

# 英汉机器翻译中英语单双重介词汉译

关晓薇

GUAN Xiao-wei

大连理工大学 计算机科学与工程系, 辽宁 大连 116023

Department of Computer Science and Engineering, Dalian University of Technology, Dalian, Liaoning 116023, China

E-mail: angel\_gxw@yahoo.com.cn

**GUAN Xiao-wei. Chinese translation of English single and double prepositions based on semantic pattern extending. Computer Engineering and Applications, 2009, 45(8): 32-35.**

**Abstract:** For English-Chinese machine translation, this paper introduced the improved concept of semantic pattern and the extending forms of English single prepositions. Semantic Pattern Bases of English single and double prepositions were constructed, and a new semantic analysis and Chinese translation algorithm based on Semantic Pattern Extending for English single and double prepositions were presented. The algorithm was combined with Single and Double Preposition Set Phrase and Sentence Semantic Pattern Bases, General Double Preposition and Special Single and Double Preposition Semantic Pattern Base to resolve the problem. The test results show that this method has a good effect on resolving Chinese translation of English single and double prepositions in general and sports contexts.

**Key words:** preposition semantic pattern; semantic pattern extending; single and double prepositions; semantic analysis; Chinese translation of English prepositions

**摘要:**面向英汉机器翻译,提出改进的介词语义模式相关概念和单重介词扩展形式,构建英语单双重介词语义模式库,提出一种基于语义模式扩展的英语单双重介词语义分析和汉译算法。将该算法与单双重介词固定搭配语义模式库、介词相关句式语义模式库、一般双重介词语义模式库和特殊单双重介词语义模式库相结合,对英语单双重介词进行总体语义分析和汉译。实验表明,利用该方法能有效解决一般文体和体育体裁中的英语单双重介词汉译的问题。

**关键词:**介词语义模式;语义模式扩展;单双重介词;语义分析;英语介词汉译

**DOI:**10.3778/j.issn.1002-8331.2009.08.010 **文章编号:**1002-8331(2009)08-0032-04 **文献标识码:**A **中图分类号:**TP391

## 1 引言

在英汉机器翻译中,英语介词的多义和语义消歧问题一直是研究的重点和难点。在自然语言处理领域中,介词短语附着(PPA)长期以来一直是国内外学者们研究的重点之一。尽管在词义消歧(WSD)方面已经做了很多工作,但专门面向计算机应用的介词语义研究却不多。

Voss<sup>[1]</sup>主要研究了英语空间介词的词干意义,但没有研究介词的词汇意义。Litkowski<sup>[2]</sup>通过利用有向标注图对机读字典中的介词定义建立模型来分析英语介词语义。O'Hara等<sup>[3]</sup>对Penn Treebank和FrameNet中的介词搭配进行语义角色标注并对其分类。Alam<sup>[4]</sup>利用语境信息提出了两种决策树来判断英语介词“over”的语义。Litkowski等<sup>[5]</sup>建立了对英语介词语义和句法特征的全面描述系统。Saint-Dizier<sup>[6]</sup>构建PrepNet,对英语及其他西方语言的介词进行原始语义标注。Hartrumpf等<sup>[7]</sup>基于Multi-Net示例对德语介词短语进行自动语义分析。Boonthum等<sup>[8]</sup>利用WordNet和LCS仅对英语介词“with”进行语义消歧。

Naskar等<sup>[9]</sup>研究了英-孟加拉机器翻译中英语介词的语义翻译问题。陈强等<sup>[10]</sup>采取在词典和规则中加入介词的语义属性和上下文相关函数的方法来解决英汉机器翻译中英语介词多义问题。关晓薇等<sup>[11]</sup>提出基于语义模式分解的英语单重介词语义分析和汉译方法。目前,专门针对英汉机器翻译中的英语介词的语义分析和汉译的研究仍不多。

本文主要针对英汉机器翻译中英语单重介词和双重介词的语义分析和汉译问题,提出了利用改进的单双重介词语义模式库和语义模式扩展算法对一般体裁下的介词进行语义分析和汉译的新方法。同时专门针对体育体裁,引入特殊语义模式库。实验结果显示该方法对处理一般体裁和体育体裁中的英语单双重介词的语义分析和汉译问题较有效。

## 2 改进的介词语义模式相关概念

首先对文献[11]中有关语义模式的定义和相关概念作如下改进:

**基金项目:**国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2006AA012140);

国家自然科学基金(the National Natural Science Foundation of China under Grant No.60573014);大连理工大学创新基金。

**作者简介:**关晓薇(1979-),博士生,主要研究领域为自然语言处理,机器翻译。

**收稿日期:**2008-11-11 **修回日期:**2008-12-23

**定义 1 语法类型:**指具有某一词性的类别。用  $N$  或  $n, A, V$  分别表示名词、形容词和动词。这里  $n$  表示不直接影响介词语义确定的名词。

**定义 2 语义类型:**指具有某一语义特征的类别。用  $N_x, A_x, V_x$  来表示, 下标  $x$  可用某一语义特征来代替。例如:  $N$  交通工具为表示交通工具类名词。

**定义 3 实量:**指语义模式中固定的词语部分。在介词相关短语中将介词和与介词搭配的动词看作实量。用  $P$  代表介词,  $V$  代表动词。在具体自然语言中用具体词语表示。

**定义 4 变量:**指语义模式中可被替代的词语部分。分为语法类型变量(用语法类型表示的变量)和语义类型变量(用语义类型表示的变量)。其中语义类型变量直接影响介词语义的确。具有相同语义类型的变量可以进行相互替换, 并可以代入同一语义模式。

**定义 5 介词简单语义模式(SSP):**指由实量和语法类型变量按照一定的位置相互插入, 能够反映一定句法结构和语义信息, 并且能够进行变量代入的翻译模板。其中实量用  $V, P$  表示。

**定义 6 介词完整语义模式(CSP):**指将简单语义模式中的实量用具体词语表示, 变量用语义类型和  $n$  表示所构成的翻译模板。英语和汉语介词完整语义模式分别是将实量用英语或汉语词语表示。文中规定具有相同语义类型的词语出现在相同的完整语义模式中。

**定义 7 语义模式的语义类型(ST):**指整个语义模式所归属的语义类型。包括  $S$ (句子),  $N, B$ (状语/定语),  $T$ (时间),  $P$ (地点)等。

### 3 英语单重介词扩展形式

#### 3.1 相关定义

**定义 8 单重介词:**指在语言片断中出现的唯一介词。这里研究的单重介词主要包括: about, against, among, as, as to, at, behind, between, by, for, from, in, into, of, on, out of, over, to, towards, under, with, within, without 等 23 个简单介词。

**定义 9 双重介词:**指在语言片断中顺序出现的两个介词。

**定义 10 单重介词扩展形式:**指从  $P$  开始, 向其左或向其右逐层扩充动词、名词和形容词, 所得到的符合英语语法规则的所有可能的搭配组合形式。用  $V, N, n, A, P$  的组合来表示。

**定义 11 单重介词实际扩展形式:**指符合实际语料中英语介词真实用法的扩展形式。

**定义 12 单重介词临时扩展形式:**指除实际扩展形式以外的扩展形式。规定:  $N, n, A$  不是  $N, n$  的向左扩展必要成份;  $V, A$  不是  $V$  的向左扩展必要成份;  $N, n, A$  不是  $A$  的向左扩展必要成份。

#### 3.2 单重介词扩展过程及扩展形式

##### 3.2.1 右-左扩展

从  $P$  开始, 先向右扩展, 再向左扩展。

(1) 以  $P$  为基点向右扩展为  $PN, Pn$ ;

(2) 以  $P$  为基点向左扩展为  $NPn, NPn, nPN, nPn, VPn, VPn, APn, APn$ ;

(3) 继续向左扩展为  $VNPn, VNPn, VnPN, VnPN, nVPn, nVPn, VAPn, VAPn$ ;

(4) 继续向左扩展为  $nVNPn, nVNPn, nVnPN, nVnPN, nVPn, nVPn, nVAPn, nVAPn$ 。

共生成 11 种实际扩展形式。

##### 3.2.2 左-右扩展

从  $P$  开始, 先向左扩展, 再向右扩展。

(1) 以  $P$  为基点向左扩展为  $NP, nP, VP, AP$ ;

(2) 分别向左扩展为  $VNP, VnPN, nVP, VAP$ ;

(3) 继续向左扩展为  $nVNP, nVnPN, nVP, nVAP$ ;

(4) 继续向右扩展为  $nVNPn, nVnPN, nVnPN, nVNPn, nVPn, nVPn, nVAPn, nVAPn$ 。

共生成 10 种实际扩展形式。

综合以上两种扩展方法, 最后共生成 12 种实际扩展形式:

$PN, NPn, nVP, nVPn, nVNP, nVnPN, nVNPn, nVnPN, nVnPN, nVnPN, nVAPn$ 。

### 4 英语单双重介词语义模式库和语义模式扩展算法

#### 4.1 单双重介词固定搭配语义模式库

单重介词固定搭配包括以下几种实际扩展形式:

(1) 全部  $nVNPn$  ( $n$  take care of  $n$  照顾), 全部  $nVNPn$  ( $n$  have the call of the market 市面繁荣), 全部  $nVnPN$  ( $n$  take  $n$  into account 考虑), 全部  $nVNP$  ( $n$  take it all in 注意倾听), 全部  $nVPn$  ( $n$  abide by  $n$  遵守), 全部  $nVPN$  ( $n$  go to pieces 崩溃);

(2) 部分  $NPn$  (a kind of  $n$  一种), 部分  $PN$  (at a loss 困惑地), 部分  $nVAPn$  ( $n$  be good at  $n$  擅长)。

双重介词固定搭配包括以下几种实际扩展形式:

(1) 部分动-介-介连用:  $nVPPn$  ( $n$  get on with  $n$  继续做);

(2) 部分介-名-介连用:  $PNPn$  (in order to  $v$  为了),  $PNPn$  (on behalf of  $n$  代表);

(3) 介...介...连用:  $PnPN$  或  $PVingPVing$  (from  $n$  to  $n$  从...到... )。

固定搭配语义模式库中, 将以上的实际扩展形式中的  $V, A, P, N$  用具体单词表示作为实量,  $n$  表示成语义类型变量。例如: 语义模式  $N_n$  take care of  $N_n$ :  $N_n$  照顾  $N_n$ 。

固定搭配语义模式库存放: 英语单重介词和双重介词固定搭配 SSP 和汉语译文 SSP。

除介词固定搭配所属的实际扩展形式以外的其余的实际扩展形式, 是本文提取介词简单语义模式和语义模式扩展算法建立的基础。因此, 介词固定搭配库的引入大大降低了语义模式扩展算法的复杂度。

#### 4.2 英语单重介词主要语义模式库的构建

基于大量的英汉双语句子级对齐语料, 提取介词语义模式, 分别构建四个主要语义模式库: 介-名语义模式库, 名-介语义模式库, 形-介语义模式库, 动-介语义模式库。每一语义模式库均分别存放: 英语 SSP, 英语 CSP, 汉语 CSP, 默认语义模式, ST, 变量个数, 英语句子, 汉语句子。分别在这四个语义模式库中为每一种介词构建独立的语义模式表和对应的变量表。

##### 4.2.1 介词语义模式提取原则

语义模式中的变量可以是单词, 短语, 现在分词, 分句或句子等。将实量相同, 变量语义模式相同, 实量、变量的句法位置相同的提取成同一语义模式。汉语 CSP 则是将英语实量对应的汉语译文提取成实量, 变量不变, 实量和变量的位置遵从汉语译文词语间的句法位置。

##### 4.2.2 介-名、名-介和形-介语义模式库

介-名语义模式库: 将  $PN$  的  $P$  作为实量,  $N$  提取成语义类

型变量。

名-介语义模式库: 将  $NP_n$  的  $P$  作为实量,  $N$  提取成语义类型变量,  $n$  提取成语法类型变量。

形-介语义模式库: 将  $nVAP_n$  的  $V$  和  $P$  作为实量,  $A$  提取成语义类型变量,  $n$  提取成语法类型变量。

变量表存放: 作为语义类型变量的英语名词和形容词, 及其对应汉语译文, 变量语义类型。

#### 4.2.3 动-介语义模式库

动介短语与名词的语义搭配具有一定的特殊性, 例如: 有的同一动介短语当与不同语义类型的名词搭配时会表现出不同的语义。因此动-介语义模式中将除  $V$  和  $P$  外的  $N$  和  $n$  都提取成语义类型变量。所提取的语义模式主要属于两种实际扩展模式:  $nVnP, nVnP_n$ 。

变量表存放: 作为语义类型变量的英语名词, 及其对应汉语译文, 变量语义类型。

#### 4.3 单重介词语义模式扩展算法

步骤 1 设介词片断为  $W_1 W_2 W_3 P W_4$ ;

步骤 2 确定基点  $P$ ;

步骤 3 将  $W_3$  与名-介语义模式库匹配;

若匹配  $N$  成功, 则得到实际扩展形式为  $NP_n$ , 转步骤 7, 否则转步骤 4;

步骤 4 将  $W_3$  与形-介语义模式库匹配;

若匹配  $A$  成功, 则得到实际扩展形式为  $nVAP_n$ , 转步骤 7, 否则转步骤 5;

步骤 5 得到扩展形式为  $P$  或  $P_n$ ;

步骤 6 将  $W_2$  与动-介语义模式库匹配;

步骤 6.1 若匹配  $V$  成功, 则得到扩展形式为  $nVnP$ , 转步骤 6.2, 否则转步骤 6.3;

步骤 6.2 若  $W_4$  为空, 则得到实际扩展形式为  $nVnP$ , 转步骤 7, 否则转步骤 6.3;

步骤 6.3 得到实际扩展形式为  $nVnP_n$ , 转步骤 7;

步骤 6.4 将  $W_4$  与介-名语义模式库匹配;

步骤 6.5 若匹配  $N$  成功, 则得到实际扩展形式为  $PN$ , 转步骤 7, 否则转步骤 6.6;

步骤 6.6 显示输出失败, 转步骤 8;

步骤 7 输出实际扩展形式;

步骤 8 结束。

#### 4.4 介词相关句式语义模式库

介词常常出现在一些固定的句式中。例如: “It’s a waste of time to vote. 投票表决是浪费时间。”提取语义模式: it be a  $N$  行为 to  $V$ ;  $V$  是  $N$  行为。

库中存放: 英语介词相关句式 CSP, 汉语译文 CSP, ST, 变量个数, 英语句子, 汉语句子。

变量表存放: 作为语义类型变量的英语名词或形容词, 及其对应汉语译文, 变量语义类型。

#### 4.5 特殊单重介词语义模式库

这里只针对体育体裁的特点, 建立该体裁下特殊单重介词语义模式。例如: “ $N_{\text{比赛名}}$  in  $N_{\text{地点}}$ : 在  $N_{\text{地点}}$  举行的  $N_{\text{比赛名}}$ ”和特殊双重介词语义模式, 例如: “in  $N_{\text{比赛项目}}$  at  $N_{\text{比赛名}}$ : 在  $N_{\text{比赛}}$  的  $N_{\text{比赛项目}}$  中”。还有一种情况, 即某介词出现在一般语境和出现在体育语境的短语时, 二者语义不同, 前者只表示短语表层含义,

而后者表示深层内涵, 例如:  $N_{\lambda}$  be\* on the score sheet 在体育体裁中表示“ $N_{\lambda}$  攻入一球”。

库中存放: 英语 CSP, 汉语 CSP, ST, 变量个数, 英语句子, 汉语句子。

变量表存放: 作为语义类型变量的英语单词, 及其对应汉语译文, 变量语义类型。

#### 4.6 双重介词语义模式库

双重介词具有如下的扩展形式:

(1)  $n[V n][P_1 N_1][P_2 N_2]$ ;

(2)  $n[V[N_1 P_1 n][P_2 N_2]]$ ;

(3)  $n[V[N_1 P_1 n]P_2 N_2]$ ;

(4)  $n[V[N_1 P_1][N_2 P_2 n]]$ ;

(5)  $n[V n][P_1[N_1 P_2 n]]$ ;

(6)  $n[[V P_1 n]P_2 n]$ 。

对于(1)~(4), 从左到右依次处理每个介词。对于(5), 即  $P_1$  和  $P_2$  都与  $V$  搭配, 则将其提取成语义模式。例如: 语义模式  $n_1$  talk with  $n_2$  about  $n_3$ ;  $n_1$  与  $n_2$  谈论  $n_3$ 。建立一般双重介词语义模式库。库中存放英语 CSP, 汉语 CSP, ST, 变量个数, 英语句子, 汉语句子。

### 5 英语单双重介词语义分析和汉译

#### 5.1 基于语义模式扩展的英语单重介词语义分析算法

```
begin
while <英语句子非空>
begin
<英语单词形态还原>;
<从左到右依次标记  $P_1, P_2, \dots, P_n$ >;
<根据语义模式扩展算法依次对  $P_1, P_2, \dots, P_n$  进行扩展, 得到实际扩展形式, 即英语 SSP>;
<将变量语义类型和英语实量单词代入 SSP, 得到英语 CSP>;
<若显示输出失败, 则匹配语义模式库中相应的默认语义模式>;
<若匹配成功, 则将默认模式作为英语 CSP, 否则显示输出失败>;
end
<输出英语 CSP 或显示输出失败>;
end
```

#### 5.2 英语单重介词汉译算法

```
begin
while <英语 CSP 成功输出>
begin
<在语义模式库中搜索英语 CSP 对应的汉语 CSP>;
<取出汉语 CSP 中变量的语义类型>;
<到相应的变量库中匹配该语义类型>;
<将变量用英语单词原形代入>;
<在变量表中匹配该语义类型对应的该英语单词>;
<若成功匹配, 则输出该英语单词对应的汉语译文; 否则显示输出失败>;
end
<输出汉语 CSP 和该介词的汉语译文或显示输出失败>;
end
```

#### 5.3 英语单重介词语义分析和汉译总体算法

步骤 1 取英语句子非空;

步骤 2 将英语单词形态还原;

步骤3 从左到右依次标记  $P_1, P_2, \dots, P_n$ ;

步骤4 对  $P_1$  进行处理;

步骤5 将含  $P_1$  的句子片断与单重介词固定搭配语义模式库进行匹配;

若匹配成功, 则输出该英语固定搭配 SSP 和汉语译文 SSP, 转步骤 10, 否则转步骤 6;

步骤6 将含  $P_1$  的句子片断与介词相关句式语义模式库进行匹配;

若匹配成功, 则输出该英语句子 CSP 和汉语句子 CSP, 转步骤 10, 否则转步骤 7;

步骤7 将含  $P_1$  的句子片断与特殊单重介词语义模式库进行匹配;

若匹配成功, 则输出该英语特殊单重介词 CSP 和汉语特殊单重介词 CSP, 转步骤 10, 否则转步骤 8;

步骤8 基于语义模式扩展的英语单重介词语义分析法;

步骤9 英语单重介词汉译算法;

步骤10 结束。

## 5.4 英语双重介词语义分析和汉译总体算法

步骤1 当英语单重介词语义分析和汉译总体算法处理  $P_1$  结束;

步骤2 若显示输出失败, 则转步骤 3, 否则转步骤 4;

步骤3 将含  $P_1$  和  $P_2$  的句子片断与双重介词固定搭配语义模式库进行匹配;

若匹配成功, 则输出该英语双重介词固定搭配 SSP 和汉语译文 SSP, 转步骤 6, 否则转步骤 4;

步骤4 将含  $P_1$  和  $P_2$  的句子片断与特殊双重介词语义模式库进行匹配;

若匹配成功, 则输出该英语 CSP 和汉语 CSP, 转步骤 6, 否则转步骤 5;

步骤5 开始英语单重介词语义分析和汉译总体算法处理  $P_2$ ;

步骤6 结束。

对  $P_3, P_4, \dots, P_n$  的语义分析和汉译算法同理可推。

## 6 实验

### 6.1 实验设置

从“英汉对照例句集”(共 400 句)(一般体裁语料)和“奥运会”(前 300 句)(体育体裁语料)的双语语料进行开放测试。人工统计实验结果。

表 1 显示了测试语料的基本信息。

表 1 测试语料

|        | 语料      | 句对总数 | 含介词句数 | 介词出现次数 |
|--------|---------|------|-------|--------|
| 开放测试 1 | 英汉对照例句集 | 400  | 229   | 306    |
| 开放测试 2 | 奥运会     | 300  | 270   | 717    |

表 2 显示了开放测试 1 和开放测试 2 的结果。表中第一行中的①表示介词翻译结果正确, 且与原汉语译文相同; ②表示由于英汉句对不对齐, 翻译结果与原汉语译文不同, 但经人工判断也是正确翻译, 不改变句义; ③表示未成功匹配各语义模式库, 显示输出失败, 即翻译失败; ④表示未成功匹配变量表, 显示输出失败, 即翻译失败; ⑤表示虽成功输出结果, 但经人工判断是错误翻译。

表 2 开放测试结果

|        | ①   | ②  | ③  | ④ | ⑤  | 正确率/(%) |
|--------|-----|----|----|---|----|---------|
| 开放测试 1 | 283 | 10 | 8  | 2 | 3  | 95.75   |
| 开放测试 2 | 624 | 49 | 18 | 5 | 21 | 93.86   |

$$\text{正确率} = [(\text{①} + \text{②}) / (\text{①} + \text{②} + \text{③} + \text{④} + \text{⑤})] * 100\%$$

## 6.2 实验结果分析

对实验结果进行分析, 发现导致翻译失败有以下原因:

(1) 语义模式不全。例如: “She was with child. 她怀孕了。”被按照语义模式“ $N_A$  be with  $N_A$ ”错误地译成“她带了孩子。”其实该句对应的英语句子是“*She was with a child.*”所以应该添加新语义模式“ $N_A$  be with child;  $N_A$  怀孕了”。

(2) 句子中的动-介短语被其他介词短语分开。例如: “I heard[over the weekend] of a terrible fuss.”中的“hear”和“of”被分开, 导致翻译错误。

(3) 动-介短语中的动词被省略。例如: “I was not beaten by him, but by her. 我不是被他打了, 而是被她打了。”中的第二个“by”前省略了“was not beaten”, 导致翻译错误。

(4) 在体育体裁语料中, 由于变量表中的变量单词不全造成翻译失败。另外“with”的使用非常广泛, 常常不用翻译出来或翻译方法灵活, 因此必须增加“with”的语义模式。

## 7 结语

本文提出了改进的语义模式的概念, 归纳了英语单双重介词的扩展形式; 介绍了单双重介词主要语义模式库的构建过程, 并提出了一种基于语义模式扩展的介词语义分析和汉译算法。将该算法与单双重介词固定搭配语义模式库、介词相关句式语义模式库、一般双重介词语义模式库相结合, 提出介词语义分析和汉译总体算法。实验结果证明该方法能够有效地解决一般语境下英语单双重介词汉译的问题。另外, 引入体育体裁特殊单双重介词语义模式库, 使得该方法更适于解决体育语境下英语单双重介词汉译的问题。

为进一步提高正确率, 需在将来的工作中完善各语义模式库和体育体裁特殊单双重介词语义模式库, 并建立除体育体裁外的其他特定语境下的特殊单双重介词语义模式库。

## 参考文献:

- [1] Voss C R. Interlingua-based machine translation of spatial expressions[D]. Baltimore: University of Maryland, 2002.
- [2] Litkowski K C. Digraph analysis of dictionary preposition definitions[C]//Proceedings of the ACL Workshop on Word Sense Disambiguation, Philadelphia, Pennsylvania, 2002: 9-16.
- [3] O'Hara T, Wiebe J. Preposition semantic classification via Penn Treebank and FrameNet[C]//Proceedings of the Computational Natural Language Learning (CoNLL-2003), Edmonton, Canada, 2003.
- [4] Alam Y S. Decision trees for sense disambiguation of prepositions: Case of over[C]//HLTNAACCL, Computational Lexical Semantic Workshop, Boston, 2004: 52-59.
- [5] Litkowski K C, Hargraves O. The preposition project[C]//Proceedings of the 2nd ACLSIGSEM Workshop on Prepositions, Colchester, UK, 2005.
- [6] Saint-Dizier P. PrepNet: A framework for describing prepositions: Preliminary investigation results[C]//Proceedings of IWCS 6, Tilburg, 2005.