第 33 卷第 5 期　　兰州文理学院学报(社会科学版)　　Vol. 33, No. 5
2017 年 9 月　　Journal of Lanzhou University of Arts and Science (Social Sciences)　　Sept. 2017

机器翻译中的歧异性研究现状综述

余 碧 燕
(西安交通大学 外国语学院,陕西 西安　710049)

【摘　　要】语际差异是无法消除的客观事实,因而歧异性成为翻译研究必须面对的问题,尤其是对于高度注重形式化语言的机器翻译来说。本文从歧异性的定义、分类以及国内外研究成果三个方面对这一概念进行梳理介绍,以便让国内翻译研究人员对翻译歧异性有更深入的了解。对于英汉翻译歧异性研究而言,形成跨语系的完整的理论模式存在一定困难,期待各领域重视理论研究、具体问题具体分析、实现跨学科合作发展,从而推动机器翻译实现长足进步。

【关　键　词】翻译歧异性;机器翻译;综述
【中图分类号】H085　　【文献标识码】A　　【文章编号】2095-7009(2017)05-0106-06

Current Studies on Translation Divergence in Machine Translation at Home and Abroad

YU Bi-yan
(*School of Foreign Languages, Xi'an Jiaotong University, Xi'an* 710049, *China*)

Abstract: The omnipresence of cross-linguistic distinctions makes divergence a core problem in translation, especially for Machine Translation (MT) which requires highly formalized languages. Previous studies have made great effort to discuss how to define translation divergences, how to specify them, and how to solve divergence problems in MT. In terms of the translation between English and Chinese, the formalization of an integrated theoretical framework about translation divergences is rather difficult since they are of different language families. It is highly expected that the theoretical research on translation divergences would be paid more attention to, that the specific issues between English and Chinese would be analyzed depending on their situations, and that the interdisciplinary cooperation would be realized, by which MT will achieve long-term progress.

Key words: translation divergence; machine translation; research summarize

一、引言

机器翻译(Machine Translation,简称 MT)是计算语言学(Computational Linguistics)的一个分支,涉及计算机、认知科学、语言学、信息论等学科,是人工智能的终极目标之一,兼有重要的科学研究价值和实用价值。自 20 世纪 90 年代起,MT 经历了一个快速发展的繁荣时期,翻译质量得到了突飞猛进的提高。尽管如此,在某些方面,MT 与人工翻译尚有距离。现实情况对 MT 的需求仍在不断上升,国内外研究人员对 MT 的探索和改进也从未停歇。

困扰 MT 的问题很多,其中一大问题是翻译歧异性(Translation Divergence)。在 20 世纪 80 年代,欧洲各国之间的政治经济一体化进程加快,与此同时,各国语言交流的迫切性日益增强。在此背景下,欧洲共同体(European Community)成立 EUROTRA 项目[1],旨在设计一个操作系统,用以当时欧共体内法语、德语、荷兰语、意大利语、英语、丹麦语、希腊语七种语言之间的翻译。项目设计时考虑的一大因素是语言习惯多样性(Diversity)的特点,因此一大挑战便是既要建立各种语言之间互通的接口表征(Interface Representation),又要兼顾每一种语言的地道性。七种语言

【收稿日期】2017-05-20

【作者简介】余碧燕(1992—),女,浙江余姚人,西安交通大学硕士研究生,主要从事语料库语言学、机器翻译、计量语言学研究。

两两互译，共42个语对的歧异性问题在当时受到关注。

MT发展至今，各国学者对翻译歧异性进行了大量的理论与实证探讨，并为解决MT中的歧异性问题作了不少努力与尝试，成果颇丰。但目前为止，国内翻译界关于MT的歧异性问题研究较少。有鉴于此，本文对各国学者关于翻译歧异性的研究进行梳理，以便让国内翻译研究人员对此有更深入的了解，在此基础上共同讨论该领域的发展及对翻译研究的启示。

二、翻译歧异性基本概念

歧异性是翻译中的一个普遍现象，在MT领域尤其受到关注。关于翻译歧异性的定义，学界没有严格统一的说法。Dorr认为，从某一自然语言到另一自然语言的翻译往往造成译文语句在形式、结构上与原文大相径庭，这便是翻译歧异性，亦即语际差异（cross-linguistic distinctions）[2]。Dorr & Voss将广义的歧异性定义为源语与目的语结构相异或表达不同信息的所有翻译句对，并将如此定义的歧异性比作MT中句对输入输出规范的“黑匣子（black box）”[3]。与之相比，狭义的歧异性是指自然语言表征（Language-to-Language）之间，或自然语言与中介语（Language-to-Interlingua）之间的映射（mapping），用来反映语言内隐假设或理论原则，Dorr & Voss将此类歧异性比作MT中句对输入输出规范的“透明玻璃箱（glass box）”[3]。而按照Mishra & Mishra的理解，源语中词汇、句法相似的句子转换成目标语后在词汇、句法上并不相似，该现象为语言歧异性[4]。Sinhal & Chandak则认为，歧异性不仅指语言本身的差异，同时也包括文本形式的差异[5]。

关于歧异性产生的原因，也是众说纷纭。Lin et al.认为歧异性产生于源语与目的语之间不同的词法（morphology）和词汇使用[6]。Sinhal & Chandak认为歧异性产生于源语与目的语内在的不兼容性（inherent incompatibility）[5]。虽然说法不一，但归根结底，其原因便是语言差异。

翻译歧异性的相关概念有翻译错配（mismatches）和不对应（non-correspondence）。先说翻译错配。Kameyama et al.认为，当某一语言需要区分的语法在另一语言的语法里没有加以区分时，翻译就产生了错配[7]，或者说，翻译错配就是目的语中无字符串来对应源语的信息表征。比如英语中可数名词有“数”（number）的区别，但日语没有，而日语要求的敬语表达英语中也常常缺乏，翻译时就会产生错配现象。错配不仅表现在语法形式上，文化因素的影响更为显著。比如汉语中“功夫”“太极”“五行”等在英语里都没有对应的字符串。Kameyama et al.认为，翻译歧异性只是词汇错配的一种。Barnet et al.也对翻译歧异性和翻译错配进行了分析比较。他们认为，翻译歧异性情况下，译文与原文表达同一个意思，但因从不同角度来阐述，最自然的译文与原文在某些方面会有差异，如句子结构。换句话说，译文与原文的语义解释（semantic interpretation）一致，但语义内容（semantic content）会有不同。而在翻译错配的情况下，因源语与目的语词汇不对称或在名词的性、数、动词的时态、体态、礼貌用语等方面的不同，译文表达的信息与原文不完全一致，或增多，或减少。Barnet et al.认为翻译歧异性与翻译错配都是典型的语际迁移问题，但是在很多情况下，前者包括后者[8]。

另一个相关概念是不对应（non-correspondence）。Xu & Li以汉译英为例，将其定义为翻译中信息的增加、删减或其他结构、语义上的调整[9]。该现象源于汉英词法、句法的差异。汉语以字为基本语言单位，而英语则是以词为单位，英语一个单词往往对应汉语若干个汉字，如smoke往往译成“吸烟”二字；同时，汉语中词的划分，尤其是多字词的划分界线较为模糊，因此在翻译时较难实现完全的对应，难免有增加、删减的情况，与印欧语系语言对译时尤为如此。另外，由于汉语属于意合语言，汉语主语、宾语可以从上下文中直接推断出来，有时并非必需，其他句法成分顺序与英语也不尽相同，因此翻译时会有结构和语义调整。如此一来，便体现在翻译的歧异性上了。

翻译不仅仅是语言之间字词、句段的一一对应，还与语境、社会习俗紧密相连。不同于人类使用的自然语言，MT程序处理的是人为设计的形式语言，将很多不定因素排除在外，因此非常缺乏意译（free translation）的能力。受歧异性的影响，MT在某些情况下经直译（literal/word-to-word translation）产生的译文生硬、尴尬，令人啼笑皆非。因此，翻译歧异性成为MT的关键问题。

三、翻译歧异性类别

在长期关于翻译歧异性的研究中，学界形成的一个共识是，想要解决 MT 中的歧异性问题，就先要对不同的歧异性进行识别和分类，这样才能分门别类，针对性地编写算法供机器学习，最终提高 MT 的质量。

对翻译歧异性最早进行系统研究的当属 Bonnie J. Dorr。Dorr 首先在研究英语、德语、西班牙语之间的翻译时，提出五类歧异性，并分别通过举例进行定义，其中包括，结构歧异性（structural divergence）指英语中作动宾的名词短语译成西语中的介词短语；合并歧异性（conflational divergence）指德译英时英语单个词的词义等于德语两个词的组合；词汇歧异性（lexical divergence）指英译西中表达同一个意思的谓词不一样；范畴歧异性（categorial divergence）指英译德中谓语由形容词变成名词；主题歧异性（thematic divergence）指英语中的宾语译作西语中的主语。其中，主题歧异性又包括论元调序（reordering of arguments）和谓词调序（reordering of predicates）两大类。论元调序较为简单，即源语与目的语中主语与宾语的互转。谓词调序则包括晋级（promotion）和降级（demotion）两种情况，晋级指源语作补语的成分译作目的语中的谓词，而降级则相反，源语中的谓词译作目的语中的修饰成分[10]。

Dorr 在原来的基础上，将主题歧异性下属的谓词调序所包含的晋级与降级两种情况独立出来，将原来的五类歧异性发展为关于词汇——语义歧异性的七大语言学分类，并将原来通过例子进行的解释进行提取概括，作了更为抽象的定义，分别为：主题歧异性，指论元换位；晋级歧异性（promotional divergence），指逻辑上的修饰成分经翻译后成为主动词；降级歧异性（demotional divergence），指逻辑主语经翻译后作为内部论元；结构歧异性，指句子成分之间的逻辑关系的改变；合并歧异性，指源语必需的论元经翻译后，其意义被合并入目的语的单个谓词中；范畴歧异性，指源语转换为目的语后，作谓语的词类不一样；词汇歧异性，此项歧异性可视为前六类的附带效应，因为任意一种歧异性必然引起词汇选择上的改变[11]。

除词汇——语义歧异性外，Dorr 还讨论了由句法、惯用语、篇章知识、专业领域知识或其他常识引起的翻译歧异性[12]，并就这些语言知识与词汇——语义歧异性的关系作了分析讨论[13]。但这些讨论并未形成一个相对完整的系统框架，其应用性也不如词汇——语义歧异性那样适用于分析多种语言，因此后人借鉴讨论较少。

Dorr 所提的关于词汇——语义歧异性的七大语言学类型有相当重要的理论指导意义。然而，由于 Dorr 的研究仅以英、西、德三种语言为基础，不一定适应于其他语言，难免存在局限性。随后，大量的实证研究采用此分类作为框架进行探讨，不仅证实了 Dorr 的观点，同时也在各自的基础上针对歧异性分类提出了改进与完善。Mahesh et al. 认为，Dorr 提出的分类无法将所有的歧异性包括在内，“只是探索复杂的翻译歧异性领域的一个开始”[14]。他们的研究以英语与北印度语之间的翻译为例，关注因语法系统相异而引起的歧异性，指出结构歧异性所涉及的内容应当扩展。Mishra & Mishra 将歧异性按照内容分为传统型与典型型。传统型即 Dorr 提出的七大分类，典型型则包括语言学、社会语言学、心理语言学、连词与小品词的作用、分词、动名词以及社会文化等方面的内容。另外，他们认为歧异性现象因语言而特异，某类型的歧异性不会发生在所有语言之间[4]。Sinhal & Chandak 认为歧异性可分为两大类，即词汇——语义歧异和句法歧异，两者是互补的，前者亦即 Dorr 的分类，完全由词汇属性造成，而后者与词汇的实际使用无关，是由语言的句法属性决定[5]。

四、翻译歧异性解决方案的研究成果

为了解决 MT 中的歧异性问题，各国学者与专业研究人员从理论、技术等各方面作了大量的尝试与努力。以下研究成果按照发表时间先后顺序排列。

Habash & Dorr[15] 认为，解决歧异性问题的前提是源语和目的语之间有明显对称的语言知识，需要词汇、结构等信息的匹配。他们介绍使用 GHMT（generation-heavy machine translation）系统来解决翻译歧异性问题，该方法无需转换规则或复杂的中介语表征，而是依赖于目的语的词汇意义、范畴变化、次范畴框架等信息，将源语进行句法分析后与目的语进行匹配。

Mahesh & Sinha[16]指出目前机器翻译的研究开发人员虽然已经意识到识别并解决翻译歧异性问题非常关键，但是要设计一个整体的方案还是存在很大困难。相对来说，制定一个针对于某一个语言的方案更为可行。因此，他们利用北印度语的构词法，以动词的曲折变化为例，提出一项技术方法用以辨识某些歧异性模式并提供方法解决英译时的歧异性问题。

Goyal & Sinha[17]采用 Dorr 所提的关于翻译歧异性的分类作为讨论的基础框架，举例分析英语与梵语、北印度语与梵语两对语言之间的歧异性，同时讨论了框架外的歧异性模式，包括英语与梵语之间在语序影响语义、语态、动名词和分词小句实现形式、形态、敬语以及时间表述等方面的差异。他们认为，需要尽量多地研究各种语言之间的歧异性，认识各种歧异性模式，这样才有助于解决翻译时由歧异性导致的问题，促进机器翻译的发展。

Mishra & Mishra[4]举例讨论了英语与梵语之间与连词、小品词、分词、动名词以及社会文化相关的歧异性，并介绍了他们研发的英译梵机译系统。该系统由两部分组成：基于规则的模型和通过人工神经网络模型进行的词典匹配。并通过实验说明该系统能解决英译梵中大部分与连词、小品词、分词、动名词以及社会文化相关的歧异性问题。

Saboor & Khan[18]采用 Dorr 的分类对乌尔都语与英语之间的六类歧异性进行举例分析，并针对 EBMT（Example Based Machine Translation）介绍了一项算法用以识别这些歧异性，同时建议用<DIV>对歧异性句子进行标注。

Sinhal & Chandak[5]认为机器翻译往往需要调整，识别歧异性则是进行调整、实现有效翻译的关键。歧异性问题总是需要针对性的解决措施。建议可采用框架转移法（framing transfer rules）或参数化映射（parameterized mappings）来解决机器翻译中的歧异性问题。

Kulkarni et al.[19]认为制定一个一般应对方案将所有语言之间的歧异性问题都解决是不太现实的，应着眼于具体语言之间的问题，因此讨论英语与马拉地语之间主要因句法结构规则相异而引起的歧异性问题。

Feng[20]举例介绍了中、英、西、德等语际翻译时存在的几类歧异性问题，包括词汇选择、时态、主题关系、结构、词类等方面。针对这些问题，作者建议使用基于共现簇（co-occurrence cluster）的方法来解决。若机器翻译的歧异性问题没那么大，作者又提出若干建议，包括在源语与目的语结构相似的情况下不关注句子意义，按目标设定不同的接受阈值，半自动标注，控制原文歧义，与其他文本处理技术结合等，以此来减少甚至解决歧异性问题。

纵观前期研究，可获得如下一些启示：第一，理论研究对于 MT 中歧异性问题的解决有不可替代的重要作用。尽管 MT 研究需要系统、程序的完善，但是翻译从根本上说是与文字打交道的过程，尤其是对于需要高度形式化语言的机器来说，更离不开语言学理论的支撑。因此在开发机译系统的同时，不能忽视理论研究。第二，具体问题具体分析。各国学者对歧异性问题的探讨都是着眼于具体的语言差异，并且针对性地设计解决方案。第三，MT 的发展需要跨学科的共同努力，包括语言学、计算机、认知科学、信息论等各个学科，仅靠其中任意一个都无法实现 MT 的长足发展。

五、翻译歧异性研究反思

语际差异的存在是无法消除的客观事实，因而歧异性是翻译必须面对的问题。然而，作为研究对象的语言是何其复杂，加之全世界语言数量庞大（据德意志民主共和国出版的《语言学及语言交际工具问题手册》统计，现今世界上有 5651 种语言），在人们开始自觉去认识歧异性现象的时候，想要用一个完善的理论模式概括出歧异性现象的本质，难免具有局限性，因而争论和理论模式存在缺陷是不可避免的。研究之初，学界便认识到，形成能应用于世界所有语言的统一的理论模式（universal mode）是不可能的，便逐渐把主要精力转移到对歧异性现象的具体分类上，针对性地开展研究。

国外进行翻译歧异性研究一大优势是，目前所研究的各种语言大都属于印欧语系，如印度的梵语、欧洲的希腊语、拉丁语。它们大都是屈折语，广泛利用词缀和词干元音音变来表达语法意义，显示出系统的相似点，如名词和大部分形容词都有性、数、格的变化；动词词根相似且大都有时

态、语态和语体的变化，主语和动词在变化中互相呼应；等等。虽然这些语言历经发展，对这些特点的保留程度不一，有些保留得较为完整，如德语、俄语，有些在形态稍有简化，如英语，但这些系统的相似性还是为各语言之间的翻译歧异性研究提供了基础，方便形成一个系统的理论模式。

相比之下，跨语系的翻译歧异性研究就显得较为困难。尤其是汉藏语系本身的特点没有统一，所含的语言类型各异，千姿百态，虽然经过各国语言学家的努力，这些语言的特点陆陆续续、深浅不一地被了解，但是它们各自的演化脉络还不清楚，各语言群体之间的共性和特性以及远近关系也还很不清楚，至于整个语系的特点和演变脉络更是难以把握[21]。在这样的基础上，形成跨语系的翻译歧异性的系统理论还任重道远。

六、结语

机器翻译之所以要关注歧异性，根本原因是机器无法像人一样在翻译过程中采用意译。机器翻译的现实情况往往是英语的输出结果较中文输出更为通顺合理，这也说明机器需要形式化的语言学知识。然而，国内机器翻译研究一直以来更为关注机译系统本身的发展，大量研究侧重系统测评、模型建立、算法匹配、译后编辑工具研发等，而对于理论的应用关注较少。相比于英语，汉语很多的语法规则缺乏直接的表现形式，但是这并不等于汉语不能形式化。针对机器翻译的理论探索还需进一步努力。

【参考文献】

[1]Johnson R, King M, Des T L. *Eurotra: a multilingual system under development* [J]. Computational Linguistics, 1985, (2-3):155-169.

[2]Dorr B J. *Machine translation divergences: a formal description and proposed solution* [J]. Computational Linguistics, 1994, 20(4):597-633.

[3]Dorr B J, Voss C R. *Constraints on the space of MT divergences* [C]. Building Lexicons for Machine Translation, Papers from the 1993 Spring Symposium, Technical report SS-93-02, Stanford University, Stanford, California:43.

[4]Mishra V, Mishra R B. Divergence patterns between English and Sanskrit machine translation [EB/OL] (2009-8-17) [2017-3-21]. https://www.researchgate.net/publication/228733192_Divergence_patterns_between_English_and_Sanskrit_machine_translation

[5]Sinhal R A, Chandak M B. *Divergence: A Challenge in Example Based Machine Translation* [M]. Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011. 2012:805-812.

[6]Lin S C, Wang J C, Wang J F. (2004). Translation Divergence Analysis and Processing for Mandarin-English Parallel Text Exploitation[EB/OL]. (2017-3-22). http://anthology.aclweb.org/O/O05/O05-1029.

[7]Kameyama M, Ochitani R, Peters S. *Resolving translation mismatches with information flow*[C]. Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1991: 193-200.

[8]Barnett J, Mani I, Rich E. *Reversible Machine Translation: What to Do When the Languages Don't Match Up* [J]. Reversible Grammar in Natural Language Processing, 1994:255, 321-364.

[9]Xu J, Li X. *Structural and semantic non-correspondences between Chinese splittable compounds and their English translations: A Chinese-English parallel corpus-based study* [J]. Corpus Linguistics & Lingustic Theory, 2013, (1):79-101.

[10]Dorr B J. *Solving thematic divergences in machine translation*[C]. Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1990:127-134.

[11]Dorr B J. *The use of lexical semantics in interlingual machine translation*[J]. Machine Translation, 1992, (3):135-193.

[12]Dorr B J. *Interlingual machine translation: a parameterized approach* [M]. Natural language processing. MIT Press, 1994:429-492.

[13]Dorr B J, Voss C R. *Machine translation of spatial expressions: defining the relation between an interlingua and a knowledge representation system* [C]. Eleventh National Conference on Artificial Intelligence. AAAI Press, 1993:374-379.

[14]Mahesh R, Sinha K, Thakur A. *Translation Divergence in English-Hindi MT* [J]. Eamt, 2005: 245-254.

[15]Habash N, Dorr B. *Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation*

[C]. Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research To Real Users. Springer-Verlag, 2002:84-93.

[16]Mahesh R, Sinha K. Using rich morphology in resolving certain Hindi-English machine translation divergence[EB/OL]. (2017-3-25). http://www.mt-archive.info/MTS-2007-Sinha.

[17]Goyal P, Sinha R M. *Translation Divergence in English-Sanskrit-Hindi Language Pairs*[C]. Sanskrit Computational Linguistics, Third International Symposium, Hyderabad, India, January 15-17, 2009. Proceedings. DBLP,2009:134-143.

[18]Saboor A, Khan M A. *Lexical-semantic divergence in Urdu-to-English Example Based Machine Translation*[C]. International Conference on Emerging Technologies. IEEE,2010:316-320.

[19]Kulkarni S B, Deshmukh P D, Kale K V. *Syntactic and Structural Divergence in English-to-Marathi Machine Translation*[C]. International Symposium on Computational and Business Intelligence. IEEE, 2013:191-194.

[20]Feng, Z. Translation divergence in machine translation[EB/OL]. [2006-03-20](2017-04-01). http://lingviko.net/feng/forum-fzw.

[21]孙宏开. 汉藏语研究中的一些问题[J]. 语言科学, 2006,(1):49-51.

[责任编辑:王敬儒]