中文全词消歧在机器翻译系统中的性能评测

王 博 [1] 杨沐昀 [1] 李 生 [1] 赵铁军 [1]

摘 要 独立的词义消歧模型性能已经获得很大提高, 但是对于独立消歧模型在机器翻译系统中应用的必要性和作用一直存在着不同的观点. 为了从更为一般性的角度评价这个问题, 本文突破了具体模型的限制, 通过在不同类型汉英机器翻译系统中引入不受特定条件约束的高精度全词消歧过程, 对词义消歧在机器翻译系统中的影响进行了较为充分和全面的评价. 实验结果证明词义消歧模型不仅本身具有一定的翻译能力, 而且可以提高不同类型的机器翻译系统的整体性能. 同时也说明当前的翻译系统在消歧能力上还有较大的提升空间.

关键词 词义消歧, 机器翻译, 全词
中图分类号 TP391

Evaluation of All-words WSD for Chinese in Machine Translation

WANG Bo[1] YANG Mu-Yun[1] LI Sheng[1] ZHAO Tie-Jun[1]

Abstract Although remarkable improvements have been seen in the independent word sense disambiguation (WSD) models, there are still debates about the necessity to integrate the WSD models with the machine translation (MT) systems. To settle the question in a general view, we break the restrictions from specific models and a simulative perfect all-words WSD process is imported into MT systems of different types to acquire a sufficient and general evaluation. Experiment results indicate that a fine WSD process not only yields considerable translation quality itself but also obviously improves the MT systems. In addition, this work also reveals that current MT technologies still have much room to improve in selecting the best translation.

Key words Word sense disambiguation, machine translation (MT), all-words

词义消歧 (Word sense disambiguation, WSD) 的主要任务是在上下文中判断词汇的正确词义. 经过多年的发展, 尤其在 2000 年英语词义消歧系统开发, 测试和评价语料 —Senseval 的出现之后, WSD 的研究取得了快速发展, 目前无论是英语词义消歧还是起步较晚的汉语词义消歧都逐渐成为研究热点, 并已经形成一些较为成熟的模型, 其中既有基于规则的方法, 也有基于统计的方法. 在基于统计的方法中, 又可分为有指导的方法和无指导的方法, 各种方法在知识获取、精度、规模等方面各有优劣, 但是总体来讲词义消歧模型的性能已经达到了一定高度. 以汉语词义消歧为例, 鲁松在 2002 年使用无指导方法获得的平均消歧精度已达到 83.13 %[1], 而有指导的方法的精度还会更高.

另一方面, 词义消歧往往被视为一个中间任务, 最终要服务于诸如机器翻译 (Machine translation, MT)、信息检索等高级应用. 在独立的 WSD 模型的性能已经获得提高的情况下, 接下来受到关注的问题是 WSD 是否可以引入实际应用? 对这个问题的研究目前还处于起步阶段. 在各种自然语言处理 (Natural language process, NLP) 应用中, MT 作为 WSD 的一个主要应用领域, 是检验 WSD 在实际应用中的性能的一个重要对象. 对于 WSD 是否有助于提高 MT 系统的性能这一问题的研究尚不多见. 为了验证 WSD 对 MT 的影响, 在近期的工作中, Wu D K[2] 首先将 WSD 应用于统计机器翻译 (Statistical machine translation, SMT) 系统, 结果显示 WSD 的引入反而降低了 SMT 系统的整体性能, 随后 Cabezas[3] 在其工作中采取了不同的策略, 结果使得引入了 WSD 的 SMT 系统的性能获得了少许的提升. 二人的工作都对如何将 WSD 模型引入 MT 系统作了有益的尝试, 并且验证了当前 WSD 模型对 MT 系统可能产生的影响. 从他们的工作中可以发现, 虽然 WSD 和 MT 系统都已经具有较为成熟的方法, 但是二者的有效结合却仍存在着问题. 同时 Cabezas 的工作也表明, 即使在 WSD 模型及 MT 系统的性能基本不变的情况下, 仅对训练语料及结合方式进行改善, 也会使得 WSD 对 MT 的帮助有明显的提升. 这说明, 在具体模型、系统以及方法的约束下, 只能够对特定条件下的 WSD 对 MT 系统的影响作出阶段性的评测, 而难以从更宏观的角度对这一影响作出充分评价, 然而这种评价对于

收稿日期 2006-12-21 收修改稿日期 2007-08-01
Received December 21, 2006; in revised form August 1, 2007
国家自然科学基金 (60375019, 60773066) 资助
Supported by National Natural Science Foundation of China (60375019, 60773066)
1. 哈尔滨工业大学计算机科学与技术学院机器智能与翻译研究室 哈尔滨 150001
1. Machine Intelligence and Translation Laboratory, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001
DOI: 10.3724/SP.J.1004.2008.00535

回答是否应当将独立的 WSD 模型引入 MT 系统, 以及独立的 WSD 模型会对 MT 系统产生怎样的影响都是十分必要的.

为了实现非特定条件下的评价, 本文构造了一个面向机器翻译的消歧过程, 这是一个模拟的、高精度的和面向全词汇 (All-words) 的中文消歧过程. 有别于具体消歧模型, 模拟消歧过程不具备模型内部的复杂算法, 而从输入输出的角度对理想的消歧功能进行模拟, 从而实现了较高的精确率和召回率. 由于对于 MT 系统而言, 独立的 WSD 模型是一个黑箱过程, 所以模拟消歧过程不仅可以在评测中替代具体模型, 而且可以突破具体模型的限制, 进而实现更为充分的评价. 随后, 本文将模拟过程分别引入了不同类型的 MT 系统, 并对独立的模拟过程以及引入了该过程的 MT 系统的翻译性能进行了评测, 实验结果充分显示了在无约束的条件下, 独立的 WSD 模型的自身翻译能力及其对 MT 系统可能产生的影响.

本文的具体内容安排如下: 第 1 节介绍独立的词义消歧模型及机器翻译中的词义消歧的特点和现状; 第 2 节介绍构造高精度全词消歧过程的方法; 在第 3 节中将模拟过程分别引入了基于规则的机器翻译 (Rule-based machine translation, RBMT) 和 SMT 系统; 在第 4 节中通过一组实验对消歧过程的独立翻译能力及其对翻译系统的影响进行了评测和分析; 最后在第 5 节中对本文的工作进行了总结.

1 独立的 WSD 模型与 MT 中的 WSD

自然语言处理的最终目的是让机器像人一样理解并使用自然语言, 要达到这个目的有很多问题亟待解决, 如: 准确地在上下文中理解最小的语义单元—“词” 的含义就是其中之一, 这个过程被称为 “词义消歧”. 一直以来词义消歧的研究主要以独立于具体应用的消歧模型为主, 在这类研究中, 消歧过程通常被抽象为一个分类问题: 分类的对象是被称为 “词” 的单个词汇或词组构成的语义单元, 而分类的目标是一组根据语言学知识设定的词义标记. 给定一组与上下文相关的特征, 独立消歧模型对输入的词汇根据这些特征标注特定的词义标记. 这个过程可以形式化地描述为: 令词语 W 具有 N 个词义, W 在特定的上下文 C 中的正确词义是 S'. 词义消歧的任务就是在 N 个词义中确定 S'. 每个词义 S_i 和上下文 C 存在着强度不同的关系 $R(S_i|C)$, 而 S' 同 C 的关系应该是最强的. 消歧模型通过计算 C 和每个 S_i 之间的 R 来确定 s'. 整个过程可描述为: $S = \arg\max(S_i|C)$. 近年来, 国内外对词义消歧的研究, 尤其是对基于统计方法的消歧模型的研究发展迅速, 具有代表性的有: Yarowsky[4] 及 Mihalcea[5] 提出的单语语料增强方法, Li C 和 Li H 在此基础上提出的双语增强方法[6], 李涓子提出的基于最大熵模型的方法[7] 以及 Wu D K 提出的 Kernel PCA 方法[8] 等, 以上方法均为有指导的方法. 在无指导的方法中, 黄昌宁和李涓子依照《同义词词林》的语义分类体系在大规模语料库中自动获取分类器的方法[9] 及鲁松的向量空间模型方法也取得了较好的效果. 在消歧性能方面, 目前有指导的消歧方法在较小规模的歧义词集合上可以达到接近 90 % 的平均精度, 而无指导的方法也可以达到 80 % 左右的平均精度[10]. 虽然独立消歧模型的性能已经达到一定高度, 但是若作为实际系统的一部分词义消歧会有怎样的表现, 仍然有待考察. 而机器翻译系统目前正成为实践消歧模型的一个热点.

机器翻译中的词义消歧也被称为 “译文选择”, 因为这时一个源语言词的词义最终是由其在目标语中的译文来表示的, 而决定一个词的正确词义就是根据上下文在其若干可能的译文中进行选择. 因此, 机器翻译中的词义消歧任务具有一定的独特性: 首先, 在经典词义消歧中, 词义的表达需要一套约定的词义集合, 这样的词义集合往往以一部义类词典的形式给出, 如中文的 HowNet 和英文的 WordNet. 由于词义的定义是先验的语言学知识, 这就造成了词义表达体系存在着规模受限和标准不一致的问题, 从而给依赖于义类词典的消歧方法带来了困难. 与此不同, 由于在机器翻译中词义最终由目标语译文来表达, 这使得机器翻译中的词义消歧可以采取两种不同的策略. 第一种是首先依据一部义类词典决定词汇的词义, 再根据词义决定译文, 吴德凯在其工作中便采取了这种方法. 他首先对每个中文词汇针对其在 HowNet 中的 “Def” 进行消歧, 然后再针对选定的 “Def” 选择译文. 这种方法的不足是将义类词典的局限性引入了消歧过程. 另一种方法是直接使用译文作为消歧目标, Cabezas 在其工作中采用了这种方法, 他使用双语词对齐语料对消歧模型进行训练, 并直接对译文进行消歧. 其次, 经典的词义消歧的另一个瓶颈是标注语料的规模. 由于词义系统的特点和自动标注的准确率不高, 使得词义标注语料难以形成较大的规模. 在吴德凯的工作中由于需要对中间语义进行消歧, 所以使用了一个较小的训练集 (20 个中文歧义词, 平均每个歧义词约 37 个实例), 这在一定程度上影响了模型的训练质量. 但是在机器翻译中, 可以使用词对齐的平行双语语料作为训练数据, 而大规模的平行双语语料不难获得, 这使得构造更高精度的消歧模型成为可能. 在 Cabezas 的工作中, 正是利用了机器翻译中消歧过程的这两个特点, 在大规模词对齐的双语语料上直接针对译文训练消歧模型, 并将其应用到 SMT 系统当中, 最终提升了 SMT 系统的整体性能.

虽然 Cabezas 的方法获得了较好的结果, 但是如前言中所述, 在对当前消歧模型的使用中, 该方法给测试结论带来了一定的局限性. 一方面, 当前消歧模型在精确率上尚不十分理想; 另一方面, 在召回率上也存在着不足. 在吴德凯的评测中以一个较小的歧义词集合作为目标, 而 Cabezas 在西班牙语上的测试虽然是面向开放集合的, 但是在测试集中的词汇覆盖率也仅为 56 %. 同时, 作为目前机器翻译研究的热点, 上述的测试工作均选择了 SMT 系统作为研究对象, 但是若要实现较为全面的评价, RBMT 系统作为另一种类型的系统也应当加以考虑 (事实上, 目前较为成功的商用系统多数属于 RBMT 系统).

综上所述, 在特定的义类词典、语料规模、消歧模型及单一翻译系统的约束下, 虽然能够对现阶段 "词义消歧对 MT 系统的影响" 进行很好的评价, 但是难以对该问题在根本上作出一般性的回答. 因此, 本文尝试在突破特定条件约束的情况下, 从机器翻译中词义消歧的最终目的入手, 构造一个模拟的无约束消歧过程, 并将其应用到不同的翻译系统当中, 以期对在机器翻译系统中引入消歧过程的必要性和可能产生的影响作出更加充分的评价. 为了实现这一目的, 本文将在一个双语语料上通过准确的词对齐构造模拟的高精度全词消歧过程, 而后将模拟消歧过程分别以独立的和与翻译系统相结合的两种方式应用于该语料上的翻译任务上, 从而在翻译过程中实现无约束的消歧功能.

2　高精度的全词消歧及其与 MT 的结合

得到高精度全词消歧过程的关键在于获取准确词义. 在词义标注资源不足的情况下, 利用词汇级对齐的双语资源来获取词义信息是近年来出现的一种新方法. 这种方法利用词汇间的语义不对称性, 获取对齐词汇间的语义交集来实现词汇语义的定位. 这种方法的不足在于, 双语的语义不对称性是有限的. 例如, SENSEVAL-2 任务所有 29 个英语名词中, 有近四分之一的英语词汇与某些汉语词汇在全部义项上完全重合 (如 material (材料)), 对于这样的词汇, 就无法对语义进行准确定位. 但是, 正如第 1 节中所述, 由于机器翻译中的词义最终以目标语译文的形式来表达, 从而使得机器翻译中的词义消歧既可以是一个显式的语义消歧过程, 也可以是一个隐式的译文消歧过程. 而这两种过程的最终表现都是正确的目标语译文. 所以, 如果旨在构建一个以获取正确译文为目的的模拟消歧过程, 则可以不考虑不对称性缺失的问题, 因为对齐信息总是可以提供正确的译文. 在以往工作中, 词对齐语料仅用于进行训练, 而本文直接利用词对齐的映射来构造模拟的消歧过程, 然后将这个过程应用于面向该对齐语料的翻译任务. 这是一个类似于封闭测试的过程, 与封闭测试不同的是, 该过程在对齐词汇上能够保证 100 % 的准确率, 若进而能够使得对齐词汇在语料中具有较高的覆盖度, 便可以满足充分评价的要求. 这里, 我们假定源语言词汇的正确译文总是包含在作为其所在源语言句子的标准答案的目标语句子中, 并且源词汇与其译文之间的联系可以由词对齐给出. 一旦得到了这种联系, 就可以利用它来构造消歧过程. 这个过程可以形式化地表示为

$$\{cp_i\}_{i=1,\cdots,n} = CS,\ A(cp_i) = ep_j \tag{1}$$

$$Simulate(cp_i) = revise(ep_j) \tag{2}$$

$$TS = Translation(Simulate(cp_0) \\ Simulate(cp_1), \cdots, Simulate(cp_n)) \tag{3}$$

其中, CS 表示一个中文句子, cp_i 表示 CS 中的一个中文词 (组), A 表示 CS 及其对应英文句子上的一个对齐关系, ep_j 表示 cp_i 在 A 中的对应目标语词 (组), $Simulate(\cdot)$ 表示模拟消歧过程, $revise(\cdot)$ 表示对目标语词 (组) 的一个预处理过程. TS 表示 CS 的译文, 而 $Translation(\cdot)$ 表示与模拟过程相关的翻译过程. 下面我们从词对齐开始介绍构造模拟消歧过程的具体方法.

2.1　词对齐

如式 (1) 所示, 构造模拟消歧过程的第一步是构造对齐关系, 本文所需要的词对齐关系应当具有两个特点: 1) 对齐必须是准确的; 2) 对齐对所包含的词汇在句对中具有较高的覆盖度. 为了满足这两个特点, 本文在语料的选择和对齐方法上采取了相应措施. 用于对齐的语料为 NIST04 测试语料, 该语料包含了 1 788 个中文句子 (源语言) 及 4 个英文标准答案 (目标语). 所采用的词汇对齐方法是一个半自动化过程. 首先进行自动词对齐, 其目的是: 1) 降低人工校对的难度; 2) 在人工校对前对 4 组答案进行初步筛选. 所使用的自动对齐模型是由吕雅娟[11] 提出的统计与规则混合模型. 首先将中文句子分别与 4 个英文答案组成 4 组双语语料分别进行自动对齐. 为了提高对齐的覆盖率, 对于每个中文句子, 我们在其 4 个答案中选择包含对齐对最多的一句作为人工校对的基础. 接下来, 对筛选出的句对进行人工校对, 校对过程主要遵循以下两个原则:

1) 最小化原则: 如果一个词 (组) 可以被独立对齐, 那么它就不作为一个词组的一部分来被对齐. 即

$$(cp_i, ep_j) \in A \Rightarrow \forall (cp_i \in cp_j, ep_k), (cp_j, ep_k) \notin A \tag{4}$$

2) 词组的定义: 词对齐有可能存在一对多或多对多的关系, 这时对齐对的单侧或双侧是一个连续

或不连续的词组. 这里将词组定义为 3 种可能形式: a) 名实体; b) 固定搭配; c) 两个相邻停止词之间的一个连续片断, 且该片断的任何子串都不能够被独立地对齐.

以上原则用来保证对齐对中的词 (组) 是符合词义消歧中 "词" 的概念的较小的语言单位. 最终的对齐结果包含了 51 136 个中文单词, 其中 45 577 个单词是独立的或作为词组成员出现在对齐对中, 约占全部的 89.1 %. 虽然没有覆盖 100 % 的词汇, 但是基本满足了全词词义消歧的要求.

2.2 构造消歧过程

2.2.1 不受限的消歧过程

如式 (2) 所述, 在获得了对齐关系之后, 需要定义 $revise(\cdot)$. 在不受限的消歧过程中, $revise$ 仅被定义为取 ep_j 的原形 (如 did 的原形是 do, books 的原形是 book, $\cdots$), 而不作其他限制. 取原形是因为词汇在上下文中的正确变形不是 WSD 的功能, 所以需要进行还原. 特别地, 对于未被对齐的源语言词 (组), 将空串作为输出, 即 $Simulate(cp_i) =$ " ". 对于标点符号, 则使用一个对照表进行转换而不基于对齐对.

2.2.2 字典受限消歧过程

许多机器翻译系统需要依赖一个字典, 在这种系统中, 无论所使用的消歧模型多么完美, 都只能够在字典所提供的译文当中进行选择. 为了测试模拟消歧过程在这类系统中的作用, 我们假定不受限消歧过程的输出受到一部双语字典的约束, 从而得到一个字典受限的消歧过程. 在这个过程中不再直接使用不受限消歧过程的 $revise(\cdot)$ 的输出 (记为 $TransA$), 而是通过以下过程来重新定义受限过程的 $revise(\cdot)$ 的输出 (记为 $TransB$):

```
If (cp_i 包含在字典词条中)
    If (cp_i 的字典译文包含 TransA )
        TransB = TransA
    Else
        TransB = cp_i 在字典中的第一
                 个译文 (最常用译文)
Else
    TransB = 空串
```

我们可以用一个语料库中的实例来说明这两种消歧过程, 下面为一个语料库中的中英句对及其对齐对:

Education/1 is/2 the/3 cornerstone/4 of/5 modern/6 civilization/7./8
教育/1 是/2 现代/3 文明/4 的/5 基石/6 ./7
(1:1); (2:2); (4:6); (5:5); (6:3); (7:4);

对于汉语词汇 "基石", 使用不受限消歧过程选择其译文时, 输出为 "cornerstone", 而使用字典受限消歧过程选择其译文时, 发现字典中虽然存在 "基石" 这一词条, 但是其可选译文中不存在 "cornerstone" 这一译文, 所以这时将 "基石" 在字典中的第一个译文 "foundation stone" 作为输出.

2.3 将消歧过程引入 MT

下面介绍如何将模拟消歧过程引入不同类型的机器翻译系统. 这里我们选择了两种系统. 一种是 RBMT, 该类型的系统是目前商用系统中最为多见的类型; 另一种是 SMT, 该类型系统是目前性能最好的翻译系统, 也是研究的热点.

2.3.1 将消歧过程引入 RBMT

RBMT 系统一般分为两类: 基于转换的系统和基于中间语的系统. 这两种系统具有较大的相似性, 主要体现在都是依靠层次模型, 通过层间转换, 逐步完成由源语言到目标语的转化. 二者的主要区别在于, 在基于转换的系统中, 源语言与目标语的转化是在某一特定层次直接发生的 (如词汇层); 而在基于中间语的系统中, 这种转换是通过抽象的中间语言来实现的. 本文中, 我们选择了基于转换的 RBMT 系统作为消歧过程的引入对象, 因为: 1) 目前尚没有一个较为成熟的中间语及中间语系统可供选择; 2) 由于模拟消歧过程是建立在源语言与目标语词汇直接转换的基础上的, 这与基于转换的 RBMT 系统的方法相吻合. 如前所述, 一个典型的基于转换的 RBMT 系统具有清晰的层次结构, 每层独立地处理特定的语言现象, 各层之间不存在功能重叠. 换言之, 一个层次对于其前趋和后继层次而言是一个黑箱过程. 而词汇层作为一个特殊的层次, 往往又是两种语言接驳的地方. 它一般利用字典和规则集合为位于句法分析结果中的叶结点的每个词选择一个译文. 由于知识获取的瓶颈, 规则往往不能满足实际应用的需要, 从而使得系统在词汇层的消歧能力较弱. 但是若将本文中的模拟过程以黑箱的方式替代原有的词汇层, 便可替换系统原有的译文选择功能. 具体方法只需首先除去系统字典和消歧规则, 然后将句法分析结果中每个叶结点上的词作为模拟过程的输入, 而将其输出作为选定译文交至翻译过程的下一个步骤. 在式 (3) 的模型中, 引入了模拟过程的 RBMT 系统可形式化地表示为

$$\begin{aligned} TS = Translation(\cdot) = opLev_0(opLev_1(\cdots opLev_n \\ (Simulate(leaf_0), Simulate(leaf_1), \\ \cdots, Simulate(leaf_j)))) \end{aligned} \tag{5}$$

其中, $opLev_i$ 表示第 i 层的转换操作, $leaf_j$ 表示第 i 个叶结点.

2.3.2 将消歧过程引入 SMT

与 RBMT 不同, SMT 的翻译过程不具有层次结构, 以基于词组的 (Phrase-based) SMT 系统为例, 它利用统计方法在训练语料中获取不同长度词汇序列的翻译概率, 然后利用这些序列来组合译文, 使得组合后的译文的概率最大化. 在这个过程中, 系统知识的获取完全依赖于从训练语料中自动学习. 但是, 在实践中, 我们对一些特殊语言单元 (如名实体) 的翻译具有一定的先验知识, 于是 German (2003) 将这种先验知识引入到基于单词的 SMT 当中, 而 Koehn[12] 将其方法发展到了基于词组的 SMT 当中. 在 Koehn 的方法中一个子序列的译文可以在解码前以 XML 标注的方式嵌入到源语言句子中, 在解码过程中, 该译文被赋予最大概率 "1", 而句子的其他部分仍根据统计概率来计算, 这样先验知识便被融入到 SMT 当中. 在我们的研究中, 模拟消歧过程相对于 SMT 系统而言, 正是一种先验知识, 该过程对源语言中的特定序列 — 词汇给出了特定译文. 因此我们同样可以借助于 XML 标注的方法, 将模拟过程提供的知识引入 SMT 的翻译过程当中. 具体方法是, 对于源语言句子中的词汇, 通过 XML 标注的方法将其译文指定为该词汇在模拟过程中的输出译文, 然后再在标注后的语料上进行解码. 例如:

源语言句子: 教育是现代文明的基石.
标注后的句子:

```
<n english=
"education"> 教育 </n>
<n english= "be"> 是</n> <n english=
"modern">现代</n> <n english =
" civilization "> 文明</n> <n
english="of"> 的 </n>
<n english="cornerstone">基石</n>.
```

3 评测及分析

本节分别对理想的模拟消歧过程的独立翻译能力及其对 MT 系统的影响进行评测, 如第 2 节所述, 用于测试语料和用于构造消歧过程的语料相同, 均为 NIST04 的测试语料.

3.1 消歧过程的独立翻译测试

对消歧过程独立翻译能力的测试旨在评价消歧过程在不借助其他功能的情况下的独立翻译能力. 一般来讲, 从词汇的角度看, 机器翻译的任务主要包括 3 部分: 1) 选择正确的译文 (可为空); 2) 增加必要的词汇; 3) 为译文选择正确的排列顺序. 在这 3 部分当中, 词汇消歧所能够完成的仅包含第 1 部分. 所以, 若要对消歧过程的独立翻译能力进行评测, 就需要利用该过程根据词汇义项和上下文确定译文, 而屏蔽另两个任务带来的影响. 以此为依据, 本文设计了以下的实验:

实验 1 首先测试不受限消歧过程的独立翻译能力. 测试中使用第 2.2 节中介绍的不受限消歧过程获得源语言句子中每个词 (组) 的译文, 然后将译文按照源语言词 (组) 的顺序排列形成目标语句子, 而不进行其他处理, 如句法转换. 特别的, 如果源语言词组是不连续的 (如 "在 $\cdots$ 里"), 那么该译文将被放在词组的第一个单词在源语言句子中的位置上. 例如:

源语言句子: $cw1\ cw2\ cw3\ cw4, cw5\ cw6\ cw7$.
对齐对: $(cw1, ew2), (cw2, ew1\ \ ew5), (cw4\ \ cw7, ew6), (cw5, ew3\ ew4), (cw6\ cw7, ew8\ ew10)$.
目标语译文: $ew2\ ew1\ ew5\ ew6, ew3\ ew4\ ew8\ ew10$.

实验 2 测试字典受限消歧过程的独立翻译能力. 实验 2 的方法与实验 1 类似, 所不同的是所采用的消歧过程是第 2.2 节中介绍的字典受限消歧过程. 这里所使用的字典取自后续实验将使用的 RBMT 系统. 该字典包含 88 373 个中文词条及其所对应的若干英文译文.

若仍以第 2.2 节中的句对为例, 汉语句子 "教育是现代文明的基石." 在实验 1 中的翻译结果应为 "education be modern civilization of cornerstone.", 而在实验 2 中的结果将变为 "education be modern civilization of foundation stone.".

3.2 消歧过程对 MT 系统的影响

3.2.1 对 RBMT的影响

实验 3 和实验 4 分别测试不受限消歧过程和字典受限消歧过程对 RBMT 系统的影响. 选择的系统是本实验室开发的一个基于转换的 RBMT 系统 — MTS2003[13], 该系统参加了 2004 年的 NIST 评测. 在该系统的字典中, 测试语料中出现的词平均有 2.53 个译文, 最多的一个有 26 个译文. 这里依照第 2.3 节中介绍的方法, 将两种不同的模拟消歧过程引入到 MTS2003, 并使用该系统对测试语料进行翻译.

3.2.2 对 SMT 的影响

实验 5 测试不受限消歧过程对 SMT 系统的影响. 选用的系统是由 Koehn 提出的 Pharaoh[12]. 翻译模型的训练语料为筛选过的 Linguistic data consortium (LDC) 双语语料. 原始的 LDC 语料曾被选作 2004 年 NIST 评测的训练语料. 筛选的过程是首先找出 1 788 句测试语料中所包含的歧义词 (歧义词定义为该词在 HowNet 中对应了至少两个不同的 Definition[14]), 然后在 LDC 语料中选出至少包含

了一个上述歧义词的句子, 最后获得的训练语料包含了 840 229 个中英句对. 语言模型的训练语料为上述筛选后的语料的英文部分. 最后依照第 2.2 节所述方法将不受限的模拟消歧过程引入到 Pharaoh, 并进行测试. 同样以第 2.2 节中的句子为例, 汉语句子 "教育是现代文明的基石." 在实验 3 和实验 4 中的翻译结果分别为 "The education is cornerstone of modern civilization." 和 "The education is foundation stone of modern civilization.". 而该句在实验 5 中的翻译结果为 "education be cornerstone of modern civilization.". 这里将原始的 MTS2003 系统在测试语料 (NIST04 1 788 句双语句对) 上的翻译结果作为 baseline1, 将 SMT 模型在未标注过的原始语料上的翻译结果作为 baseline2. 表 1 给出了各实验结果的 BLEU-4 得分. 作为参考, 在 2004 年 NIST 机器翻译评测中, 得分位居前列的若干系统在相同测试集上取得的较好及最佳得分约在 0.2 ∼ 0.3 左右 (由于 NIST 评测未公布参赛系统细节, 故此处未能作详细说明).

表 1 Baseline 及各实验中翻译结果的 BLEU-4 得分

Table 1 BLEU-4 scores of translations in experiments and baseline

翻译结果	BLEU-4 得分
原始 RBMT (baseline1)	0.11
不受限 WSD (实验 1)	0.25
字典受限 WSD (实验 2)	0.09
RBMT + 不受限 WSD (实验 3)	0.18
RBMT + 字典受限 WSD (实验 4)	0.12
SMT 在原始语料上 (baseline2)	0.1966
SMT 在 XML 标注后的语料上 (实验 5)	0.2097

3.3 分析与讨论

首先应当注意的是, 本文所构造的消歧过程是一个模拟过程, 用来揭示理想的消歧过程的独立翻译能力及其对 MT 系统的影响, 并不代表目前的消歧模型所能达到的性能.

实验 1 的结果获得了最高的 BLEU 得分, 甚至与 NIST 评测的最佳得分相当 (0.2 ∼ 0.3). 获得该分数的参赛系统多为基于词组的统计机器翻译系统, 总体来讲, 与实验 1 中所使用的独立的不受限模拟消歧过程相比, 参赛系统在词序安排上具有优势, 但在消歧能力上劣势明显. 而词序与正确的译文在 N-gram 匹配中都具有重要的影响. 不受限模拟过程通过精准的译文选择弥补了词序调整的不足, 使得其性能可与完整的 SMT 系统相比拟. 这说明一个理想的消歧过程本身就具有较好的翻译能力, 若在此基础上进一步完善调序能力, 或在 SMT 系统中合理地引入独立的消歧能力, 都有望大大提高 SMT 系统的整体性能.

实验 2 的结果得分最低, 主要原因是词典的覆盖率. 本文实验中, 测试语料共包含了 8 936 个不同词汇, 其中 6 301 个词汇出现在字典中, 约占 70.5 %. 而对于 8 936 个词汇在语料中的 51 136 次出现, 其中有 41 037 次出现在字典中, 约占 80.3 %. 更进一步, 在这 41 037 次出现中, 有 32 766 次出现的对齐译文包含在其字典译文中, 约占 79.87 %, 所以最终的义项覆盖率仅为 64.1 % (80.3 % × 79.87 %). 以上分析说明字典的义项覆盖率是消歧过程的重要瓶颈.

实验 3 和实验 4 的得分均高于原始系统, 这说明消歧过程可以明显提升 RBMT 系统的性能. 类似地, 消歧过程的引入同样对 SMT 系统有了少许的改善, 虽然改善不十分明显 (BLEU 得分提高 6.7 %), 但是说明词义消歧过程可以对 SMT 系统性能起到正面作用. 以上分析均说明了 WSD 研究对于 MT 系统的重要性: MT 性能的改善不能够忽略消歧问题, 而独立的消歧过程是解决这一问题的有效途径.

值得注意的另一方面是, BLEU 评测是基于 N-gram 的, 它对译文与标准答案的匹配要求比较严格, 这使得即使译文是合理的, 但若其与标准答案的匹配度较低, 则得分也较低. 这可以从一方面解释为何实验 3 及实验 5 的得分会低于实验 1. 经分析, 对于嵌入 MT 系统中的消歧过程, 其输出会被 MT 系统进行诸如调序、加入虚词以及词性变换等操作, 这些操作在两种情况下会造成 BLEU 得分的降低: 1) 操作错误; 2) 操作正确, 并提高了结果的可理解度, 但却降低了与标准答案的 N-gram 匹配度.

为了说明上述情况, 这里设计了以下实验: 从测试语料中选择这样的句子, 其在实验 3 中的得分低于在实验 1 中的得分, 并且在实验 3 中至少被施以了调序的操作. 在这样的句子中, 随机选择 100 句, 要求 2 名实验者对这 100 个句子在实验 3 和实验 1 中的翻译结果进行打分: 按照可理解度从低到高, 分为 0, 1, 2, 3 四个分数. 结果显示, 有 59 个句子在实验 3 中的结果的人工打分高于或等于其在实验 1 中的结果的打分, 这说明了 BLEU 得分作为一种标准的评测方法虽然可以揭示翻译结果的总体质量, 但并不总是能够完全表征每个句子的可理解程度.

4 结论

本文在准确词对齐的基础上构造了一个应用于汉英翻译的模拟高精度全词词义消歧过程. 在目前词义消歧模型在规模和准确度方面尚不十分理想的情况下, 实现了对汉英翻译中高精确率和高召回率的全词消歧的性能评测. 实验结果表明一个理想的消歧过程不仅本身具有较好的独立翻译能力, 而且

还可以明显提高 RBMT 系统和改善 SMT 系统的性能. 这个结果说明了设计一个良好的独立词义消歧模型, 并以适当方法将其引入到 MT 系统中的必要性. 另一方面, 实验也显示将消歧过程引入 MT 系统仍存在着诸如字典规模瓶颈等有待解决的问题.

References

1 Lu Song, Bai Shuo, Huang Xiong. An unsupervised approach to word sense disambiguation based on sense-words in vector space model. *Journal of Software*, 2002, **13**(6): 1082−1089
(鲁松, 白硕, 黄雄. 基于向量空间模型中义项词语的无导词义消歧. 软件学报, 2002, **13**(6): 1082−1089)

2 Carpuat M, Wu D K. Word sense disambiguation vs. statistical machine translation. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. 387−394

3 Cabezas C, Resnik P. Using WSD Techniques for Lexical Selection in Statistical Machine Translation, Technical Report CS-TR-4736/LAMP-TR-124/UMIACS-TR-2005-42, College Park, University of Maryland, USA, 2005

4 Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, Massachusetts: Association for Computational Linguistics, 1995. 189−196

5 Mihalcea R. Bootstrapping large sense tagged corpora. In: Proceedings of the 3rd International Conference on Languages Resources and Evaluations. Canary Islands, Spain: 2002

6 Li H, Li C. Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, 2004, **30**(1): 1−22

7 Li Juan-Zi, Huang Chang-Ning. An improved maximum entropy language model and its application. *Journal of Software*, 1999, **10**(3): 257−263
(李涓子, 黄昌宁. 语言模型中一种改进的最大熵方法及其应用. 软件学报, 1999, **10**(3): 257−263)

8 Wu D K, Su W F, Carpuat M. A kernel PCA method for superior word sense disambiguation. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain: Association for Computational Linguistics, 2004. 637−644

9 Huang Chang-Ning, Li Juan-Zi. A language model for word sense disambiguation. *Applied Linguistics*, 2000, **3**: 85−90
(黄昌宁, 李涓子. 词义排歧的一种语言模型. 语言文字应用, 2000, **3**: 85−90)

10 Lu Zhi-Mao, Liu Ting, Li Sheng. The research progress of statistical word sense disambiguation. *Acta Electronica Sinica*, 2006, **34**(2): 333−343
(卢志茂, 刘挺, 李生. 统计词义消歧的研究进展. 电子学报, 2006, **34**(2): 333−343)

11 Lv Ya-Juan, Zhao Tie-Jun, Li Sheng, Yang Mu-Yun. English-Chinese word alignment based on statistic and lexicon. In: Proceedings of the 6th Joint Symposium of Computational Linguistics. Taiyuan, China: China Computer Federation, 2001. 108−115
(吕雅娟, 赵铁军, 李生, 杨沐昀. 统计和词典方法相结合的双语语料库词对齐. 第六届计算语言学联合学术会议. 太原, 中国: 中国计算机学会, 2001. 108−115)

12 Koehn P. A beam search decoder for phrase-based statistical machine translation models: user manual and description for version 1.2 [Online], available: http://www.isi.edu/licensed-sw/pharaoh/manual-v1.2.ps, July 16, 2004

13 Jiang Hong-Fei, Yang Mu-Yun, Zhao Tie-Jun. Olympics oriented RBMT vs EBMT. *Journal of Chinese Information Processing*, 2006, **20**(z1): 71−74
(蒋宏飞, 杨沐昀, 赵铁军. 面向奥运的汉英 RBMT 与 EBMT 研究. 中文信息学报, 2006, **20**(z1): 71−74)

14 Dong Zhen-Dong, Dong Qiang. Introduction to HowNet [Online], available: http://www.keenage.com/zhiwang/c_zhiwang.html, July, 2006

王 博 哈尔滨工业大学博士研究生. 主要研究方向为词义消歧, 机器翻译. 本文通信作者.
E-mail: bowang@mtlab.hit.edu.cn
(**WANG Bo** Ph. D. candidate in School of Computer Science and Technology at Harbin Institute of Technology. His research interest covers word sense disambiguation and machine translation. Corresponding author of this paper.)

杨沐昀 哈尔滨工业大学副教授. 主要研究方向为自然语言处理, 机器翻译.
E-mail: ymy@mtlab.hit.edu.cn
(**YANG Mu-Yun** Associate professor in School of Computer Science and Technology at Harbin Institute of Technology. His research interest covers natural language processing and machine translation.)

李 生 哈尔滨工业大学教授. 主要研究方向为自然语言处理, 机器翻译.
E-mail: shengli@mtlab.hit.edu.cn
(**LI Sheng** Professor in School of Computer Science and Technology at Harbin Institute of Technology. His research interest covers natural language processing and machine translation.)

赵铁军 哈尔滨工业大学教授. 主要研究方向为自然语言处理, 机器翻译.
E-mail: tjzhao@mtlab.hit.edu.cn
(**ZHAO Tie-Jun** Professor in School of Computer Science and Technology at Harbin Institute of Technology. His research interest covers natural language processing and machine translation.)