

语义组合与机器翻译

张霄军 著



科学出版社
www.sciencep.com

语义组合与机器翻译

张霄军 著

科学出版社

北 京

图书在版编目(CIP)数据

语义组合与机器翻译 / 张霄军著. —北京: 科学出版社,
2010

ISBN 978-7-03-029201-8

I. ①语… II. ①张… III. ①机器翻译-研究 IV. ①H085

中国版本图书馆 CIP 数据核字 (2010) 第 198690 号

责任编辑: 阎 莉 / 责任校对: 何晨琛

责任印制: 赵德静 / 封面设计: 无极书装

联系电话: 010-6403 0529 电子邮箱: yanli@mail.sciencep.com

科学出版社出版

北京东黄城根北街16号

邮政编码: 100717

<http://www.sciencep.com>

中国科学院印刷厂印刷

科学出版社编务公司排版制作

科学出版社发行 各地新华书店经销

*

2010年10月第 一 版 开本: A5(890×1240)

2010年10月第一次印刷 印张: 9 1/8

印数: 1—2 000 字数: 330 000

定价: 42.00 元

(如有印装质量问题, 我社负责调换〈科印〉)

本书得到以下出版基金和项目资助：

陕西师范大学出版基金

国家社会科学基金项目

(项目编号：07BYY050、10CYY006)

国家自然科学基金项目

(项目编号：60903123、60773173)

陕西师范大学校级重点项目

(项目编号：995377)

序

不管采用语言学规则，还是利用统计学知识，语义始终是自然语言处理绕不过去的瓶颈，词义的模糊性、歧义性和多样性是影响自然语言理解的主要原因。随着计算语言学的发展，越来越多的学者加入到了语义研究的行列，语义分析、语义标注、词义消歧、语义知识库建设等研究工作开展得越来越广泛。

张霄军博士也加入到了这一行列，他的研究工作主要体现在词义消歧方面。不过，他所做的词义消歧研究与他人有所区别：他关注的是跨语言的词义消歧，即寻找跨语言的语义对应关系。这样，他的研究对象就既包括源语词语的语义消歧，也包括目标语译词的语义消歧，但这样的工作绝不是源语研究和目标语研究的简单“叠加”——虽然只是研究对象从单语变成了双语，但涉及的研究内容却比单语语义研究多了不止一倍。毕竟，能够很好地应用于单语词义消歧的方法和模型，未必就适用于跨语言的词义消歧。

张博士想到的解决方案是利用语义的组合模式。这是一种很朴素但很实用的思想：如果源语和目标语的单词之间的语义对应关系难以确定的话，那我们为何不扩大对象——将单词扩展到几个连续的词语呢？单词的特定义项难以确定，那么几个组合在一起的词语序列的义项则是相对固定的。这就是语义组合。组合型语义关系是指词义之间的相互依存关系，在这种依存关系中，往往有一个词义居于中心地位，对其他词义有组配要求。通常说的动词配价、名词配价就反映了这种关系。这种关系既可以是动态地表现在句子之中，也可以是静态地存在于那个居于中心地位的词义之中。因此，语义组合模式中既包括词汇的语义特征组合，也包括语法语义特征。在语义组合研究过程中，细心的张博士还发现了“主观性语义特征”，并将其也纳入语义组合模式之中。

然而，实践这一思想的过程却是艰苦而枯燥的。归纳所有词汇的语义组合模式显然是不现实的，我要求他把常用的动词和名词尽量考察完全。他做了一段时间后跟我提出能否只做动词。考虑到工作量过于繁重，

我同意了，但我告诫他：“要想穷尽地归纳出一个核心动词所有的语义组合模式，至少得考察 100 条例句。”这是我对他的要求，也是忠告。他做到了，而且是 double！对于每个动词，他考察的不是 100 条例句，而是 100 对英汉平行的例句！当然，这需要不少的时间和精力。尽管他非常刻苦和勤奋，但直到毕业答辩，任务还是没有完成——预期完成的 800 多个英语多义动词，他只归纳了 400 多个。

但是，看到他交来的博士论文时，我理解了。他的研究目的不仅仅是要找到源语和目标语动词相对应的语义组合模式，而且要将这种研究方法应用到机器翻译实践中去。因此，他将机器翻译中的“译词选择”作为论文的另一重要内容。换句话说，他在完成了源语动词的语义消歧和目标语译词的语义消歧之后，还完成了目标语译词的生成。最后，为了验证这种方法的有效性，他又将“译文质量评测”纳入了研究范畴，设计了验证实验，实验结果表明利用语义组合模式进行译词的语义选择是有效的。

至此，从理论分析到语义组合模式提取，再到语义组合模式应用，最后到方法有效性检验，整篇论文的结构就很完整了，结论也很有说服力，论文盲审和毕业答辩都顺利过关。然而，总归有些遗憾，布置的任务毕竟没有完成啊！毕业以后剩下的 300 多个动词他会坚持做下去吗？

2009 年岁末，张博士寄来书稿《语义组合与机器翻译》，嘱我写序。我看到在这本由其博士论文“基于语义的动词译词选择与译文评价”修改而成的学术专著中，800 多个英语多义动词的 5 000 多个语义组合模式已经全部总结了，“双语动词语义词典（Bilingual Verb Semantic Dictionary, BiVSD）”建设完成，感到由衷的欣慰。张博士是勤奋的。我希望他继续努力，把他的研究成果应用于英汉机器翻译实践，在语义计算方面做出更多更好的成绩。

陈小荷

2009 年岁末于南京白云园

前 言

语义障碍是自然语言处理的一块拦路石,也是影响提高机器翻译译文质量的因素之一。机器翻译中的“义障”主要在于自然语言词汇语义的不确定性,这种不确定性往往使得翻译系统难以做出恰当的译词选择。从评测的角度来看,在译文质量优劣的评价因素中,语义因素无疑也是权重最高的评判标准。在语言本体研究中,语义研究一直是重中之重,前人运用不同的理论和方法,对自然语言的语义及语义关系进行了不同层面的描写,编纂了各种释义词典(对词汇语义的描写)、义类词典(对词汇聚合语义关系的描写)和搭配词典(对词汇组合语义关系的描写)等;在自然语言处理研究中,语义的处理方兴未艾,计算语言学家也依据不同的语义体系,构建了大型语义知识库和机器可读的语义词典等。然而,上述研究仍不能满足机器翻译对译词选择的语义分析的要求,更无法为量化的译文质量评价提供切实可行的语义依据。为此,我们有必要为机器翻译这个特定的自然语言处理领域做专门的语义研究。

本书在前人语义研究的基础上,总结了英-汉机器翻译系统中英语动词的语义特征组合模式,在此基础上提出了一种基于“语义特征组合模式”的动词译词选择方案。首先,通过知识库融合的方法从英语 FrameNet 和英汉双语 HowNet 中生成了常见的英语多义动词 839 个,在大量真实语料和词典例句语料的基础上对其语义特征组合模式进行了归纳总结,共提取英语多义动词的 5 092 种语义组合模式。在每种模式中,我们都给出了该英语动词相对应的汉语可选译词。在此基础上构建了供机器翻译译词选择所“专用”的“双语动词语义词典(BiVSD)”,利用 BiVSD 中这 5 000 余种语义组合模式及其所涉及英语动词的汉语可选译词,辅以动词语法语义特征,机器翻译系统就可以准确判断出该动词在一个具体语境中的恰当译词,从而选择出恰当的译词。确定了基于语义组合模式的译词选择方案之后,本书围绕机器翻译过程对原语分析、译词选择和译文生成各个阶段进行了分析,界定了命题、小句和句子的定义,提出了主观性语义特征的概念并分析其在译词选择中的作用,并具体分析

了两种特殊结构中的动词译词选择方法。最后,以个案研究的方式,对含有英语多义动词 beat 的 176 句英文原句进行动词的译词选择,实验结果译词选择正确率为 89.2%,模式匹配正确率为 98.8%,模式匹配覆盖度为 94.9%。

在译文评价方面,本书首先对上述译词选择方法进行比对评价:选取 800 多个动词中的 200 个典型多义动词,从语料库中抽取含有这 200 个动词的例句共计 1 000 个例句(每个动词抽取含有该动词的例句 5 句),建立语义选择测试集。分别通过语义特征组合模式、机器翻译系统和人工选择三种方法选择每句原文中动词的恰当义项,将这三种形式的译词选择结果分别与参考译文中的译词进行比较,得到每种形式与参考译文译词的吻合度。比对实验结果表明语义特征组合模式的译词选择结果吻合度为 92.7%,远远高于机器翻译译词选择结果(52.3%),也高于人工选择的结果(88.06%)。此外,本书还提出一种基于模糊集的多语言点译文质量综合评价方法,该方法综合了待评价译文的语义、语法、语用和语篇诸因素,利用模糊数学的方法对同一译文的多个评价结果进行统计判断,并做个案研究验证了该方法在译文评价上的有效性。

Abstract

Word sense is one of the obstacles of natural language processing (NLP), and the factor which weakens the translation quality of machine translation (MT), also. The sense obstacle of MT is the uncertainty of the word sense of natural language, which makes the selection of the target words difficult to the translation system. On the MT evaluation, semantic element is the most important one in all the judgement elements of translation quality evaluation. Linguists focus on meaning studies always. They describe the word sense and word sense-relations in different views with different methods and compile all kinds of paraphrasable dictionaries (to describe word sense), thesaurus dictionaries (to describe word paradigmatic sense relation), and collocation dictionaries (to describe word associative sense relation); computational linguists concentrates on the word sense studies recently and they construct large-sized sense-knowledge database and machine readable sense dictionaries. However, all the above sense studies can not meet the requirement of MT system to the target word selection, and provide a strong semantic support to evaluate translation quality, either. Therefore, it is necessary to do the professional studies on word sense for MT.

Benefited from the present word sense studies, this book points out a method to select the target word based on "semantic feature combination models (SFCM)" after extracting the models of English verbs manually for E-C MT system. Firstly, we select 839 English polysemic verbs from FrameNet and HowNet, extract their SFCMs from lots of real sentences in corpus and sentence examples in dictionaries. Till now, 839 verbs were all analyzed and 5092 models were extracted. In every model, we give out the available target words of the English verb. Then we collect these models to form the "Bilingual

Verb Sense Dictionary (BiVSD)", which is the special use to the target word selection of MT system. MT system selects the target word via the more than 2000 model in BiVSD. Aided with the grammatical semantic feature, system can generate the proper target word of the original verb. It also analyzes every stage of the procession of MT which includes original text analysis, target word selection and target text generation three stages after setting up the selection method, defines the concept of proposition, clause and sentence, creates the new concept of "subjective semantic feature (SSF)" and describes its roles in MT target word selection. The exceptions are also mentioned in this dissertation. At last, it selects the verb *beat's* target word selection in 176 English sentences. The precision of the selection is 89.2%, the precision of model match is 98.9% and the recall of model match is 94.9%.

On the evaluation of translation quality, it contrasts different selection approaches to test the efficiency of the above method: selection 200 polysemic verbs contained in 1000 sentences, which is the test set. To select the target word of these verbs in these sentences via SFCM method, MT selecting approach and testee selecting approach seperatively, and to contrast their selecting results with the keys, we may get the match degree to the keys: SFCM method is 92.7%, which is much better than MT approach (52.3%), and better than testee approach (88.06%) too. Otherwise, this book points out an integrative evaluation method to translation quality, which is based on fuzzy set. The method integrates semantic, syntactic, pragmatic and context elements to judge the evaluation results of the multi target texts of the same one original text. And it illustrates the method to show its efficiency in translation quality evaluation.

目 录

序.....	i
前言.....	iii
Abstract	v
第 1 章 绪论	1
1.1 翻译、语义与评价.....	1
1.2 动词概述.....	8
1.3 本书的研究框架.....	13
1.3.1 研究目的、对象、目标与假设	13
1.3.2 本书的研究方法与步骤.....	15
1.3.3 本书的结构.....	16
第 2 章 动词语义描写	18
2.1 动词语义描写研究.....	18
2.1.1 动词语义描写理论研究.....	18
2.1.2 动词语义描写方法研究.....	24
2.2 典型的动词语义特征描写体系.....	28
2.2.1 典型的英语动词语义描写体系.....	28
2.2.2 典型的汉语动词语义描写体系.....	29
2.3 本书动词语义特征描写体系.....	35
2.3.1 动词语义特征提取.....	35
2.3.2 动词语义关系描写.....	36
2.3.3 动词语义描述体系.....	36
2.4 动词语义描述体系示例.....	37
第 3 章 动词语义组合模式	41
3.1 动词语义特征.....	41
3.1.1 义项描写表.....	41
3.1.2 组合描写表.....	47

3.2 动词语义特征组合模式	53
3.3 主观性语义特征	57
3.3.1 主观性、主观量和主观化	57
3.3.2 客观性语义特征与主观性语义特征	59
3.3.3 主观性语义特征与机器翻译译词选择	60
3.4 动词的语法语义特征	63
3.4.1 改进的句子语义模型	63
3.4.2 英语动词句语法意义系统	64
3.4.3 英汉动词语法语义特征对照表	68
3.5 双语动词语义描写词典	70
3.5.1 上节内容回顾	70
3.5.2 知识库融合	72
3.5.3 BiVSD 建设进展	76
3.6 本章小结	77
第4章 机器翻译与词义消歧	79
4.1 机器翻译概述	79
4.1.1 机器翻译简史	79
4.1.2 机器翻译在中国的进展	82
4.1.3 机器翻译系统的主要方法	83
4.1.4 机器翻译系统的难点及研究思路	87
4.2 词义消歧	89
4.2.1 词义消歧及其意义	90
4.2.2 词义消歧方法	90
4.3 机器翻译译词选择	92
第5章 动词的译词选择	93
5.1 动词译词选择流程	93
5.1.1 译词选择策略	93
5.1.2 动词译词选择的流程	95
5.2 原语分析	98
5.2.1 命题 P	98
5.2.2 小句 C	99

5.2.3 句子 S	110
5.3 译词选择	111
5.4 译词生成	123
5.4.1 词类转译的主要类型	123
5.4.2 可选语法译词生成	124
5.5 特殊结构中的动词译词选择	126
5.5.1 “ $v_1 c v_2$ ”结构中的动词译词选择	126
5.5.2 独动词句中的动词译词选择	127
5.5.3 复句处理	127
5.6 动词译词自动选择的个案实现	128
5.6.1 译词选择步骤和算法	128
5.6.2 m_i 的确定	129
5.6.3 M_i 和 M_n 的确定	129
5.6.4 特殊结构翻译模式	131
5.6.5 实验语料来源	131
5.6.6 实验结果及分析	132
5.7 本章小结	133
第 6 章 动词句译词选择评价与译文评价	135
6.1 译词选择方法与译文质量评价	135
6.1.1 译词选择方法综述	135
6.1.2 译文质量评价综述	139
6.2 译词选择结果比较与分析	145
6.3 多语言点译文质量综合评价方法	150
6.4 个案研究及进一步思考	153
第 7 章 总结	161
7.1 已经完成的工作	161
7.2 对前人研究工作的改进	162
7.3 本研究工作的不足	164
7.4 进一步研究计划	166
参考书目	168

一、外文参考书目	168
二、中译参考书目	170
三、中文参考书目	171
四、工具书	176
五、网络资源	177
附录 1 英语多义动词	178
附录 2 英语动词语义组合模式(样例)	183
附录 3 含有动词 beat 的英语源语句	188
附录 4 动词译词选择测试源语句	195
附录 5 译词选择评价结果	232
后记	273

绪 论

1.1 翻译、语义与评价

本书旨在运用确定性的语义特征组合模式来描述不确定的动词语义，探讨一种基于语义转换规则的机器翻译的动词译词选择方法，进而实现机器翻译译文质量的用户评测。选题涉及三大相对独立而又绝对联系的内容——翻译、语义和评价。

1. 翻译

巴别塔的传说是人们对不同语言各自起源的无限憧憬，而“重建巴别塔”则一般被视为翻译的目标。在这个构建通天之塔的过程中，要消除各语言之间的隔阂，就必须找到其“共通的地方”。换言之，语言的共通性是翻译可译性的基础，而语言的差异性则是翻译不可译性的限度。

美国语言学家罗曼·雅各布森(Roman Jakobson)在谈及翻译时指出“语言之间之所以能够互相翻译，是因为不同语言之间总有共通的地方”。雅各布森认为，各种语言都具有同等的表达能力，“任何认知经验及其分类在任何存在的语言中都能表达”(Jakobson, R., 1959)。因此，他反对英国哲学家罗素对词语意义的看法，即要理解词的意义必须要有认知经验。雅各布森从符号学出发，认为人们对词义的理解，首先取决于人们对赋予词的意义理解。也就是说，词是语言中的一个语言符号，其意义也是人们赋予它的，这就是语言的符号学观点。进而，雅各布森采用皮尔士的符号和意义理论，指出任何语言符号的意义在于进一步把它翻译成其他可代替的符号，尤其是翻译成更为发达的符号。这一观点，以及他对翻译的分类和翻译对等的解释，构成了翻译的符号学理论基础。

雅各布森的这个理论成为西方翻译理论的经典，奈达、乔治·斯坦纳、德里达、A. J. 格雷马斯等人都曾专门对此进行过讨论，费罗道夫、奈达等直接继承和发展了他的语言共通性理论，并将其转化成为翻译理论中的“等值”和“对等”理论。在乔姆斯基的深层结构理论中，作为不同语音输入的表层语句，其“可译性”条件取决于它们能否被“翻译”为同一个作为“语义输入”的深层句子。我们可以推而广之，不同语种(如英语和汉语)之间是否具有“可译性”，完全取决于它们在语言上是否具有相同的深层结构，即相同的“语义输入”。

在我国传统译论中，“信”是其最基本和最核心的概念。“以‘信’为本的传统翻译思想有很浓的趋同意识。”(郑海凌，2006)按照严复本人对“信、达、雅”的解释来看，所谓“信”是指“意义不悖(背)本文”，即忠实。但同时，严复也指出“求信已大难矣，……，为达，即所以为信也。”达，通顺。就信与达的主从关系而言，以信为本，以达求信，意义对等始终是翻译的第一准则。后来，严复的“信、达、雅”翻译标准的“修正版”，如“信、达、切”(刘重德)、“透、化、风”(黄药眠)、“‘学思得’-‘信达雅’”(郁达夫)、“信与顺”(鲁迅)、“三似论”(陈西滢)、“化境说”(傅雷)等无一不是围绕着“信”——目标语所表达的语义要尽可能符合源语表达的语义。

机器翻译是人工智能学科的一个分支，机器翻译活动是一种类人行为。因此，人工翻译中源语与目标语之间的共通性也同样适用于机器翻译。本书讨论的机器翻译系统主要是基于规则的¹。基于规则的机器翻译系统有三种类型：直接翻译类型、转换翻译类型和中间语翻译类型。直接翻译类型的系统一般适用于语言结构相似的不同语言互译，根据源语与目标语翻译单元的对应语义关系直接匹配翻译；转换翻译类型的系统需要通过双语对比构建复杂的结构映射与转换规则，转换时需要一部双语对应词典；通过中间语言翻译类型的系统把源语文本用人工设计的无歧义的中间语言表示出来，然后把中间语言所表达的意义用目标语言的词汇和句法结构表示出来。在基于规则的这三种类型的翻译系统中，都离不开一个共同的元素——“对应语义关系”，只不过在通过中间语翻译类型的系统中多了一个“语义编码”与“语义解码”的过程而已。随着

1 机器翻译的类型及方法详见 4.1.3。

计算机运行速度的提高、存储容量的大规模增加和大规模电子文本获取的可能,基于统计和基于实例的翻译系统逐渐成为目前机器翻译的主要类型。基于统计的翻译系统可以分成三种类型。第一类是基于平行概率语法的统计机器翻译方法。其基本思想是用一个双语平行的概率语法模型,同时生成两种语言的句子,在对源语言句子进行理解的同时就可以得到对应的目标语言句子。第二类是基于信源信道模型的统计机器翻译方法。其基本思想是把机器翻译看成是一个信息传输的过程,用一种信源信道模型对机器翻译进行解释。假设一段目标语言文本 T ,经过某一噪声信道(编码)后变成源语言文本 S ,那么翻译的目标就是要将 S 还原成 T ,这也就是一个解码的过程。第三类是基于最大熵的统计机器翻译方法。最大熵方法是一种通用的统计建模的方法,用来模拟随机事件的分布。把该方法用在机器翻译系统中就需要选择一组源语语言特征,使得其与语料库中目标语样例中的分布完全一致,同时又保证这个模型尽可能的“均匀”(也就是使模型的熵值达到最大),以确保除了这一组特征之外,这个模型没有其他的任何偏好。事实上这是一种直接翻译模型,所依靠的是一个大规模双语平行语料库。在源语与目标语的“共通性”上,基于平行概率语法的统计机器翻译方法与转换翻译的规则机器翻译方法类似,基于信源信道模型的统计机器翻译方法与通过中间语翻译的规则机器翻译方法相同,基于最大熵的统计机器翻译方法与直接翻译类型的规则机器翻译方法无异。而基于实例的机器翻译系统则更像人工翻译流程:首先把输入的句子正确地分解为一些短语碎片,接着把这些短语碎片翻译成其他语言的短语碎片,最后再把这些短语碎片构成完整的句子。每个短语碎片的翻译是通过类比的原则来实现的,因此基于实例的机器翻译系统又被称为“基于类比的”机器翻译系统。在具体翻译之前,计算机应该预先存储一定数量的翻译实例,由系统来识别和检索与待翻译句相匹配的实例句,并根据它们之间的相似度大小人工修改译文。完成翻译后得到的翻译句对又可存储到翻译实例中去作为后来者的参考实例。事实上,基于实例的机器翻译方法还是通过“中间语”来进行翻译的,只不过这儿的“中间语”不是人工语言,而是翻译实例库中的翻译实例,省略了语义编码和解码的过程,却增加了人工修正的步骤。

综上所述,无论国外翻译理论,还是我国传统翻译理论,都在强调翻译中不同语言的“共通性”;无论是人工翻译,还是机器翻译,其翻译