规则和统计相结合的中文地址翻译方法

于 淼，吕雅娟，苏劲松，李贤华
中国科学 院计算技术研究所 智能信息处理重点实验室，北京 100190
E-mail: { yumiao, lvyajuan,sujinsong,lixianhua}@ict.ac.cn

摘 要：本文研究了一种规则和统计相结合的中文地址翻译方法。首先利用区划词典、关键字词典和模式表进行分词及词语类型标注，并根据词语类型划分地址单元；然后，以统计翻译模型为基础结合少量的翻译词典和人工模板对地址单元进行翻译；最后，将地址单元的翻译结果以逆序粘合在一起，形成最终译文。实验表明，利用本文的方法翻译中文地址能够取得较好的翻译效果。

关键词：中文地址；机器翻译；地址单元；统计翻译模型

Chinese Address Translation Using Combination of Rules and Statistical Method

Yu Miao, Lü Yajuan, Su Jinsong, Li Xianhua
Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing 100190
E-mail: { yumiao, lvyajuan,sujinsong,lixianhua}@ict.ac.cn

Abstract: This paper presents a method for Chinese address translation with combination of rules and statistics. First we do word segmentation and word-type tagging using division dictionary, keyword dictionary and pattern table. And then the address is chunked into several address units according to word-type. We translate address unit with a small account of dictionary and artificial template based on statistical hierarchical phrase translation model. Finally, the address unit translation results are glued together in reverse order. Experiments show that these approaches achieve good translation results.

Keywords: Chinese address; machine translation; address unit; statistical machine translation model

1 引言

随着信息技术的发展，不同语言之间的沟通和交流变得越来越重要，在此背景下机器翻译等自然语言处理技术得到长足发展。作为当前研究热点之一，命名实体翻译技术正广泛应用于诸多自然语言处理任务中，如机器翻译、跨语言信息检索等。在机器翻译中命名实体的翻译也成为影响机器翻译性能的重要因素之一。在人名的翻译任务利用音译技术基本完成之后，地址等非音译信息的翻译成为命名实体翻译研究的重点，对命名实体翻译具有重要的意义。

目前，研究地名翻译的工作很多[1][2]，但是针对完整地址翻译的工作比较少，大致可以分为两类方法。一类是基于规则的方法[3][4]，虽然这种方法翻译的准确性很高，但是规则的撰写耗时耗力。另一类是基于统计的方法，基于统计的方法[5]主要包括基于字的、基于词的以及基于短语的方法，虽然它具有很强的模型学习能力，无需人工介入，但受限于地址双语语料规模的有限性，基于统计的方法难以充分发挥它的优势，以达到较高的翻译质量。针对这些问题，本文提出了规则和统计相结合的中文地址翻译方法，实现在训练语料规模匮乏的条件下，尽可能减少人工参与，使得系统在拥有较好的模型学习能力的同时完成高质量的中文地址翻译。

本文通过深入研究中文地址的组成特点，提出了基于规则的中文地址分词及类型标注方法。通过收集总结区划词典、关键字词典、模式表，然后据此进行正向最大匹配得到中文地址分词及词语类型标注。其中区划词典是本文的一个重要工作，我们整理了大到省、市、区、县小到村、镇、乡等近 16 万词条，能够很好的切分缩略地址而且能够解决大部分切分歧义。在得到准确率较

高的分词结果之后，本文根据分词结果的词语类型设计了中文地址分块方法，从而将原始地址划分为具有逻辑独立性的小地址单元。最后，在统计层次短语翻译模型[8][9]基础上加上少量的翻译词典和人工模板以地址单元为单位进行翻译，再将翻译结果以逆序粘合在一起形成最终译文。

2 中文地址分词与类型标注

2.1 中文地址组成特点

形式上，中文地址由一系列地址单元组成，每个地址单元包含行政区划、机构名或者地址辅助信息。以下是中文地址的两个典型实例：

（1）江苏省南通市海安县海安镇海化路 28 号。

（2）北京市海淀区五道口华清嘉园 7 号楼。

（3）江苏通州市龙坝村人民路北 10 米处。

通过对大量实例进行归类、分析、总结，本文得到中文地址的 4 个组成特点：

（1）中文地址由较小的独立的地址单元组成。

（2）标准的中文地址大都由地名加关键字组成 如：“某县”、“某镇”、“某路”、“某号”。

（3）中文地址中会直接出现地名而后面省略了关键字 如：“五道口” 全称应该是“五道口街道”。

（4）有些地址单元存在复杂多变的辅助地址信息 如：“..路北..米处”。

根据以上分析，我们总结出中文地址固有的词语类型：

1. 强关键字(StrongKey)：关键字且前面必须有地名相搭配，如：“镇”、“村”、“路”。

2. 弱关键字(WeakKey)：关键字可独立出现也可前有地名相搭配，如：“山东省曲阜市小雪经济园区”中的关键字“经济园区”前有地名“小雪”相搭配，“湖南省娄底市娄星区经济园区百宝街”中的关键字“经济园区”是独立出现的，所以关键字“经济园区”是弱关键字。

3. 大行政区划(LargeDivision)：行政区划名称，如：“省”、“市”、“区”的名称。

4. 小行政区划(SmallDivision)：行政区划名称，如：“村”、“镇”、“乡”的名称。

5. 普通地名(CommonName)：未在行政区划表中出现的普通地名，如：“人民路”中的“人民”。

6. 数词(Number)：如“28 号”中的“28”。

7. 方位词(Direction)：如“南段”、“北侧”、“东口”、“西首”等。

8. 模式(Pattern)：用以表达地址辅助信息。如：“…与…交叉口”，“..路北..米处”。

2.2 行政区划表

本文整理的行政区划表以中华人民共和国行政区划代码为依据，给出了大到省、市、区、县，小到村、镇、乡、社区居委会等行政区划的名称和行政区划编码。我们从网上下载到的原始行政区划表如表1所示：

表 1 原始行政区划表

区划编码	区划名称	区划编码	区划名称
110000000000	北京市	110101001000	东华门街道办事处
110101000000	东城区	110101001001	多福巷社区居委会

（1）区划名称

我们将行政区划分为大行政区划（区划编码后六位为 0）和小行政区划（区划编码后六位不为 0）的原因是对于小行政区划的关键字有很多简称和变化，如果一一存储很费空间。如：“东华门街道办事处”可以写成“东华门街道办”，“东华门街道”，“东华门一街”，“东华门二街”等。对

于大行政区划，为了支持简称，我们既存储名称加关键字的形式又存储只有名称的形式。例如："北京市"和"北京"。经我们整理后的区划表如表2所示。

丰富的区划表有助于帮助我们解决切分歧义问题。例如："清水村镇宁西路"既可以切分成"清水村/镇宁西路"又可以切分成"清水村镇/宁西路"，已有的方法[6]是利用层次关系表，但是此例中"路"不仅隶属于"村"也隶属于"镇"，因此解决不了该切分歧义。我们的方法是利用区划表进行正向最大匹配切分，而"清水村"即在表中，歧义问题就可以有效的解决。

（2）区划编码

区划编码由12位数字组成，组织成树状层级目录，从区划编码可以看出区划之间的隶属关系。引入区划编码的优点在于可以减少切分错误。我们的区划表支持区划简称的同时也容易产生切分错误，尤其是两个字的简称很容易出错，例如："广西平南镇江滨大道"中的"镇江"容易被当成是"镇江市"的简称，因此被误切为"广西/平南/镇江/滨大道"，但是有了区划编码这种错误就不会发生，因为镇江的编码"321100000000"并不是广西的编码"450000000000"的子码。

2.3 关键字表

这里我们总结的关键字包括：强关键字、弱关键字以及方位词。表中除了有关键字的名称还有关键字的类型以及关键字前的词语类型约束(无约束用None表示)，如表3所示：

表2 整理后的行政区划表

区划编码	区划名称	区划编码	区划名称
110000000000	北京市	110101000000	东城
110000000000	北京	110101001000	东华门
110101000000	东城区	110101001001	多福巷

表3 关键字表

关键字名称	关键字类型	词语类型约束
房	StrongKey	Number
经济园区	WeakKey	None
北侧	Direction	None

这里存储关键字前的词语类型约束能够帮助我们减少切分错误，像表中的关键字"房"既有可能出现在"101房"（作为关键字）又可能出现在"鹅房街"（作为非关键字），我们通过检查"房"前面的词语是不是数字即可判断是否进行关键字切分。

2.4 模式表

中文地址辅助信息大多由数词、方位词组成。本文通过对大量的中文地址进行观察与分析，总结出如图1所示的模式表，用以刻画复杂多变的地址辅助信息，提高切分的正确率。

2.5 分词及词语类型标注方法

本文利用上述收集和总结的区划表、关键字表以及模式表对中文地址进行正向最大匹配，从而得到分词及词语类型标注结果。具体某个窗口内的切分步骤如下：

（1）查区划表

若存在于区划表中并且当前的区划编码隶属于前一个区划编码则进行切分并检查区划编码的低六位，若低六位均为0，则标记词语类型为大区划（LargeDivision），否则标记词语类型为小区划（SmallDivision）。若不存在于区划表中，则转第二步。

（2）查关键字表

若存在于关键字表中则检查关键字前的词语类型是否符合约束，若是则进行切分并标记词语类型为关键字类型，否则转第三步。

（3）查模式表

若存在于模式表中，则进行切分并标记词语类型为模式（Pattern），否则转到下一窗口。

3 中文地址单元划分方法

所谓地址单元划分，是指把已经分好词的中文地址划分成较小的独立的地址单元。之所以要进行地址单元划分是因为中文地址是按照独立地址单元逆序进行翻译，符合翻译规律。其次，基于层次短语的翻译模型对长距离调序处理不好，若按照整体翻译很难达到较高质量的译文，如：分词正确的中文地址“北京市 宣武区 广安门 北街 20 号楼”在不分块的情况下，利用层次短语模型会被翻译成“Building Guang ' anmen 20 North St, Xuanwu District, Beijing”。从例子可以看出规则在“广安门 北街 20 号楼”中都使用了，导致调序混乱，若是限制只在地址单元中使用，则不会出现这种情况，因此地址单元划分是十分必要的。本文的地址单元划分方法主要是对已分好词的词语类型进行地址单元枚举。枚举规则如下：

（1）LargeDivision

（2）SmallDivision|CommonName|Num + StrongKey +(Direction)

（3）(SmallDivision|CommonName) + WeakKey +(Direction)

（4）Pattern

其中，“|”代表或，“+”代表前后词语类型的连接，“()”代表括号里的词语类型可有可无。例如，分好词并且标记词语类型的中文地址“河南省/LargeDivision 商丘市/LargeDivision 睢阳区/LargeDivision 南京/CommonName 路/StrongKey 中段/Direction 159/Number 号/StrongKey”的地址单元划分结果为[河南省] [商丘市] [睢阳区]'[南京 路 中段] [159 号]。

4 中文地址翻译方法

层次短语翻译模型[7]是当今统计机器翻译的主流模型之一，该模型融合了传统短语翻译模型和句法翻译模型的优点，使得翻译性能相比传统短语翻译模型有了较大幅度的提高，在具有较强调序能力的同时又避免了句法分析带来的分析错误和系统负担。

中文地址结构大都是由嵌套的短语组成，非常适合用层次短语模型进行翻译。例如“人民 路 东侧”可以用“X1 路|||X1 Road”和“X1 东侧|||East of X1”两条规则翻译即可。但是由于训练语料较小，抽取出来的有用规则数量不多，所以本文在层次短语模型的基础上引入词典和人工模板对中文地址单元进行翻译。词典相当于不带变量的规则即词汇化规则，如“蒙古|||Inner Mongolia”。人工模板相当于带变量的规则，如“X1 与 X2 交叉口|||Intersection of X1 and X2”。具体词典和人工模板与 CYK 的解码融合过程如图 2 所示：

X 路北 Y 米处
X 与 Y 交叉口
甲 X 号
X 号 Y 之一

图1 模式表

```
For len = 1 to n
  For i = 1 to n
    j = i + len – 1
      Select dict_rules,template_rules,common_rules in span(i,j)
```

图2 词典和人工模板与 CYK 的融合

其中 dict_rules 代表词典，template_rules 代表人工模板，common_rules 代表从双语对齐的训练语料中抽取的规则。最后将翻译结果以逆序粘合在一起形成最终译文。

5 实验及分析

本部分对中文地址分词、地址单元划分，以及翻译三个模块的效果分别进行了测试。本文所使用的样例来自于平时收集的 12,699 条企业地址信息。

5.1 中文地址分词测试

对于分词模块，本文所使用的测试集从 12,699 条数据中随机抽取的 200 条，使用的资源有含 153719 条数据的区划表，含 633 条数据的关键字表，含 27 条数据的模式表。分别测试了分词的正确率、召回率和 F 值并和 ICTCLAS 的分词结果进行了对比。实验结果，如下表 4 所示：

表 4 分词模块测试结果

	正确率	召回率	F 值
ICTCLAS	76.39%	68.22%	72.07%
本文方法	94.53%	95.74%	95.13%

表 5 翻译模块测试结果

	开发集/测试集
ICTCLAS	0.7039/0.7133
本文分词	0.7444/0.752
本文分词+地址单元划分	0.7626/0.7732

从表 4 可以看出，和 ICTCLAS 相比，文中系统的正确率、召回率和 F 值分别提高 18.14%、27.52% 和 23.06%。分词错误大都是由一些地名特殊而导致的，例如“北京市海淀区上园村 3 号”，此处的“上园村”并不是一个真正的村，因此发生切分错误，若把“上园村”作为普通地名添加到关键字表中，则可得到正确切分。

5.2 中文地址单元划分测试

我们从 12,699 条数据中随机抽取 300 条作为测试集，先进行分词，然后通过人工挑选出 200 条分词完全正确的结果作为最终的测试。正确率为 97.71%，召回率为 98.35%，F 值为 98.02%。结果表明，分块错误多是由模式表数据不足造成，如：“乍浦 镇 雅山 西路 21 号 东起 1-5 号”被分块的结果为[乍浦 镇] [雅山 西路] [21 号 东起] [1-5 号]，正确的分块为[乍浦 镇] [雅山 西路] [21 号] [东起 1-5 号]。此例方位词“东起”被划分到与“21 号”为一块，因为方位词大都修饰前面的词语，但此例是个特例，因此若在模式表中加入“东起 X1 号”即可得到正确分块。

5.3 中文地址翻译测试

本文将 12,699 条数据随机分成三份，分别作为训练集（10162 句）、开发集（1270 句）和测试集（1267 句）。使用 srilm 工具训练四元语言模型。本文共做了三组测试，第一组用 ICTCLAS 分词，解码用普通的层次短语模型。第二组用本文的分词方法，解码也是普通的层次短语模型，第三组用本文的分词加地址单元划分方法，解码用引入词典和人工模板的层次短语模型，其中词典含 9656 个词条，人工模板含 490 个模板。采用 BLEU 评分作为翻译性能的评价指标，结果如表 5。

从结果可以看出，本文分词比 ICTCLAS 分词更加准确，抽取的规则更好，从而提高了翻译质量。而在分词的基础上进行地址单元划分，改善了句子的长距离调序能力。

本文对翻译错误以及翻译不准确的地址实例进行分析，发现错误主要有以下几种类型：

（1）分词或地址单元划分错误

中文地址：鹤岗市工农区东解放路南三道街 9 号

参考答案：No.9, Nansandao Street, Jiefang Road, East of Gongnong District, Hegang

翻译结果：No.9, Nansandao Street, Dongjiefang Road, Gongnong District, Hegang

由于分词时“_东_”未被切分，所以导致翻译错误。由于一个字的方位词“东”、“南”、“西”、“北”容易引起切分歧义，所以本文未把他们加入到关键字表中，对他们的切分处理的并不好。

（2）词典或人工模板不足

6 总结及未来的工作

本文针对中文地址的组成特点，提出了规则和统计相结合的中文地址翻译方法。首先利用区

划词典、关键字词典以及模式表进行正向最大匹配分词并给出每个词的词语类型。在分词结果的基础上本文又给出了根据词语类型进行地址单元划分的方法，从而将原始地址划分为具有逻辑独立性的小地址单元。最后，在层次短语模型基础上加上少量的翻译词典和人工模板以地址单元为单位进行翻译，再将翻译结果以逆序粘合在一起形成最终译文。实验表明，本文提出的中文分词、地址单元划分以及翻译方法都十分有效，能够完成高质量的中文地址翻译。

为了进一步提高翻译性能，还需要把中文地址分词做的更加细致，尽量消除切分歧义并且考虑自动获取关键字、模式、翻译词典的方法，扩大我们现有的资源。

参 考 文 献

[1] 孙乐乐. 中文地名翻译浅谈. 科技经济市场，2006，p358-359.

[2] 彭月华，张向京. 地名短语文化内涵与翻译. 中国科技翻译，2008，p54-57.

[3] 王松，中文机构名称及地址的汉英翻译方法研究. 哈尔滨工业大学硕士学位论文. 2008：57：67.

[4] 苗文彦，高精度中文机构名称与地址机译策略研究. 哈尔滨工业大学硕士学位论文. 2009.

[5] 王松，杨沐昀，赵铁军. 基于统计的命名实体翻译. 第七届中文信息处理国际会议论文集. 2007, 10.

[6] 苗文彦，赵铁军，周向根等. 面向机器翻译的中文机构地址切分方法. 计算机科学. 2009，(04).

[7] David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In proceedings of ACL 2005.