DOI:10.3969/j.issn.1673-4785.201211023
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20130515.0927.005.html

引入复述技术的统计机器翻译研究综述

胡金铭[1,2],史晓东[1,2],苏劲松[3],陈毅东[1,2]
(1.厦门大学 信息科学与技术学院,福建 厦门 361005; 2.厦门大学 福建省仿脑智能系统重点实验室,福建 厦门 361005; 3.厦门大学 软件学院,福建 厦门 361005)

摘 要:基于对引入复述技术的统计机器翻译研究现状的分析,提出具有研究价值的课题方向.首先归纳了复述的概念,总结了引入复述技术的统计机器翻译各类方法.然后对复述知识在统计机器翻译中的模型训练、参数调整、待译语句改写和机器翻译自动评测等方面应用的主流方法进行了概括、比较和分析,说明了复述与统计机器翻译是紧密相关的,强调了复述在统计机器翻译应用中的关键问题是复述的正确性和多样性.最后指出提高复述资源的精确度、建立复述与机器翻译的联合模型、采用新方法解决稀疏问题等是有待进一步研究的课题.

关键词:复述技术;机器翻译;统计机器翻译
中图分类号: TP391 **文献标志码**:A **文章编号**:1673-4785(2013)03-0199-09

中文引用格式:胡金铭,史晓东,苏劲松,等.引入复述技术的统计机器翻译研究综述[J].智能系统学报, 2013, 8(3): 199-207.
英文引用格式:**HU Jinming, SHI Xiaodong, SU Jinsong, et al. A survey of statistical machine translation using paraphrasing technology[J]. CAAI Transactions on Intelligent Systems**, 2013, 8(3): 199-207.

A survey of statistical machine translation using paraphrasing technology

HU Jinming[1,2], SHI Xiaodong[1,2], SU Jinsong[3], CHEN Yidong[1,2]
(1. School of Information Science and Engineering, Xiamen University, Xiamen 361005, China; 2. Fujian Key Laboratory of the Brain-like Intelligent Systems, Xiamen University, Xiamen 361005, China; 3. College of Software, Xiamen University, Xiamen 361005, China)

Abstract: In this paper, the research team discussed possible new prospective research directions of paraphrasing technology in statistical machine translation (SMT), based on reviews of state-of-the-art technology. First the research team introduced the concept of paraphrases, and next a summarization of the latest progress utilizing paraphrasing technology in SMT was conducted. Finally, conclusions were drawn, data was compared and an analysis of the main issues of incorporating paraphrases into SMT, including translation model training, parameter tuning, input sentences rewriting and machine translation evaluation was performed. The results proved that there is an inherent connection between paraphrasing and SMT. The results also point out that the correctness and diversity of paraphrasing are the key issues to apply paraphrasing to SMT. It was highly noted that the improvement in the quality of paraphrasing resource, the establishment of a joint model of paraphrasing and machine translation and the new proposed approach to solve data sparseness are problems which need further study.

Keywords: paraphrasing technology; machine translation; statistical machine translation

机器翻译(machine translation, MT)是利用计算机程序,实现从一种自然语言到另一种自然语言的翻译.它属于计算语言学(computational linguistics)的范畴.经过数十年的研究,机器翻译在理论和实践方面都有了较大的进步.从方法论的角度来看,目前的主流研究使用基于统计的方法.统计机器翻译(statistical machine translation, SMT)是通过对大量双语平行语料库的统计分析来构建统计翻译模型,

收稿日期:2012-11-16. **网络出版日期**:2013-05-15.
基金项目:国家科技支撑计划资助项目(2012BAH14F03);国家自然科学基金资助项目(60573189,61005052);福建省自然科学基金资助项目(2006J0043).
通信作者:史晓东. E-mail:mandel@ xmu.edu.cn.

并使用该模型进行翻译.早期的研究使用噪声信道模型[1-2],当前的主流统计模型是对数线性模型[3].对数线性模型由若干特征组成,每个特征都反映了翻译概率的一个方面,该模型由于可以包含更多的反映翻译概率的信息而受到了广泛关注.从事机器翻译研究的学者正尝试将不同的语言学、统计学特征加入到对数线性模型中,使翻译系统更加强大.而反映语言多样性的复述技术(paraphrasing technologies)也被用来改善机器翻译的效果.

随着自然语言处理各项底层技术的不断成熟和发展,复述(paraphrases)作为自然语言处理中一种非常普遍的现象,受到了越来越多研究者的关注.刘挺[4]、赵世奇[5]等国内学者也都对复述技术研究进行了详细综述.很多学者试图给复述一个精确的定义,早在 20 世纪 80 年代,语言学家 Halliday 和 De Beaugrande 等认为复述是“概念上的近似等价”,但互为复述的 2 个语言片段的可替换程度(interchangeability)始终没有确切的标准[6-7]. Barzilay 等[8-9]把复述看作传达相同信息的可替换形式. Glickman 等[10]则认为复述现象反映了语言多变性的核心,复述是对应到相同意义的等价表达.鉴于上述观点,笔者认为复述就是在同一种语言内有相同语义但有不同表达形式的语言片段,它反映了人类语言的灵活多样性,同时也为自然语言处理的研究难点提供了更多的解决方法.

统计机器翻译的实质是对大规模的双语语料进行统计,提取有助于文本翻译的规则.这些规则使得翻译系统可以较好地处理字面上的直译,但其并没有真正意义上的意译能力,即无法翻译未知文本.随着时间的推进,科技发展、知识增长,语言也在不断地进化,不可能存在包含所有语言现象的语料库.然而,复述技术可以将未知文本片段转化成语料库中出现的同义表述;那么,适时地引入复述技术便可以提高翻译系统的性能.

目前由于统计机器翻译的研究热点是对数线性模型,因此将复述技术引入统计机器翻译的研究也多数围绕对数线型模型展开.基于对数线性模型的统计机器翻译大致可以分为 4 个阶段:翻译模型的训练、特征参数的调整、译文的搜索解码、翻译质量的自动评价.本文介绍了复述与统计机器翻译的概念,并对复述技术与统计机器翻译中各个阶段内容的联系进行概述,最后对引入复述技术的统计机器翻译研究进展及前沿课题进行分析评述,概括并凝练出具有研究价值的课题方向,希望对统计机器翻译领域的研究有所裨益.

1 复述在统计机器翻译中的研究现状

近年来,许多学者将复述应用到信息抽取、文本生成、自动问答、自动文摘等多个相关研究领域中.如图 1 所示,对复述在自然语言处理的部分子课题中的文献资料做粗略统计(数据来自 Google 学术搜索),可见,其中讨论得最为广泛的是复述在机器翻译研究中的应用.

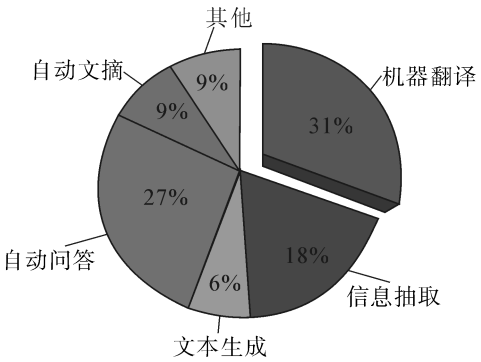


图 1 复述在自然语言处理子课题的应用统计

Fig.1 Statistics of using paraphrases in sub-subject of NLP

复述是单语同义文本的表达形式转换,而机器翻译则是跨语言同义文本的表达形式转换.它们的共通性也使得机器翻译中的理论和方法可以用于解决复述问题,因此有基于 MT 的复述生成方法[11-13].同样,复述技术也可以解决机器翻译问题.

在 21 世纪初,机器翻译中基于统计方法逐渐趋于主导地位.在研究过程中,越来越多的学者发现语料资源不足会极大影响统计翻译系统的翻译质量,复述便成为了一个解决办法.复述可以从更为广泛的语料中获取,如同义词词典、单语可比语料、单语平行语料等,更多的单语知识可以改善翻译系统性能.从方法角度上讲,将复述引入到统计机器翻译的研究集中在改进其 4 个阶段,引入到前 3 个阶段是为了提升翻译效果,而对于自动评测主要是为了提升机器评价和人工评价的一致性.为了更直观地对比前 3 种途径翻译效果的提升程度,图 2 列出了各方法在 BLEU 值上的提升比.因为各学者选取的实验数据并不一致,结果对比可能略有出入.但从图 2 中可以发现,对待译语句的改写可以更好地提升翻译质量(图中的参数调整部分,因为数据都来自 Madnani 的研究,故命名为“年份.人工参考译文数量”,“H”前的数字表示开发集的人工参考译文数量).下面从 4 个方面分别介绍引入复述的统计机器翻译研究的国内外发展现状.

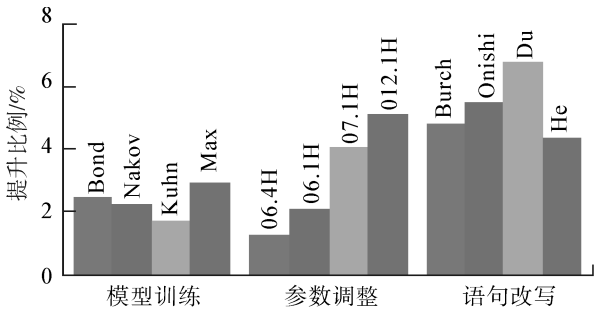


图 2　各方法效果对比

Fig.2　Comparison with BLEU on various methods

1.1　复述改善模型训练

训练数据不足会引起数据稀疏,引入复述知识,对已有的训练数据或者规则表进行处理可以改善这一问题.通常有 2 种途径:1)对训练数据的平行句对生成复述从而扩充训练数据的规模;2)利用短语间的复述关系平滑翻译模型的概率估计使其更加准确.

统计机器翻译的模型训练是通过大规模的双语平行语料获得.由于语言的多样性,训练集不能覆盖所有的语言现象,对稀有语种而言尤为明显.当无法直接获得更多训练语料时,研究者利用复述技术扩充训练集的规模,提高模型的覆盖率.基本思想是对双语平行句对(f,e)的源端f生成句法等价的句级复述f',f'与目标端e重新组合构成新句对(f',e)加入到训练集中.Bond 针对词序、时态等语言学现象并结合句法信息生成复述[14].Nakov 则对名词短语进行复述[15],首先识别句中的名词短语,利用人为定义的包含句法信息的复述规则,仅当句子中发现符合复述转换规则结构的名词短语时才生成复述.Nakov 不但扩充训练集,还对已训练的规则表进行类似实验,结果表明对短语表进行复述并没有对训练数据进行复述的效果好.这是因为规则表是经过分词、对齐等前序步骤后得到,其中已含有噪声;同时对规则表复述没有考虑句法信息及上下文信息,新生成的翻译规则可能并不合理.

短语概率作为 SMT 的一个非常重要的特征,传统方法使用最大似然估计,通过词频的累加来计算,如式(1)所示,式中#表示频次统计.这种方法的不足之处是,当短语出现次数较少时,其概率估计会出现较大误差.Kuhn 和 Max 引入复述技术来进行平滑翻译模型概率估计的研究.

$$P_{RF}(e_i \mid f)=\frac{\#(f,e_i)}{\sum_j \#(f,e_j)}. \tag{1}$$

Kuhn 利用短语聚类来进行平滑处理[16],如式(2)所示:

$$P_{PC}(e \mid f)= P(e \mid C(e)) \times P(C(e) \mid C(f))= \frac{\#(e)}{\#C(e)} \times \frac{\#(C(e),C(f))}{\#C(f)}. \tag{2}$$

式中:$C(e)$、$C(f)$分别代表目标端和源端的短语类.研究者认为复述片段含义相同,不应分别进行概率估计,应对同类短语一并计算.可以验证,当P_{RF}为 0 时,P_{PC}不为 0.所以当e出现的频次很小时,P_{PC}会有更好的概率估计.他提出了利用基于短语共现次数和基于词序的 2 种相似度计算来进行短语聚类的方法,获得了很好的效果.

Max 针对短语概率估计提出了 2 个观点:1)一个合适的短语需要更多地参与到概率估计;2)复述可以用来优化概率估计[17].他利用源端f的上下文相似度的计算代替传统的频次统计,上下文相似度偏低的短语,其概率的估计也会较低,则相应译文可取度降低.如式(3)所示:

$$P_{\text{cont}}(e_i \mid f)=\frac{\sum_{\langle f_k,e_i\rangle}\text{sim}(\text{Cont}(f),\text{Cont}(f_k))}{\sum_{\langle f_k,e_j\rangle}\text{sim}(\text{Cont}(f),\text{Cont}(f_k))}. \tag{3}$$

$$P_{\text{para}}(e_i \mid f)=\frac{\sum_{\langle p_k,e_i\rangle}\text{sim}(\text{Cont}(f),\text{Cont}(p_k))}{\sum_{\langle p_k,e_j\rangle}\text{sim}(\text{Cont}(f),\text{Cont}(p_k))}. \tag{4}$$

式中:f是测试集中待译的源短语,f_k是f在训练集中出现的第k个特例,e_j表示f_k的所有可能译文,e_i是f_k的特定译文,Cont(f)是指f的上下文.P_{cont}通过比较测试语句中短语f的上下文与译文为e_i的特例f_k的上下文的相似度,来估计e_i是f译文的概率.式(4)利用复述对式(3)进行补充,作为另一个特征加入到模型中.p_k是f的复述,$\langle p_k,e_i\rangle$是训练集中的短语对.同样,考虑上下文信息来估计e_i是f译文的概率.式(3)解决了 Max 提出的第 1 个问题,使上下文信息更接近短语主导概率的估计,式(4)则缓解了上下文种类较少带来的数据稀疏问题.

1.2　复述提高调参效果

目前统计机器翻译的参数调整大多采用最小错误率训练方法[18].通常使用基于n元组匹配的 BLEU[19]等评测指标作为最小错误率.因此在调参过程中所使用的开发集规模越大、多样性越强、参考译文数量越多,n元组匹配的准确性就越高,调参的效果也就越好.基于这个思想,Madnani 引入复述知识,对开发集的参考译文进行扩展,来增加参考译文的多样性[20].首先,利用层次短语系统训练出双语

层次规则[21],如式(5)~(7)所示;其次,利用基于枢轴法(pivot-based)的复述获取,抽取单语层次规则,如式(8)~(9)所示.

$$X \rightarrow < X_1 \text{建} X_2; X_1 \text{ to build } X_2 >, \quad (5)$$

$$X \rightarrow < X_1 \text{建} X_2; X_1 \text{ to construct } X_2 >, \quad (6)$$

$$X \rightarrow < X_1 \text{建} X_2; X_1 \text{ to formulate } X_2 >, \quad (7)$$

$$X \rightarrow < X_1 \text{ to build } X_2; X_1 \text{ to construct } X_2 >, \quad (8)$$

$$X \rightarrow < X_1 \text{ to build } X_2; X_1 \text{ to formulate } X_2 >. \quad (9)$$

获得单语层次规则后建立单语的翻译模型,通过该模型的解码对已有的人工参考译文进行复述扩展,并加入到开发集中进行调参.2007 年 Madnani 又做了进一步补充[22],生成参考译文的 n-best 复述译文,并利用启发式规则进行过滤.但经实验发现,n 取到 3 以上便会由于复述带来的噪声使得调参效果变差.针对这些不足,Madnani 在 2011 年提出细化复述生成过程[23],在不改变参考译文原意的前提下使生成的参考译文复述和机器译文有尽可能多的字面匹配,并将其加入到在线调参过程中,使翻译质量有所提高.虽然 Madnani 在不断细化复述的生成,但其方法仍有几点不足之处:1)用单语翻译解码来生成复述句,缺少对一些错误复述的过滤;2)由于单语开发集的稀缺,单语翻译模型调参的准确性有待考证;3)词语对齐、复述生成、单语翻译等前序步骤带来的噪声传播也会对翻译产生负面影响;4)n-best 多样性随 n 的数目增加逐渐变小,而引入的噪声却起了主要作用.

1.3 复述改写待译语句

利用复述生成技术,对机器翻译系统的输入语句进行改写[24-25].尤其是对于口语翻译而言,将形式灵活且不规范的口语语句改写为规范的书面语语句,无疑会降低翻译系统的处理难度[26-27].对于资源不足的语言对,翻译系统无法翻译出包含未知词汇的待译语句,却能够翻译意义相近而没有未登录词的复述句.因此可以引入复述知识,改写待译语句,使系统能够翻译原本无法翻译的句子.

Callison-Burch 提出使用枢轴法获取复述来替换待译语句中未知的词和短语,并使用该复述的译文作为翻译结果[28]. Marton 也开展了类似的研究[29],不同的是 Marton 从单语语料获取复述,他们的研究局限于只替换待译语句的未知片段.这种不考虑句法信息的替换极有可能导致复述语句语法不通、语序不畅、语义混乱.Mirkin 则利用 WordNet 得到文本蕴含和复述规则,利用上下文模型对复述打分,翻译前 k 条规则生成的复述,并用语言模型为译文进行打分,最后选择分数较高的译文[30].其优点是不仅利用了人工知识 WordNet,还利用上下文判断复述句是否合理,避免盲目改写,但缺点是系统解码过程更加复杂.

Onishi 和 Du 利用短语级复述构建待译语句的复述词图(word lattice)[31-33].图 3 是“the exercise will continue”的词图结构,图中双圆圈和实线箭头分别代表待译语句最初的节点和单词,单圆圈和虚线箭头分别代表复述扩充的节点和单词.

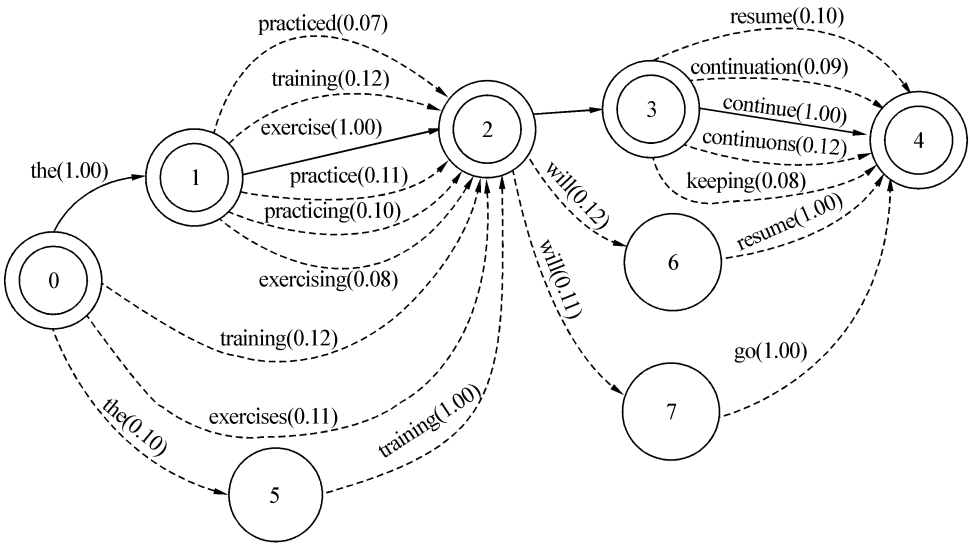


图 3 输入语句的复述词图构建示例[32]

Fig.3 An example of how to build a paraphrase lattice for an input sentence[32]

构建词图的好处是不用区分待译语句中的未知词和已知词,而是让翻译系统的解码器根据词图自行搜索最优翻译结果,提高容错性.这样可以构造比 Callison-Burch 方法更为流利的复述输入语句,其缺陷在于构造词图时过多的边数会导致复杂度成倍提升.此外,部分不当替换不但会增大词图的搜索空间而且也不能改善翻译效果,需进行适当的剪枝.He 的研究[34]与 Du 相似,他采用一种正向翻译与反向翻译相结合的方法获取复述.正向翻译就是源端到目标端的翻译过程,反向翻译则是目标端到源端的翻译过程.他利用一次正向翻译的译文 T_1 和经过反向翻译后再正向翻译的译文 T_2 作为抽取复述的单语平行语料,然后通过启发式规则的过滤,利用层次短语系统的规则抽取方法[21]构建复述规则,之后再构造词图.这样做的好处有 2 点:1)不但生成了词和短语级的复述,而且可以生成句级的复述;2)因为复述规则由翻译系统得来,对于部分病态复述,经翻译系统的病态处理,会意外获得质量更好的结果,这也体现了复述和统计机器翻译融合的思想.

Resnik 通过迭代修改待译语句来解决翻译质量较差的问题[35].其方法是,对翻译系统的翻译结果进行评判,将译文中翻译较差的片段所对应的源语句片段进行复述,构造出新的输入语句,新输入语句的译文要优于原译文.该方法较 Callison-Burch 的方法[28]能更针对地构造复述,利用 TERp 中定义的多种操作来判断哪些片段应该构造复述.

1.4　复述改善机器翻译自动评测

机器翻译的自动评测一直是机器翻译研究中的难点,目前最为广泛使用的指标是 BLEU[19],它计算机器译文和参考译文间 n-gram 的匹配准确率,将其加权得到评价分数.很多学者基于 BLEU 指标改善机器翻译的自动测评.Kauchak 调查发现,NIST2004 测试集中每个句子的参考译文两两组成句对,其中 0.2%是字面完全一致的,60%至少 11 个词不同[36].这就意味着,如果参考译文的数量有限时(1 ~ 3 句),那么基于字面匹配的自动评测永远不可能达到人工评测的水平.因此,Kauchak 提出应该使参考译文更多地包含机器译文的词或短语,而这也是早期学者们改善评测技术的主要手段.他利用 WordNet 从参考译文和机器译文中识别可能构造复述的词对,测试候选复述是否在参考译文的上下文中可采纳,然后生成参考译文的复述,达到增加参考译文数量的目的.Kanayama 考虑日语相比英语更多样性和胶合性[37],利用人工定义的复述规则加以形态学分析,生成参考译文的复述,构造更多的参考译文,来提高自动评价和人工评价的一致性.因为人工定义的复述规则中没有实词的替换规则,所以该方法减少了内容词替换带来的任意性;但只能处理功能词和日文语气词,有一定局限性.Lepage 利用类似复述模板的方法生成参考译文的复述集,丰富参考译文的表达[38].Zhou 则针对 BLEU 没有考虑召回率和缺少对复述匹配的支持来进行改善,提出了基于 BLEU 的 ParaEval 评测方法[39],对 1-gram 的匹配进行修改使其支持了复述匹配,并使用单参考译文计算召回率.

Russo-Lassner 对$(\boldsymbol{x}, h)$训练线性回归模型,其中 $\boldsymbol{x}$ 是一个代表机器译文和参考译文句对间一致性的特征向量,h 是对机器译文的人工评分[40].他将机器翻译自动评测任务看作复述识别,即对比机器译文与参考译文之间的词汇、句法信息的变化,因此特征选择包括词干共现、WordNet 同义词集、动词语义类等.

Snover 基于其在 2006 年提出的 TER 评测指标,融合了可调参数、形态学分析、同义词以及复述之后,提出新的评测指标 TERp[41-42].TERp 不但将参考译文和机器译文字面相同的片段匹配,还将有相同词干或同义词的片段匹配.TERp 保留了 TER 的编辑操作——匹配、插入、删除、替换、移动,还增加了词干匹配、同义词匹配、短语替换,使评价结果与人工评价的一致性更高.

Pado 将文本蕴含(textual entailment)用在机器翻译的评测中[43].蕴含被定义为一个前提 P(premise)和一个假设 H(hypothesis)之间的二元关系,即若已知前提 P 成立可以推出 H 为真,则说 P 蕴含 H.研究者一般将复述看作蕴含的特例,因为复述是双向的,而蕴含的推理是单向的.举例说明:设 P 为“Jane is a French teacher”,H 为“Jane can speak French”,则 P 蕴含 H,H 可从 P 中推理出来,相反 P 不一定能从 H 推理出来.Pado 认为好的机器译文与参考译文是双向蕴含的,机器译文内容的缺失会破坏正向蕴含,而机器译文内容的增添又会打破反向蕴含,如果双向蕴含都不成立则认为翻译结果较差.蕴含识别可以包含更多的语义和语法知识,利用蕴含信息的“深度”匹配自然会优于简单的字面匹配评测标准.

2　复述在统计机器翻译中的应用分析

复述作为人类语言中的一个普遍现象,受到自然语言处理界学者的广泛关注.尤其在机器翻译领域,在不同的阶段引入复述技术,在一定程度上改善了翻译质量.鉴于前人的研究工作,将复述引入机器翻译的不同阶段中,确实可以改善翻译结果.但在机

器翻译中引入复述的研究还处于初级阶段,有一定的局限性,并没有实质性地改变机器翻译的框架.笔者认为具体表现在如下几个方面:1)复述抽取的质量不高,由于错误的传播将间接影响到翻译结果的好坏;2)复述生成的多样性不充分,并没有达到利用复述来丰富表达形式的目的;3)现有工作还局限在已有的统计模型框架下引入复述知识,因此复述技术与统计机器翻译系统的整体融合还需进一步的研究.

在引入复述技术的统计机器翻译研究中,虽然将复述技术运用在统计机器翻译的不同阶段,但究其本质,主要是为了解决数据稀疏的问题.能否很好地提升翻译效果,笔者认为主要有 2 个关键问题需要解决:1)复述的正确性;2)复述的多样性.

2.1 复述的正确性

引入复述改善机器翻译系统的翻译质量主要取决于复述的正确性,复述的正确性又可以体现为复述生成的准确率.如表 1 所示,有 4 句复述句,其中句子(2)~(4)这 3 句是错误的,若将其全部用于机器翻译系统中,必定会产生负面影响.如何提高复述生成的准确率是复述能否提升翻译效果的一个至关重要的问题.就目前的研究而言,还没有很好的自动评测手段来判断复述生成的好坏.笔者认为复述的正确性与复述规则的正确性和复述规则的适用性相关.如表 1 所示,句子(1)是原句,句子(2)是利用“the movies->the films”复述规则生成的复述句.可以发现这条短语复述规则是正确的,但是生成的“go to the films”并不符合英语的习惯用法.没有考虑句中上下文、句法信息与简单地使用短语级复述规则是造成句子(2)错误的根本原因.而表 1 中句子(3)、(4)使用了错误的复述规则,因此产生了语法错误.

表 1 复述句实例

Table 1 Examples of paraphrase sentence

复述句	序号	正误
Everyone often goes to the movies.	(1)	原
Everyone often goes to the films.	(2)	错
Everyone goes often to the movies.	(3)	错
Everybody goes to the movies often.	(4)	错
Everybody often goes to the movies.	(5)	对

复述一般通过同义词典、语料库、互联网等获取.根据粒度不同,又分为复述句、复述短语、复述模板和复述搭配等,在统计机器翻译中引入的复述粒度一般是句级、短语级和复述模板.对于统计机器翻译中的不同阶段,不同粒度的复述选取会有不同的效果.如改善模型训练的方法,因为要生成训练语句的句级复述,通常不能使用短语级复述规则做简单的短语替换,如表 1 中的句子(2).因为对于短语级复述规则的使用可以不受句法约束,而句级复述如果句法错误,则会导致语义混乱.对于短语级复述,将其运用生成复述句必然会引入一些语法错误、语序不畅的问题.但是在构造复述词图时,就可以使用短语级复述,从而避免上述问题的出现.使用短语级复述构建词图,可以使词图包含更多信息,使用待译语句的词图进行翻译是依据搜索解码过程对译文的评判,因此对词图的解析不但可以提高容错性也可以提升多样性.复述模板和复述搭配都包含句法信息,而目前在统计机器翻译中引入复述模板和复述搭配的研究工作还较少.笔者认为复述模板和复述搭配的结构可以很好地与句法翻译模型相结合.句法翻译模型的过分细化使数据稀疏问题显得尤为严重,而复述的多样性可以很好地解决这一问题,又因为复述模板和复述搭配的结构与句法翻译模型相似,所以在句法翻译模型中引入复述模板和复述搭配将成为未来研究的重点,这也体现了机器翻译与复述融合的思想.可见对于不同的任务,恰当的复述粒度选取会有效提升复述的适用性.

复述规则有人工定义和自动获取 2 种方法.对于人工定义的复述规则,由于考虑到了各种语言学知识,规则自身都是正确的,问题只是规则的使用是否恰当.而对于自动获取的复述规则,由于统计的误差、语料库的覆盖度等因素,导致包含过多的噪声.这就需要一种合适的途径过滤掉噪声,不但过滤质量较差的规则,而且还能够对规则的使用,即复述的生成作一定的限制.上下文信息及句法信息的引入使复述质量得到了很好的改善.通过对比上下文相近的语句找到适用于相同语言环境的复述规则来获取和生成复述,使得语义不变;利用句法知识分析复述句,使得语法正确.

2.2 复述的多样性

数据稀疏是导致现有统计机器翻译系统的翻译结果不能令人满意的主要原因,笔者认为数据稀疏性的根源来自语言的多样性,又可体现为复述生成的召回率.因为机器翻译的训练数据无法包含所有的语言现象,如果能极大地提升复述的召回率,便可以使翻译系统的覆盖度尽可能扩大,从而提升翻译效果.笔者认为语言的多样性主要表现为个体性与进化性.

1)个体性.如表 2 所示,当用英文表达“请给我一杯啤酒”时,因为个人的习惯与口语的随意性,有多种结构完全不同的表达形式时.当训练语料库中

缺少下述某种表达形式时,那么翻译系统就不会翻译相应的文本.利用复述知识来阐述下述表达形式之间的关系,使翻译系统融入更多的单语知识,且对不同表达形式的同义句翻译有更好的处理,从而可以改善数据稀疏带来的问题.

表 2　复述多样性实例

Table 2　Examples of the diversity of paraphrase

多样性例句	序号
A beer, please.	(1)
Beer, please.	(2)
Can I have a beer?	(3)
Give me a beer, please.	(4)
I would like beer.	(5)
I'd like a beer, please	(6)

2)进化性.语言多样化不单表现在个体使用的不同,同时随着时代发展,语言整体也在不断地进化.新词、新的语法结构的诞生使语言的表达多种多样.目前网络语言日趋流行,每天都可能有新词诞生,或者是旧词生新义.表 3 列举了时下流行的新词以及对应含义相近的旧词.可以发现,所谓新词绝大多数都是已存在的词语,只是用一种新奇的字词组合来进行表达.这些新词,在过时的语料库中几乎不会出现或者出现次数很少,这就会导致概率估计不准确从而影响到翻译结果.同样,古代的词汇、语法也会随着时间的推移渐渐被遗弃,古文、古诗的语料稀缺使古汉语的翻译更为困难.

表 3　新词实例

Table 3　Examples of neologism

新词	原词
神马	什么
顶	支持
囧	难堪
东东	东西

解决数据稀疏的问题已成为统计机器翻译的重中之重.统计机器翻译中数据稀疏问题主要表现在 4 个方面:1)训练集的数据稀疏导致的概率估计不准确;2)系统译文中的片段没有出现在开发集或测试集的参考译文中,影响了调参和自动评测的准确性;3)在待译语句中出现了训练集中没有出现的文字片段,对于这样的陌生文字片段,翻译系统无法处理;4)数据稀疏导致的一些预处理工作效果不佳,如分词、对齐等.

复述技术就是对一段文本片段生成意义相同的不同字面表述,可以丰富语言表达.因此召回率的提高恰好可以解决语言的个体性差异与进化性多变所带来的数据稀疏问题.

3　总结与展望

本文对引入复述的统计机器翻译研究的进展进行了综述.重点介绍了复述在统计机器翻译应用研究中的几个关键问题,包括复述改善模型训练、复述提高调参效果、复述改写待译语句和复述改善机器翻译自动评测.虽然对复述和机器翻译问题的探讨由来已久,但将复述与统计机器翻译相结合进行广泛的研究却不足 10 年.所以还存在许多值得深入探索的问题,在此提出一些值得进一步挖掘的研究方向,希望对本领域的研究有所启发.

1)尽管人们已经提出了多种方法用于获取复述句、复述短语和复述模板等资源.然而,获取的资源精确度还较低,含有的噪声太多.因此,如何找到一种有效的方法,对获取的复述资源进行过滤,并且有效地应用到统计机器翻译中,这是一个重要的研究课题.

2)虽然已有研究者将复述技术应用到统计机器翻译的不同阶段,但机器翻译和复述仍是 2 个独立的子集,没有将机器翻译与复述融合为一个模型,尤其是复述和语言模型结合的研究还不是很多[44].基于 MT 的复述生成模型和利用复述的 MT 模型,可将其整合,形成一个融合机器翻译与复述的联合模型,这样的好处是提高容错性,使翻译系统更好地运用复述知识.

3)从已有研究工作中可知,复述对于统计机器翻译的应用大部分是解决数据稀疏问题,而解决稀疏性还有很多其他方法,如把复述(同义表达)一般化为相关表达(如上位表达),就可得到更通用的模型.举个例子,把"获得性免疫缺损综合症"和"艾滋病"联系起来是复述,还可进一步泛化为"疾病".这点笔者将另文著述.

参考文献:

[1]BROWN P F, JOHN C, PIETRA S A D, et al. A statistical approach to machine translation[J]. Computational Linguistics, 1990, 16(2): 79-85.

[2]BROWN P F, PIETRA V J D, PIETRA S A D, et al. The mathematics of statistical machine translation: parameter estimation[J]. Computational Linguistics, 1993, 19(2): 263-311.

[3]OCH F J, NEY H. Discriminative training and maximum entropy models for statistical machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 295-302.

[4]刘挺,李维刚,张宇,等. 复述技术研究综述[J].中文信

息学报, 2006, 20(4): 25-32.
LIU Ting, LI Weigang, ZHANG Yu, et al. A survey on paraphrasing technology[J]. Journal of Chinese Information Processing, 2006, 20(4): 25-32.

[5]赵世奇,刘挺,李生.复述技术研究[J].软件学报, 2009, 20(8): 2124-2137.
ZHAO Shiqi, LIU Ting, LI Sheng. Rearch on paraphrasing technology[J]. Journal of Software, 2009, 20(8): 2124-2137.

[6]BEAUGRANDE D, ALAIN R, DRESSLER W. Introduction to text linguistics[M]. New York: Longman, 1981: 54-56.

[7]HALLIDAY M A K. An introduction to functional grammar [M]. London: Edward Arnold, 1985: 225-250.

[8]BARZILAY R, MCKEOWN K R. Extracting paraphrases from a parallel corpus[C]//Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. Toulouse, France, 2001: 50-57.

[9]BARZILAY R, ELHADAD N. Sentence alignment for monolingual comparable corpora[C]//Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Sapporo, Japan, 2003: 25-32.

[10]GLICKMAN O, DAGAN I. Identifying lexical paraphrases from a single corpus: a case study for verbs[C]//Proceedings of the International Conference on Natural Language Processing. Borovets, Bulgaria, 2003: 1-8.

[11]QUIRK C, BROCKETT C, DOLAN W. Monolingual machine translation for paraphrase generation[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain, 2004: 142-149.

[12]FINCH A, WATANABE T, AKIBA Y, et al. Paraphrasing as machine translation[J]. Journal of Natural Language Processing, 2004, 11(5): 87-111.

[13]ZHAO Shiqi, NIU Cheng, ZHOU Ming, et al. Combining multiple resources to improve SMT-based paraphrasing model[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Columbus, USA, 2008: 1021-1029.

[14]BOND F, ERIC N, APPLING D S, et al. Improving statistical machine translation by paraphrasing the training data [C]//Proceedings of the International Workshop on Spoken Language Translation. Waikiki, USA, 2008: 150-157.

[15]NAKOV P. Improved statistical machine translation using monolingual paraphrases[C]//Proceedings of the 18th Biennial European Conference on Artificial Intelligence. Patras, Greece, 2008: 338-342.

[16]KUHN R, CHEN Boxing, FOSTER G, et al. Phrase clustering for smoothing TM probabilities—or, how to extract paraphrases from phrase tables[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China, 2010: 608-616.

[17]MAX A. Example-based paraphrasing for improved phrase-based statistical machine translation[C]//Proceedings of the 2010 Conference in Empirical Methods in Natural Language Processing. Cambridge, USA, 2010: 656-666.

[18]OCH F J. Minimum error rate training for statistical machine translation[C]//Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan, 2003: 160-167.

[19]PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA, 2002: 311-318.

[20]MADNANI N, AYAN N F, RESNIK P, et al. Using paraphrases for parameter tuning in statistical machine translation[C]//Proceedings of the Second Workshop on Statistical Machine Translation. Prague, The Czech Republic, 2007: 120-127.

[21]CHIANG D. A hierarchical phrase-based model for statistical machine translation[C]//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Sydney, Australia, 2005: 263-270.

[22]MADNANI N, RESNIK P, DORR B J, et al. Are multiple reference translations necessary? Investigating the value of paraphrased reference translations in parameter optimization[C]//Proceedings of the 8th Conference of the Association for Machine Translation in the Americas. Waikiki, USA, 2008: 993-1000.

[23]MADNANI N, DORR B J. Generating targeted paraphrases for improved translation[J]. ACM Transactions on Intelligent Systems and Technology, 2013, 4(3): 1-26.

[24]MITAMURA T, NYBERG E. Automatic rewriting for controlled language translation[C]//Proceedings of the NLPRS 2002 Workshop on Automatic Paraphrasing: Theories and Applications. Tokyo, Japan, 2001: 1-12.

[25]YAMAMOTO K. Machine translation by interaction between paraphraser and transfer[C]//Proceedings of the 19th International Conference on Computational Linguistics. Taipei, China, 2002: 1107-1113.

[26]ZHANG Yujie, YAMAMOTO K. Paraphrasing of Chinese utterances[C]//Proceedings of the 19th International Conference on Computational Linguistics. Taipei, China, 2002: 1163-1169.

[27]SHIMOHATA M, SUMITA E, MATSUMOTO Y. Building a paraphrase corpus for speech translation[C]//Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal, 2004: 1407-1410.

[28]BURCH C C, KOEHN P, OSBORNE M. Improved statistical machine translation using paraphrases[C]//Proceedings of the Human Language Technology Conference of the

NAACL. New York, USA, 2006: 17-24.

[29] MARTON Y, BURCH C C, RESNIK P. Improved statistical machine translation using monolingually-derived paraphrases[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, 2009: 381-390.

[30] MIRKIN S, SPECIA L, CANCEDDA N, et al. Source-language entailment modeling for translation unknown terms [C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Singapore, 2009: 791-799.

[31] ONISHI T, UTIYAMA M, SUMITA E. Paraphrase lattice for statistical machine translation[C]//Proceedings of the ACL 2010 Conference Short Papers. Uppsala, Sweden, 2010: 1-5.

[32] ONISHI T, UTIYAMA M, SUMITA E. Paraphrase lattice for statistical machine translation[J]. EICE Transactions on Information and Systems, 2011, E94-D(6): 1299-1305.

[33] DU Jinhua, JIANG Jie, WAY A. Facilitating translation using source language paraphrase lattices[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, USA, 2010: 420-429.

[34] HE Wei, WU Hua, WANG Haifeng, et al. Improve SMT quality with automatically extracted paraphrase rules[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea, 2012: 979-987.

[35] RESNIK P, BUZEK O, HU Chang, et al. Improving translation via targeted paraphrasing[C]//Proceedings of 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, USA, 2010: 127-137.

[36] KAUCHAK D, BARZILAY R. Paraphrasing for automatic evaluation [C]//Proceedings of the Human Language Technology Conference of the NAACL. New York, USA, 2006: 455-462.

[37] KANAYAMA H. Paraphrasing rules for automatic evaluation of translation into Japanese[C]//Proceedings of the Second International Workshop on Paraphrasing. Sapporo, Japan, 2003, 16: 88-93.

[38] LEPAGE Y, DENOUAL E. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation[C]//Proceedings of the 2nd International Joint Conference on Natural Language Processing. Jeju Island, Korea, 2005: 57-64.

[39] ZHOU Liang, LIN Chinyew, HOVY E. Re-evaluating machine translation results with paraphrase support[C]//Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia, 2006: 77-84.

[40] LASSNER G R, LIN J, RESNIK P. A paraphrase-based approach to machine translation evaluation, Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57 [R]. College Park, USA: University of Maryland, 2005.

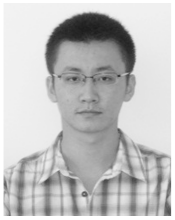
[41] SNOVER M, MADNANI N, DORR B J, et al. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric[C]//Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics. Athens, Greece, 2009: 259-268.

[42] SNOVER M, DORR B J, SCHWARTZ R, et al. A study of translation edit rate with targeted human annotation [C]//Proceedings of Association for Machine Translation in the Americas. Cambridge, USA, 2006: 223-231.

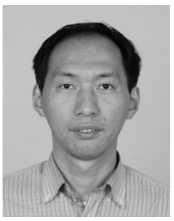
[43] PADO S, GALLEY M, JURAFSKY D, et al. Textual entailment features for machine translation evaluation[C]//Proceedings of the 4th Workshop on Statistical Machine Translation. Stroudsburg, USA, 2009: 37-41.

[44] LIU X, GALES M J F, WOODLAND P C. Paraphrastic language models[C]//Proceedings of 13th Annual Conference of the International Speech Communication Association. Portland, USA, 2012: 1-4.

作者简介:



胡金铭,男,1987 年生,硕士研究生,主要研究方向为自然语言处理、机器翻译.



史晓东,男,1966 年生,教授,博士生导师,主要研究方向为自然语言处理、机器翻译.先后主持和参与国家自然科学基金项目 3 项、国家“863”计划项目 10 余项,获福建省科技进步三等奖 1 项,发表学术论文 30 余篇.

苏劲松,男,1982 年生,讲师,博士,主要研究方向为自然语言处理、机器翻译等.