VP 调序	0.174 9	4.418 8
基线+PP 调序	0.163 8	4.251 4
基线+形态词调序	0.160 9	4.252 5
基线+VP+PP 调序	0.183 9	4.422 7
基线+全调序	0.185 9	4.419 3

从实验结果可以看出，对系统翻译性能最有影响的短语结构调序类型是动词短语调序，这主要是由汉维句子结构差异所决定的：汉语和维语中谓语动词的位置不同，汉语是 SVO 结构，维语是 SOV 结构。随着所用调序规则集合的扩展，系统翻译性能不断提高。当应用所有的调序规则时，系统翻译性能达到最好。

虽然汉语和维语中许多结构在语序上存在较大差异，但本文只针对汉语中部分短语结构进行了调序，主要是因为：(1)汉语和维语的句法结构有一定的相似性。(2)分析器对于某些结构的分析会产生错误。(3)有些结构在训练数据中很少出现，或者是对翻译性能的影响不大。所以只对汉维差异显著的结构进行了句法调序。

6.2　基于因素的系统

同 PSMT 系统相同，基于因素系统的训练过程也包括使用 Giza++进行汉维双向词语对齐训练，根据词语对齐抽取短语表。在基于因素的系统中，由于对维语进行了形态切分，最小翻译单位由词变成了更小的形态素，参照其他语言形态学处理对维语进行分析研究[3]。在训练前，先对汉维平行语料进行预处理：对平行语料进行分词和词性标注，用开发的维语词缀切分工具将维语词切分为词干|词缀。

参照 4.2 节中的方法，进行汉维翻译的训练。实验数据如表 5 所示，与表 4 进行对比分析，可以看出，在性能上，基于因素的系统要优于基线系统。将汉语句法调序和维语形态学信息同时融入到系统中，此时系统性能得到最大提升。

表 5　句法信息和形态学对翻译性能的影响

方法	BLUE 评分	NIST 评分
基于因素	0.177 3	4.394 4
基于因素+全调序	0.191 7	4.519 9

由于 PSMT 系统没有包含详尽的语言学信息(词性和形态学信息)，因此在有限的语料库上进行训练时，它不能够处

(下转第 175 页)