子句对齐及其在专利统计机器翻译中的应用

何彦青[1] 张 娟[2]
（1. 中国科学技术信息研究所，北京 100038；2. 北京联合大学，北京 100101）

摘 要：针对专利文献句子偏长的特点，将统计机器翻译中的训练语料进行子句切割获取双语的子句序列，再采用统计和规则相结合的策略来生成子句对齐，建立基于简单子句的双语语料来重新训练统计机器翻译系统，在一定程度上改善了原有双语训练语料中的短语对齐和词对齐，可以更为深入地利用平行语料中蕴含的翻译信息，应用于专利统计机器翻译中，在NTCIR-9的测试集上进行实验比较，获得较为满意的翻译效果。

关键词：子句对齐；词对齐；简单子句；专利文献；统计机器翻译
中图分类号：TP391 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2014.04.014

Sub-sentence Alignment and Its Application for Statistical Patent Machine Translation

He Yanqing[1], Zhang Juan[2]
(1.Institute of scientific and Technical Information of China, Beijing 100038; 2.Beijing Union University, Beijing 100101)

Abstract: For sentences in patent documents are often long, this paper tries to segment the training corpus of statistical machine translation into bilingual sub-sentence lists and uses statistical strategies and rules to obtain their sub-sentence alignment. Then new-generated training corpus based on simple sub-sentences is added into the training data to train statistical machine translation system. This method improves phrase alignment and word alignment in bilingual training corpus. It also digs translation information in parallel corpus more deeply and improves translation quality. This method was applied to statistical patent machine translation. Experiments were conducted on the test set in NTCIR-9 and a satisfactory translation result was obtained.

Keywords: sub-sentence alignment, word alignment, simple sentence, patent text, statistical machine translation

作者简介：何彦青*（1974- ），女，博士，中国科学技术信息研究所副研究员，研究方向：机器翻译，自然语言处理，机器学习；**张娟**（1975- ），女，硕士，北京联合大学讲师，主要研究方向：信息管理。

基金项目：国家自然科学基金项目“面向专利文献的统计机器翻译语境分析”（61303152）；“十二五”国家科技支撑计划课题“基于多源信息的电动汽车数据挖掘关键技术研究（2013BAG06B01）”；国家国际科技合作专项“面向科技文献的日汉双向实用型机器翻译合作研究”（2014DFA11350）。
收稿日期：2014年5月12日。

1 引言

机器翻译尤其是基于统计的机器翻译，因其具有语言无关性强、领域移植性好、知识获取方便、开发周期短等特点，已经成为专利文献翻译中不可忽视的技术。

统计机器翻译方法直接依赖语料库中的统计结果进行歧义消解处理和译文选择。近年来，多种类型的统计机器翻译模型的涌现，如基于短语的翻译模型[1−2]、基于层次短语的翻译模型[3]及基于句法的翻译模型[4−7]，推动了机器翻译技术的发展，提高了机器翻译的质量。专利文献独特的语境特征对统计机器翻译提出了更为严峻的挑战，其形式规范、语言严谨，具有一定的法律文件特性，句子结构复杂，多动词和嵌套子句等特性增加了机器翻译的难度。因此，非常有必要加强面向专利文献的统计机器翻译的研究。针对专利文献句子偏长的特点，本文尝试将统计机器翻译中的训练语料进行子句切割获取双语的子句序列，再采用统计和规则相结合的策略来生成子句对齐，建立基于简单子句的双语语料来训练统计机器翻译系统，这里称之为基于简单子句的统计机器翻译系统。这样在一定程度上改善了双语训练语料中的短语对齐和词对齐，可以更为深入地利用平行语料中蕴含的翻译信息，从而提高翻译质量。我们将该方法应用于专利统计机器翻译中，在NTCIR−9的测试集上进行了实验比较，获得了较为满意的翻译效果。

本文在第2节介绍目前国内外专利统计机器翻译的主要方法以及句子对齐的主要技术；第3节对子句对齐的关键算法以及基于简单子句的统计机器翻译系统进行介绍；第4节是实验数据的准备和实验结果；第5节给出结论。

2 相关工作

现有的专利机器翻译系统主要分为4种：基于规则的翻译系统、基于实例的翻译系统、基于统计的翻译系统和基于混合策略的翻译系统。研究人员也使用不同的混合策略来改善专利机器翻译的质量，例如融合规则翻译系统、翻译模板以及统计机器翻译系统获取质量的改善[8−10]，或者采用额外的技术来专门处理专利数据中的特殊特征[11]。其中，基于统计的翻译系统和基于混合策略的翻译系统是目前被多数采用的机器翻译系统。

统计机器翻译是机器翻译研究的主流。随着该技术不断成熟，越来越多的国内外研究人员和机构将其应用到专利文献翻译之中。Ehara将统计机器翻译系统作为规则系统的后编辑处理[8]；BBN采用串到依存树的统计机器翻译模型[11]进行汉英的专利翻译，利用专利文献的上下文信息来选择目标语言的语言模型；德国RWTH Aachen大学[12]在融合基于短语的统计机器翻译和基于层次短语的统计机器翻译的时候，在对数线性框架下使用三元组词汇（triplet lexicon）模型、判别性词汇模型以及源语言解码序列模型等多个额外模型；法国Le Mans大学用连续空间语言模型对统计机器翻译结果重新打分[13]；IBM对3个统计机器翻译系统（基于短语的翻译模型、直接翻译模型和句法翻译模型）进行系统融合时，从上下文数据中抽取平行语料加入到统计翻译模型的训练中[14]；韩国浦项工科大学将短语方法和句法辅助预处理方法串联来解决英语和日语的结构差异，再使用短语方法来处理词汇差异[15]；日本电报电话公司（NTT）和东京大学将不同的统计翻译系统进行融合，包括句法预排序、森林到树翻译模型等，还使用新闻领域的语料来加强词对齐的准确率[16−17]。国内的研究人员也将统计机器翻译用于专利翻译，国家知识产权局的在线汉英机器翻译系统整合了规则系统和统计机器翻译方法[9]；富士通在基于统计的翻译系统中加入了中文改写和括号处理的预处理方法[18]；东北大学将统计方法和基于实例的方法相结合来进行专利翻译[19]；中科院计算技术研究所在翻译专利文献时采用多系统重新打分的技术，针对专利文献加入了领域识别、人工编写的翻译模板以及化学领域化学表达式特殊处理等策略[20]；中国科学技术信息研究所采用基于词和短语的系统融合方法来融合不同

语言模型的统计机器翻译系统[21]。由此可见，众多的研究机构已经提出了多种有效方法并成功地将统计机器翻译系统应用于专利文献的翻译领域。

但是，这些专利机器翻译系统的研究重心大都在于多翻译模型的融合或者多翻译策略的实施，针对专利文献长句的处理还比较单一。我们的研究不改变统计机器翻译本身的模型，而是着重于将双语长句切割为短句，通过获取更为简短更为准确的双语子句翻译对来改善统计机器翻译。我们的关键技术在于句子对齐研究。目前，句子对齐的主要方法可以分为3类：基于长度的方法[22-23]、基于词汇的方法[24-25]以及混合的方法[26-28]。基于长度的句子对齐方法利用了句子长度等非词汇化的信息来建立统计模型，算法简单，实施效率高，能够独立于语言知识和其他外部资源，但是句子对齐精度低。基于词汇的句子对齐方法通过句对中词汇的分布或者外来资源（双语词典）中词汇的匹配信息来获取句对的匹配程度信息，能够生成精度更高的句子对齐，但由于双语词典的覆盖率有限，计算较为复杂，效率难以保障。混合的句子对齐方法可以同时考虑多种信息，例如句对的长度信息和词汇化信息，将多信息进行组合的方法有利于获得更为准确的句子对齐效果。我们的句子对齐与上述方法的不同之处在于，通过统计方法获取长句句对中的词对齐信息，利用词对齐来映射获取长句对中的子句对齐，这样无需利用词典等外来资源，既简单又实用。

3　关键算法

3.1　子句对齐算法

子句对齐算法是在已有的双语训练语料的基础上获取子句对齐，其目的是得到更为简短的子句翻译对。该算法利用统计方法获取双语长句的词对齐，然后使用标点对双语长句进行切割得到双语子句序列，再根据子句内的词对齐信息来评价子句的对齐程度，推导出初始子句对齐，最后用规则方法进行过滤得到最终的子句翻译对。这里以汉英平行语料为例，我们从NTCIR-9汉英专利机器翻译的训练语料中选择了一个长句对（图1），来说明子句对齐的4个步骤。值得注意的是，该句对并非完全互为翻译。

（1）获取词对齐。对已有的双语训练语料中的平行句子对进行训练，获得了词对齐信息。这里采用开源统计机器翻译系统Moses中基于短语的翻译系统中训练双语句对的词对齐方法。在对双语句对分别进行预处理和分词之后，将其放入Moses的训练流程中，运用Giza++训练词对齐，再使用Grow-Diag-Final式启发函数进行扩展获得最终的词对齐。Giza++使用EM算法来迭代计算双语词对的同现概率，这样依靠统计方法获得的词对齐主要依赖于双语平行语料的规模，规模越大，词对齐的准确率越高。虽然词对齐存在很多错误，但是由词对齐映射得到的子句对齐的准确率会相对高一些。图1中的词对齐由GIZA++自动训练而得。

（2）双语子句切割。双语子句切割采用一些常用的标点来进行。我们选择6个中文标点“。”“，”“？”“！”“；”“：”及相应的英文标点“.”“,”“？”“！”“；”“：”作为切割点。在这6种标点中，除了逗号外，其余5种标点基本都能表示一个语义的完结和另一个语义的开始。在此仍然选择了逗号，是因为逗号有的时候可以表示一个相对完整的语义，而且它出现的频率很高。在分别利用上述标点对中文句子和英文句子进行切割后，就获得中文的子句序列和英文的子句序列。图2给出了图1中的双语句对被切割后的子句序列，用“|||”表示分割。在这个实例中，中文由4个子句组成，英文由6个子句组成。很容易看出该双语句对中存在的错误，英文句子中的后3个子句在中文中没有对应的翻译。但是这样存在翻译错误的双语句对放入GIZA++中，训练出来的词对齐会包含这3个英文子句的错误的对齐。

（3）由词对齐映射得到初始子句对齐。得到中英文的子句序列后，我们采用图3中的算法从该双语句对的词对齐中映射获取初始的子句对

术语“可生物匹配聚合物”指的是聚合物，其，如碘复合物（加合物），与腈基丙烯酸酯组合物在哺乳类动物皮肤包括人皮肤上的体内应用是相容的。

the term “a biocompatible polymer” refers to polymers which, as iodine complexes (adducts), are compatible with in vivo applications of cyanoacrylate ester compositions onto mammalian skin including human skin. representative polymers include polyvinylpyrrolidone, copolymers comprising polyvinylpyrrolidone which is optionally crosslinked, and the like.

0-0 0-1 1-2 3-4 4-4 5-5 6-6 7-7 7-8 12-10 13-11 14-12 15-13 16-14 17-15 18-16 19-17 20-18 38-19 39-20 21-21 36-22 36-23 37-24 35-25 22-26 23-26 24-26 24-27 25-28 26-29 27-29 28-30 29-30 30-31 31-32 32-33 33-34 41-35 10-37 11-40 2-46 8-50

图1　双语句对以及词对齐

术语“可生物匹配聚合物”指的是聚合物，|||其，|||如碘复合物（加合物），|||与腈基丙烯酸酯组合物在哺乳类动物皮肤包括人皮肤上的体内应用是相容的。|||

the term “a biocompatible polymer” refers to polymers which, ||| as iodine complexes (adducts), ||| are compatible with in vivo applications of cyanoacrylate ester compositions onto mammalian skin including human skin. ||| representative polymers include polyvinylpyrrolidone, ||| copolymers comprising polyvinylpyrrolidone which is optionally crosslinked, ||| and the like. |||

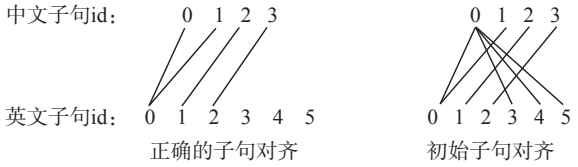
图2　中英文子句序列

齐。该算法的核心思想是：只要英文子句到中文子句中的词对齐个数占其到整个中文句子的词对齐个数的比例超过10%，就表明该子句对为对齐关系。图3中SSAlign表示初始子句对齐。根据此算法，由图1的词对齐和图2中的子句序列可以得到图4中的初始子句对齐。图4中列出了正确的子句对齐和由图3算法生成的初始子句对齐。可以看出，由于错误的词对齐，英文句子中的子句3、4和5本来为多余的内容，但是被错误对齐到中文子句0中。因此需要对初始子句对齐进行过滤。

（4）对初始子句对齐进行过滤获得最终双语子句翻译对。在生成最终双语子句翻译对的时候，对子句翻译对做如下的限制：①子句翻译对的中文或者英文必须是连续的。②子句翻译对的子句对齐必须相容，即在每一个中文子句对齐英文子句中，每一个英文子句对齐到中文子句中。如果完全遵照限制条件②，那么由于词对齐的某些错误会导致很多子句翻译对的损失，例如对于图4中的初始子句对齐，能抽取到｛c_2 <=> e_1，c_3 <=> e_2｝这2个子句翻译对，反而抽取不到｛c_0,c_1 <=> e_0｝这个子句翻译对。为此，我们放松了限制条件：对于每个英文子句片段，根据子句对齐寻找中文子句片段，如果该中文子句片段中的子句全部对齐到英文片段，则抽取该中文片段和英文片段为双语子句翻译对（这样的子句翻译对完全满足两个限制条件）；否则，如果该中文子句片段中的子句没有全部对齐到该英文片段，而是有部分中文子句对齐到了该英文子句片段以外，则计算该中文子句片段中的所有中文词对齐，计算这些词对齐到英文子句片段中的次数在它对齐到整个英文长句中次数中所占的比例，如果该比例大于0.7，就抽取该子句翻译对。在图4的例子中，子句翻译对｛c_0,c_1 <=> e_0｝的中文片段中所有中文词的词对齐总数为15，对齐到英文片段e_0中词对齐数为11，该比例约等于0.73，因而可以得到子句翻译对｛c_0,c_1 <=> e_0｝。为

输入：句子对c，e，中文子句序列$\{c_i\}$，英文子句序列$\{e_j\}$和词对齐A
输出：SSAlign
1: SSAlign＝ϕ //
2: FOR（ 每一个中文子句c_i ）{
3:　FOR（ 每一个英文子句e_j ）{
4:　in_count=0, out_count=0
5:　FOR（ e_j中每个英文词 ）{
6:　寻找该英文所对齐的中文词集合Target_word_set，
7:　FOR（ Target_word_set中每一个中文词 ）｛
8:　如果该中文词在中文子句c_i范围内，in_count++
9:　否则，out_count++｝｝
10:　计算ratio=in_count/（ in_count+out_count ）
11:　如果ratio>=0.1
12:　将子句对齐c_i ＜＝＞ e_j加入到SSAlign中 ｝｝

图3　初始子句对齐的映射算法



最终抽取的子句翻译对：
术语“可生物匹配聚合物”指的是聚合物，其，||| the term “a biocompatible polymer” refers to polymers which, 如碘复合物（加合物），||| as iodine complexes（adducts）,
与腈基丙烯酸酯组合物在哺乳类动物皮肤包括人皮肤上的体内应用是相容的。||| are compatible with in vivo applications of cyanoacrylate ester compositions onto mammalian skin including human skin.

图4　初始子句对齐

了避免重复，我们从最小的子句片段进行抽取，如果一个英文子句片段已经抽取到对应的中文翻译，包含它的更大范围的英文子句片段就不再进行抽取。这样的抽取方法也避免了对英文子句片段｛e_3, e_4, e_5｝的中文翻译的抽取，从而减少了该英文片段中的错误词对齐所导致的短语翻译对的数量。

3.2　简单子句的统计机器翻译系统

将子句对齐得到的双语子句翻译对加入到原有的训练语料当中形成新的训练语料，来训练基于短语的统计机器翻译模型，可以得到新的统计机器翻译系统，这里称之为基于简单子句的统计机器翻译系统，采用一个对数线性模型来完成最终的翻译。给定源语言句子f，翻译的过程就是搜索具有最大概率的目标翻译e：

$$e^* = \arg\max_{e} \sum_{m=1}^{M} \lambda_m h_m(e, f)$$

其中，$h_m(e, f)$为特征函数，λ_m为特征权重。每个特征函数的权重由最小错误训练算法训练得

出。

4 实验

现以NTCIR−9的汉英专利机器翻译任务为目标，对简单子句的统计机器翻译系统进行实验，并比较翻译效果。评价标准是大小写不敏感的BLEU−4[29]打分，选用最短参考答案的长度惩罚。

表1列出了所使用的所有语料的详细统计量。使用992337个平行句对作为统计机器翻译系统的初始训练语料。用于训练翻译模型的各个参数的开发集为2000个句对。测试集为2000个句对。将初始训练语料使用第3节的子句对齐方法得到子句翻译对，与初始训练语料合并在一起作为新训练集。从表1可以看出，与初始训练语料相比，汉语的平均句长减少为26个词，英文的平均句长减少为28个词。

使用Moses的基于短语的翻译模式分别对两个训练集进行训练，使用相同的开发集进行翻译参数的调整，翻译相同的测试集，表2列出了两个翻译系统的翻译结果。这里将使用初始训练语料训练的翻译系统作为Baseline，将使用新训练语料得到的翻译系统记为NewSMT。两个系统都使用了5元的语言模型[30]。实验结果表明，在开发集和测试集上，NewSMT的翻译效果都超越了Baseline。在开发集中，NewSMT的翻译结果比Baseline提高了0.26%的BLEU得分和0.58%的NIST得分。在测试集上，NewSMT的翻译结果比Bseline提高了约0.68%的BLEU得分和0.17%的NIST得分。由此可见，基于简单子句的统计机器翻译系统超越了基于短语的统计翻译系统，改善了翻译效果。这说明子句对齐算法改善了基于短语的统计机器翻译。因为将原始训练语料进行子句对齐，一方面确保了原始语料中子句之间的对齐，从而避免了其中词对齐的部分错误；另一方面在原始的长句对中，某些子句找不到对应的翻译，通过子句对齐将其去除也可以避免部分错误的词对齐。整体词对齐的改善导致了短语对齐的改善，从而保证了翻译效果的提升。

5 总结

在子句对齐中，将统计机器翻译中的训练语料进行子句切割获取双语的子句序列，采用统计方法获取词对齐，再利用词对齐获取初始子句对齐，最后使用规则抽取子句翻译对。将子句翻译对加入到原始训练语料中用来训练基于短语的专利统计机器翻译系统。在NTCIR−9的测试集上进行了实验比较，获得了较为满意的翻译效果。

表1 实验语料的统计量

数据集	语言	句子个数	词汇表	平均句长
初始训练集	中文	992337	437895	38
	英文	992337	444455	42
新训练集	中文	2886057	437895	26
	英文	2886057	444455	28
开发集	中文	2000	7465	37
	英文	2000	7897	39
测试集	中文	2000	5662	28
	英文	2000	6271	29

表2 翻译表现比较

系统	开发集		测试集	
	BLEU/%	NIST	BLEU/%	NIST
Baseline	34.62	8.6683	32.19	8.3100
NewSMT	34.71	8.7193	32.41	8.3245

目前，子句对齐算法还比较简单，未来可以改进的地方有：（1）在子句分割方面：现在仅采用了标点信息来获取子句，可以加入语法关联词来进行更符合句法意义的子句分割；（2）在子句对齐方面：只是利用了词对齐信息来评价子句对齐，可以结合词典方法来提高子句对齐的精度；（3）统计机器翻译系统：只在基于短语的统计机器翻译系统上进行应用来说明子句对齐的作用，将来还可以在基于句法的统计机器翻译系统或其他类型的翻译系统上进行后续的验证；（4）对齐粒度的信息融合：对齐可以在词级、短语级、句子级甚至篇章级的不同层次粒度上进行，不同粒度的语义对齐可以互相补益，也可以互相限制，句子对齐将来可以在语义单位的不同粒度间进行信息融合上进行更为深入的研究。

参考文献

[1] Philipp Koehn, Franz Josef Och, Daniel Maucu. Statistical Phrase−based Translation [C] //Proceeding of HLT−NAACL, Edmonton:Association for Computational Linguistics,2003: 48−54.

[2] Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. Moses: Open Source Toolkit for Statistical Machine Translation [C] // Proceedings of ACL of Demo and Poster Sessions, Prague:Association for Computational Linguistics, 2007: 177–180.

[3] David Chiang. A Hierarchical Phrase−based Model for Statistical Machine Translation [C] //Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor:Association for Computational Linguistics,2005: 263–270.

[4] Kenji Yamada, Kevin Knight. A Syntax−based Statistical Translation Model [C] //Proceedings of ACL−EACL, Toulouse:Association for Computational Linguistics ,2001: 523–530.

[5] Liu Yang, Liu Qun, Lin Shouxun. Tree−to−string Alignment Template for Statistical Machine Translation [C] //Proceedings of Coling−ACL, Sydney:Association for Computational Linguistics,2006: 609−616.

[6] Andreas Zollmann, Ashish Venugopal. Syntax Augmented Machine Translation via Chart Parsing [C].// Proceedings of the Workshop on Machine Translation, New York City:Association for Computational Linguistics,2006: 138−141.

[7] Zhang Min, Jiang Hongfei, Ai Ti AW, et al. A Tree−to−tree Alignment−based Model for Statistical Machine translation [C] // Proceedings of MT Summit XI, Copenhagen : European Association for Machine Translation,2007: 535−542.

[8] Ehara Terumasa. Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation [C] // Proceeding of MT Summit XI, Copenhagen : European Association for Machine Translation,2007:13−18.

[9] Wang Dan. Chinese to English Automatic Patent Machine Translation at SIPO [J]. World Patent Information, Elsevier Ltd ,2009, 31 (2): 137−139.

[10] 吕雅娟,付雷,董瑾,等.面向专利文献翻译的机器翻译系统的设计与实现[EB/OL].[2014−01−01]. http://www.ccf.org.cn/resources/11902017 76262/2010/05/07/2007%20Jul.Vol.5.No.4%20pp37−47.pdf.

[11] Jeff Ma, Spyros Matsoukas. BBN's Systems for the Chinese−English Sub−task of the NTCIR−9 PatentMT Machine Translation Evaluation [C] //Proceedings of NTCIR−9 Workshop Meeting, Tokyo:National Institute of Informatics,2011:579−584.

[12] Minwei Feng, Christoph Schmidt,Joern Wuebker, et al. The RWTH Aachen System for NTCIR−9 PatentMT [C]//Proceedings of NTCIR−9 Workshop Meeting, Tokyo:National Institute of Informatics,2011:600−605.

[13] Holger Schwenk, Sadaf Abdul−Rauf. LIUM's Statistical Machine Translatiion System for the NTCIR Chinese/English PatentMT [C] //Proceedings of NTCIR−9 Workshop Meeting, Tokyo:National Institute of Informatics, 2011:618−622.

[14] Lee Y, Xiang B, Zhao B, et al. IBM Chinese−to−English PatentMT System for NTCIR−9 [C] //Proceedings of NTCIR−9 Workshop Meeting, Tokyo: National Institute of Informatics, 2011.

[15] Na Hwidong, Li Jinji, Kim Sejong, et al. POSTECH's Statistical Machine Translation Systems for NTCIR−9 PatentMT Task (English−to−Japanese) [C] //Proceedings of NTCIR−9 Workshop Meeting, Tokyo: National Institute of Informatics, 2011: 652−656.

[16] Katsuhito Sudoh, Kevin Duh,Hajime Tsukada, et al. NTT−UT Statistical Machine Translation in NTCIR−9 PatentMT [C] //Proceedings of NTCIR−9 Workshop Meeting, Tokyo: National Institute of Informatics, 2011: 585−592

[17] Wu Xianchao, Matsuzaki Takuya, Tsujii Jun'ichi. SMT Systems in the University of Tokyo for NTCIR−9 PatentMT [C] //Proceedings of NTCIR−9 Workshop Meeting, Tokyo: National Institute of Informatics, 2011: 666−672.

[18] Zheng Zhongguang, Ge Naisheng, Meng Yao, et al. HPB SMT of FRDC Assisted by Paraphrasing for the NTCIR−9 PatentMT [C] //Proceedings of NTCIR−9 Workshop Meeting, Tokyo, National Institute of Informatics, 2011: 679−683.

[19] Xiao Tong, Li Qiang, Lu Qi, et al. The NiuTrans Machine Translation System for NTCIR−9 PatentMT [C] //Proceedings of NTCIR−9 Workshop Meeting, Tokyo: National Institute of Informatics, 2011: 593−599.

[20] Xiong Hao, Song Linfeng, Meng Fandong, et al. The ICT's Patent MT System Description for NTCIR−9 [C] //Proceedings of NTCIR−9 Workshop Meeting, Tokyo: National Institute of Informatics, 2011: 673−678.

[21] He Yanqing, Shi Chongde, Wang Huilin. ISTIC Statistical Machine Translation System for Patent Machine Translation in NTCIR−9 [C] //Proceedings of NTCIR−9 Workshop Meeting, Tokyo: National Institute of Informatics, 2011: 634−637.

[22] Peter F Brown, Jennifer C Lai, Robert L Mercer. Aligning Sentences in Parallel Corpora [C] //Proceedings of 29th Annual Meeting of the ACL, Berkeley: Association for Computational Linguistics, 1991: 169−176.

[23] William A Gale, Kenneth W Church.A Program for Aligning Sentences in Bilingual Corpora [C]//Proceedings of 29th Annual Meeting of the ACL, Berkeley: Association for Computational Linguistics, 1991: 177−184.

[24] Martin Kay, Martin Röchdisen. Text−translation Alignment [J]. Computational Linguistics, MIT Press Cambridge, MA, USA, 1993, 19(1): 121−142.

[25] Stanley F Chen. Aligning Sentences in Bilingual Corpora Using Lexical Information [C] // Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics, 1993: 9−16.

[26] Michel Simard,Pierre Plamondon.Bilingual Sentence Alignment: Balancing Robustness and Accuracy[J]. Machine translation, Kluwer Academic Publishers, 1998, 13(1): 59−80.

[27] 刘昕,周明,朱胜火,等.基于自动抽取词汇信息的双语句子对齐[J].计算机学报,1998,21(8):151−158.

[28] Robert C Moore. Fast and Accurate Sentence Alignment of Bilingual Corpora [J]. Machine translation:From Research to Real User, 2002, 2499: 135−144.

[29] Kishore Papineni, Salim Roukos, Todd Ward, et al. BLEU: A Method for Automatic Evaluation of Machine Translation [C]//Proceedings of the 40th Annual Meeting of the ACL, Philadelphia: Association for Computational Linguistics, 2002: 311−318.

[30] Andreas Stolcke. SRILM−An Extensible Language Modeling Toolkit [C]//Proceedings of International Conference on Spoken Language Processing, Denver: International Speech Communication Association, 2002: 901−904.

（上接第 78 页）

[8] 刘明亮,唐先明,刘纪远.基于1 km格网的空间数据尺度效应研究[J].遥感学报,2001,5(3):183−189.

[9] 国家地球系统科学数据共享平台.地球系统科学数据共享数据分类标准[S].2013.

[10] 徐冠华.实施科学数据共享,增强科技竞争力[J].中国基础科学,2003(1):5−9.

[11] 廖顺宝,蒋林.地球系统科学数据分类体系研究[J].地理科学进展,2005,24(6):93−98.

[12] 王卷乐,林海,冉盈盈,等.面向数据共享的地球系统科学数据分类探讨[J].地球科学进展,2014,29(2):265−274. DOI: 10.11867/j.issn.1001−8166.2014.02.0265.

[13] 高美荣.盐亭站1985−1994年旱地水分、养分动态观测数据集[EB/OL].[2014−06−30].http://www.geodata.cn/Portal/metadata/viewMetadata.jsp?id=100101−610041.10067.