

基于语义模式的英语动词汉译

张 瑶, 黄德根

(大连理工大学计算机科学与技术学院, 辽宁 大连 116023)

摘 要: 英语动词正确汉译是机器翻译的难点。为此, 提出一种基于动词语义模式库、固定句式库和变量库的英语动词汉译算法。介绍语义模式的提取方法, 构建常用动词的语义模式库、固定句式库和变量库, 归纳动词语义模式的3种结构形式: V语义模式, VO相邻语义模式和VO非相邻语义模式。动词语义模式库由上述3种结构形式对应的模式库构成。实验结果表明, 该方法能有效解决英语常用动词的汉译问题。

关键词: 机器翻译; 语义类型; 动词; 语义模式; 变量

Chinese Translation of English Verb Based on Semantic Pattern

ZHANG Yao, HUANG De-gen

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China)

[Abstract] Chinese translation of English verbs is one of the difficulties for machine translation. A new Chinese translation algorithm of English verbs based on semantic pattern base, fixed pattern base and variable base is presented. The extraction method of semantic pattern is introduced; the semantic pattern base, fixed pattern base and variable base of commonly used verbs are constructed. Three verb semantic patterns are concluded. The verb semantic pattern base consists of V semantic pattern base, contiguous VO semantic pattern base and non-contiguous VO semantic pattern base. Experimental results show that the method can resolve the Chinese translation of English verbs effectively.

[Key words] machine translation; semantic type; verb; semantic pattern; variable

DOI: 10.3969/j.issn.1000-3428.2011.19.049

1 概述

动词是语言中最活跃的词类, 变化多、形式复杂、多义性强。国内外学者已经从多种不同的角度对动词进行研究。语言学家从句式、语义、分类、认知等多个方面对英语动词进行研究。在自然语言处理领域, 多义动词消歧也是一个研究重点。尽管国内外学者在词义消歧方面已经做了一些工作, 但从机器翻译的角度对英语动词的语义消歧和汉译研究却很少。

文献[1]利用维基百科中的语义知识, 融合上下文语境、背景知识和语义信息等进行了词义消歧。文献[2]应用语法关系和WordNet中的名词上位层级构建了语义选择限制, 并利用概率分布和贝叶斯法则对动词进行了语义消歧; 文献[3]利用支持向量机和WordNet中名词上位的语义特征对多义动词进行了语义消歧; 文献[4]总结和改进了句法特征和语义特征, 并将其应用于英语动词的词义消歧。目前, 对英语动词的研究一般都是利用现有词典和知识库对动词前后搭配的名词进行语义标注和语义分类, 再基于规则或统计的方法达到动词消歧的目的。这些研究主要针对单个动词, 很少对动词相关词组进行研究。文献[5]以各语言的相通意义为研究对象, 提出了语义语言和语义单元的概念。

本文在语义单元理论的基础上提出了动词语义模式的概念, 对动词搭配的变量加入了语义限制。给出一种基于语义模式库、固定句式库和变量库的动词汉译算法, 对动词和动词相关词组进行语义分析和汉译。

2 语义模式、语义模式库、固定句式库和变量库

2.1 动词语义模式

定义 1 语义类型是具有相同语义特征和相同语法特征

的词语、词组或句子的集合。用 $A\{B\}$ 表示, 其中, A 表示语法特征; B 表示语义特征, 例如: 语义类型 $N\{\text{国家}\}$ 是表示国家的名词的集合, 其中“ N ”指该语义类型的语法特征是名词, “国家”指该语义类型的语义特征是国家。

定义 2 变量是指语义模式中影响动词语义的可以被替换的词语部分。变量用 X 表示。在具体的语义模式中变量可由一个或多个语义类型替换。

定义 3 实量是指语义模式中不可替换的自然语言中的字符, 用 R 表示。在动词语义模式中把动词和与动词相搭配的介词等词汇作为实量。用 V 代表动词, O 代表其他实量。 R 由 V 或 V 与 O 的组合构成。

定义 4 语义模式由一个或多个变量 X 和实量 R 按照一定顺序排列构成。英语语义模式和与其对应的汉语语义模式总是成对存在。英语语义模式和汉语语义模式中变量和实量的位置可以不同。例如 $N\{\text{人}\} \text{ sell } N\{\text{人}\} N\{\text{物}\}$, $N\{\text{人}\}$ 把 $N\{\text{物}\}$ 卖给 $N\{\text{人}\}$ 是动词“sell”相关的一对英语语义模式和汉语语义模式。

2.2 动词语义模式库

定义 5 动词语义模式库是由动词英语语义模式和与之对应的汉语语义模式组成的集合。动词语义模式库包含单一实量语义模式库即 V 语义模式库, 和复合实量语义模式库即 VO 语义模式库。 VO 语义模式库包含 VO 相邻语义模式库和

基金项目: 中央高校基本科研业务费专项基金资助项目(DUT10RW202)

作者简介: 张 瑶(1983—), 女, 博士研究生, 主研方向: 动词汉译算法, 自然语言处理, 机器翻译; 黄德根, 教授、博士生导师

收稿日期: 2011-03-30 **E-mail:** zy_dwj4@sina.com

VO 非相邻语义模式库。

单个动词可以表示一个语义, 例如: 英文“run”对应汉语语义“跑”。动词和与之相搭配的词汇构成动词相关词组, 共同表示一个语义, 例如: 英文“get up”对应汉语语义“起床”。这些动词或动词相关词组是唯一表示动词语义的最小单位, 不可再分, 为实量 R 。按照动词和动词相关词组的组合方式, 构造动词语义模式库的 3 个子集。在每个语义模式库中, 动词英语语义模式与汉语语义模式都成对出现。

V 语义模式库是单个动词作为 R 部的语义模式构成的集合。例如: $N\{人\} own N\{物\}$, $N\{人\} 拥有 N\{物\}$ 是英语动词“own”在 V 语义模式库中的一对英汉语义模式。

VO 语义模式库是由动词和其他实量按照不同排列组成的语义模式的集合, 存放动词相关词组的语义模式。

其中, VO 相邻语义模式库是由动词和与之相邻的其他实量共同作为 R 部的语义模式构成的集合。例如: $N\{人\} meet up with N\{人\}$, $N\{人\} 偶遇 N\{人\}$ 是英语动词“meet”在 VO 相邻语义模式库中的一对英汉语义模式。

VO 非相邻语义模式库是由动词和与之不相邻的其他实量共同作为 R 部的语义模式构成的集合。例如: $N\{人\} make N\{物\} out of N\{材料\}$, $N\{人\} 用 N\{材料\} 制造 N\{物\}$ 是英语动词“make”在 VO 非相邻语义模式库中的一对英汉语义模式。

2.3 变量库和固定句式库

变量库是由各种语义类型及其在英汉语中对应的具体表示构成的集合。例如语义类型 $N\{方向\}$ 对应的英汉语具体表示为: east 东方, west 西方, north 北方, south 南方, northeast 东北方, northwest 西北方, southeast 东南方, southwest 西南方。

固定句式库是由含有动词固定句式的英语句子和对应的汉语译文构成的集合。每个固定句式都是一个最小的语义单位, 不可再分。例如“存放 Not that I know of, 据我所知没有”; “How do you do, 你好等句子”。

3 常用动词的汉译算法

3.1 动词语义模式的提取方法与动词语义模式库的构建

在大量的英汉双语语料中提取语义模式。对含有常用动词的句子进行对照提取, 将其相同部分提取为实量, 句法位置相同、语义类型相同的部分提取为同一语义模式的变量。变量可以是单词、短语或句子。分别将英语语义模式的实量对应的汉语译文提取成汉语语义模式的实量, 英语语义模式的变量对应的汉语译文提取成汉语语义模式的变量。实量或变量在位置上会有所差别。

对英语动词中 211 个常用动词进行语义模式提取, 根据这些语义模式, 构建 211 个常用动词的语义模式库、固定句式库和变量库。一共提取英汉句对 8 632 对, 英汉对照语义模式 6 986 对。语义模式库主要存放英语语义模式、汉语语义模式、每个语义模式对应的变量数量和语义类型等。将常用动词相关的固定句式和对应的汉语译文直接提取构成固定句式库。变量库中存放了 237 种与常用动词相关的变量的语义类型, 以及每种变量具体的英汉语表示。

3.2 英语常用动词的汉译算法

定义 6 目标词(I)即需要进行汉译的动词

步骤 1 初始化由 $S_1 S_2 \dots S_{n-1} S_n$ 等单词由左至右排列构成的英语句子。

步骤 2 将含 I 的句子与固定句式库进行匹配; 若匹配成

功, 输出汉语译文; 若失败, 转步骤 3。

步骤 3 将 I 与 VO 相邻语义模式库的 V 部匹配; 若匹配成功, 转步骤 4; 若失败, 转步骤 6。

步骤 4 取 I 右词 M, M 与 VO 相邻语义模式库中的 O 部匹配; 若匹配成功, 转步骤 5; 若失败, 转步骤 6。

步骤 5 用语义模式匹配算法在 VO 相邻语义模式库中匹配语义模式; 若匹配成功, 转步骤 11; 若失败, 转步骤 6。

步骤 6 将 I 与 VO 非相邻语义模式库的 V 部匹配; 若匹配成功, 转步骤 7; 若失败, 转步骤 9。

步骤 7 取 I 右侧除 M 以外的词 $P_1 P_2 \dots P_n$, 分别与 VO 非相邻语义模式库中的 O 部匹配; 若匹配成功, 转步骤 8; 若失败, 转步骤 9。

步骤 8 用语义模式匹配算法在 VO 非相邻语义模式库中匹配语义模式; 若匹配成功, 转步骤 11; 若失败, 转步骤 9。

步骤 9 将 I 与 V 语义模式库的 V 部匹配; 若匹配成功, 转步骤 10; 若失败, 输出失败, 算法结束。

步骤 10 用语义模式匹配算法在 V 语义模式库中匹配语义模式; 若匹配成功, 转步骤 11; 若失败, 输出失败, 算法结束。

步骤 11 输出英语语义模式和对应的汉语语义模式并输出汉语语义模式的 R 部即是英语动词 I 的译文。

语义模式匹配算法如下:

```
begin
<将含有 R 的语义模式按照变量数量递减的顺序排列成队列 Q>;
While(Q 不为空)
begin
<取 Q 队首的语义模式>;
<在变量库中匹配 X 部各变量>;
if (匹配成功) then
<输出英语语义模式>;
<在语义模式库中搜索对应的汉语语义模式并输出>;
break;
endif
end
end
```

4 实验及分析

4.1 实验设置

在厦门大学双语语料中选取 4 万句对进行开放测试, 人工统计了从中随机抽取的 9 个单词共 1 157 个英语句子的测试结果。表 1 显示了测试语料的基本信息。

表 1 测试语料

句对总数	英语常用动词	含有该词的句数
40 000	eat	99
40 000	grow	87
40 000	hear	200
40 000	kill	56
40 000	meet	161
40 000	pay	119
40 000	read	174
40 000	receive	95
40 000	wear	166
共计	9 个词	1 157

表 2 显示了开放测试的结果。其中, ①表示翻译结果正确且与原汉语译文相同; ②表示翻译结果与原汉语译文不同, 但是也是正确翻译, 不改变句义; ③表示未匹配成功, 翻译失败; ④表示匹配成功, 但翻译结果错误。其中, 正确率 = $[(①+②)/(①+②+③+④)] \times 100\%$ 。

表 2 开放测试结果

英语常用动词	①	②	③	④	正确率/(%)
eat	83	7	7	2	90.9
grow	42	36	8	1	89.7
hear	156	22	17	5	89.0
kill	35	16	5	0	91.1
meet	87	53	20	1	87.0
pay	62	48	8	1	92.4
read	126	34	14	0	92.0
receive	70	19	6	0	93.7
wear	116	27	20	3	86.1

4.2 实验结果分析

根据实验结果可知,本文提出的基于语义模式库、固定句式库和变量库的常用动词汉译方法能够达到90%左右的正确率。导致翻译失败的原因主要有以下2种:

(1)语义模式库、固定句式库和变量库中的语义模式、固定句式和变量及其英汉语具体表示收录不全导致翻译失败。

(2)省略等特殊句式导致翻译失败。例如:

“and have you heard from him often?” “常常收到他的信吗?”

“We have heard only once.” “只收到他一封信。”

其中,第1个英汉句对中“hear from”可以正确匹配语义模式并得到正确的汉语译文;第2个英汉句对由于省略了介词“from”导致语义模式无法正确匹配翻译失败。

因此,只要继续扩充和完善语义模式库、固定句式库和变量库,利用该方法对英语动词汉译的正确率还会有一个很大的提高空间。对于省略等特殊句式造成的翻译错误情况也可以通过增加上下文语境函数来改善。

5 结束语

本文以常用动词为研究对象,构建了动词的语义模式库、

固定句式库和变量库。归纳了动词语义模式的3种结构形式,并以此构建了3种语义模式库。提出了一种基于语义模式库、固定句式库和变量库的动词汉译算法,对句子中的英语常用动词和动词相关词组进行了语义分析和汉译。实验结果表明,本文提出的动词汉译方法能有效地对英语动词进行正确的翻译。为进一步提高准确率,需要在将来的工作中对动词语义模式库、固定句式库与变量库进行不断完善,并通过引入语境分析等方法加强对省略等特殊句式的处理。

参考文献

- [1] 史天艺,李明禄. 基于维基百科的自动词义消歧方法[J]. 计算机工程, 2009, 35(18): 62-64, 66.
- [2] McCarthy D, Carroll J. Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences[J]. Computational Linguistics, 2003, 29(4): 639-654.
- [3] Buscaldi D, Rosso P, Pla F, et al. Verb Sense Disambiguation Using Support Vector Machines: Impact of WordNet-extracted Features[C]//Proc. of the 7th International Conference on Computational Linguistics and Intelligent Text Processing. Mexico City, Mexico: [s. n.], 2006: 192-195.
- [4] Chen Jinying, Palmer M. Improving English Verb Sense Disambiguation Performance with Linguistically Motivated Features and Clear Sense Distinction Boundaries[J]. Language Resources & Evaluation, 2009, 43(2): 181-208.
- [5] 高小宇,高庆狮,胡 玥,等. 基于语义单元表示树剪枝的高速多语言机器翻译[J]. 软件学报, 2005, 16(11): 1909-1919.

编辑 索书志

(上接第149页)

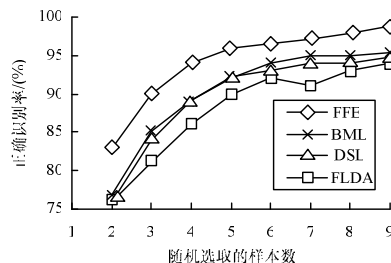


图 2 不同训练样本下人脸识别的正确率比较

从图2可以看出,FFE算法在训练样本数目变化过程中,始终保持着比较高的识别率。但其相对于其他子空间的算法计算复杂度要高一些,因为本文算法要估计类内散度矩阵零空间中的判别值,但对于同样的识别率,本文算法会更快,因为从实验1和实验2中可以看出,本文算法使用较少的特征维数就可以达到相同的识别率。

6 结束语

散度矩阵的主空间和零空间也包含了一些鉴别信息,在特征提取过程不应该被丢弃。从这2个空间里单独提取鉴别特征往往是次优的算法,所以应该在整个空间里进行特征的提取。本文算法是把类内离散度矩阵的特征向量所张成的空间分解为二个子空间,然后利用本征谱模型对这二个子空间的特征值进行正则化以避免或减轻不稳定、过拟合和推广性较差的问题,最后在整个正则化后的空间里进行特征提取和降维,这样就得到了一组有鉴别性、可靠的和低维的人脸表示子空间。经实验表明,本文算法能用较少的特征维数达到与其他算法相同的识别率,并且在特征提取上是高效的、稳

定的、更小的过拟合,具有很好的推广性。

参考文献

- [1] Kirby M, Sirovich L. Application of Karhunen-loeve Procedure for the Characterization of Human Face[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1990, 12(1): 103-108.
- [2] Moghaddam B. Principal Manifolds and Probabilistic Subspace for Visual Recognition[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002, 24(6): 780-788.
- [3] 刘典婷,华顺刚. 基于 Fisher 投影的多光谱人脸融合识别[J]. 计算机工程, 2010, 36(8): 180-182.
- [4] 程正东,各毓晋,樊 祥. FISHER 线性鉴别函数的一种推广形式[J]. 模式识别与人工智能, 2009, 22(2): 176-181.
- [5] Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces and Fisherfaces: Recognition Using Class Specific Linear Projection[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711-720.
- [6] Lu Juwei, Plataniotis K N, Venetsanopoulos A N. Face Recognition Using LDA-based Algorithms[J]. IEEE Trans. on Neural Networks, 2003, 14(1): 195-200.
- [7] Wang Xiaogang, Tang Xiao'ou. Dual-space Linear Discriminate Analysis for Face Recognition[C]//Proc. of IEEE International Workshop on Computer Vision and Pattern Recognition. Washington D. C., USA: [s. n.], 2004: 564-569.
- [8] Liu Chengjun, Wechsler H. Robust Coding Schemes for Indexing and Retrieval from Larger Face Database[J]. IEEE Trans. on Image Processing, 2000, 9(1): 132-137.

编辑 任吉慧

