北京大学学报(自然科学版) 第 53 卷 第 2 期 2017 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, No. 2 (Mar. 2017)
doi: 10.13209/j.0479-8023.2017.035

融合词法句法分析联合模型的树到串EBMT方法

王丹丹 徐金安† 陈钰枫 张玉洁 杨晓晖

北京交通大学计算机与信息技术学院, 北京 100044; † 通信作者, E-mail: jaxu@bjtu.edu.cn

摘要 针对传统的基于实例的机器翻译(EBMT)方法中系统构筑复杂度和成本较高的问题, 提出一种基于依存树到串的汉英实例机器翻译方法。与传统方法相比, 该方法只需进行源语言端的句法结构分析, 可以大大降低构筑系统的复杂度, 有效降低成本。为了提高翻译精度, 引入中文分词、词性标注和依存句法分析联合模型, 可以减少汉英 EBMT 中源语言端基础任务中的错误传递, 提高提取层次间特征的准确性。在此基础上, 结合依存结构的特征和中英语料的特性, 对依存树到串模型进行规则抽取以及泛化处理。实验结果表明, 相对于基线系统, 该方法可以提高实例对抽取质量, 改善泛化规则和译文质量, 提高系统性能。

关键词 基于实例的机器翻译; 依存树到串模型; 联合模型; 泛化模板

中图分类号 TP391

A Tree-to-String EBMT Method by Integrating Joint Model of Chinese Segmentation and Dependency Parsing

WANG Dandan, XU Jin'an†, CHEN Yufeng, ZHANG Yujie, YANG Xiaohui

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044;
† Corresponding author, E-mail: jaxu@bjtu.edu.cn

Abstract In consideration of the complexity and high cost of system construction in traditional examplebased machine translation (EBMT) methods, the authors propose a Chinese-English tree-to-string EBMT method. Compared with the traditional methods, the preposed approach just needed to implement the processing of source language parsing. Word segmentation, POS tagging and dependency parsing were jointed to relieve the affections of error propagation and failure of feature extraction at different levels. Moreover, the authors extracted and generalized bilingual word and phase alignments from examples and templates by using the dependency structure of source language. Experimental results show that the preposed method can achieve better performance significantly than baseline systems.

Key words example-based machine translation; dependency tree-to-string model; joint model; generalization template

经过 50 多年的发展, 已形成多种机器翻译方法。20 世纪 90 年代以来, 逐渐形成三大主流方法: 基于规则的机器翻译、基于统计的机器翻译和基于实例的机器翻译。基于实例的机器翻译(example-based machine translation, EBMT)方法由日本学者长尾真(Makoto Nagao)于 1981 年提出, 并于 1984 年发表[1]。长尾真提出的基于实例的机器翻译方法借鉴了日本人初学英语时翻译句子的过程, 在翻译简单句子时, 不对句子进行深层次的分析, 而是将句子分解为几个片段, 借助已有的片段翻译结果, 对每个片段进行翻译, 最后将目标语片段进行重组得到句子翻译结果[2]。

EBMT 系统的研究主要围绕翻译模型(相似实例检索、对齐与译文生成)以及为其服务的双语对

国家自然科学基金(61370130, 61473294)、中央高校基本科研业务费专项资金(2014RC040)和科学技术部国际科技合作计划(K11F100010)资助
收稿日期: 2016–07–22; 修回日期: 2016–09–30; 网络出版日期: 2016–11–28

齐实例库的构建进行[3]。在 EMBT 研究的早期，往往不对原始双语语料做预处理，而是直接基于简单句子构成的实例库构建翻译模型。Somers 等[4]使用由简单句子构成的实例库，采用以字符为单位的最短编辑距离方法，以动态规划的算法来实现相似实例检索。在开源 EBMT 平台 Cunei[5]中，先根据词串匹配长度的阈值对实例进行筛选，再根据词性、词类等特征选择最相似的实例。这种以字符或词语为对象，采用编辑距离计算词串相似度的方法虽然易于实现，但是存在数据稀疏的问题。

为了解决数据稀疏的问题，许多学者采用泛化的双语实例构建翻译模型。泛化实例的概念最早由 Och 等[6]提出，并应用在机器翻译中。他们将双语中所有的词泛化为不同的词类，对词类进行对齐处理，再将词类翻译成词，得到最终的翻译结果。虽然实例泛化的方法解决了实例覆盖率低的问题，但是基于字符串表示的相似实例检索仍属于一种表层的关系，难以反映语言的内在规律，无法获得高质量的译文。

现阶段，EBMT 普遍采用句法结构树的形式存储双语实例，借助语言学知识指导机器翻译的过程[7−10]。基于结构化实例的 EBMT 多采用依存树到依存树的实例存储方式。日本学者佐藤研制的 MBT1[7]和 MBT2[8]系统是著名的基于实例的机器翻译系统。其中 MBT2 采用依存树到依存树的方式存储翻译实例。系统根据待翻译的源语言词汇依存树检索相似实例，利用检索到的实例碎片形成源匹配表达式。在译文生成阶段，将目标匹配表达式表示成目标语言词汇依存树的形式生成译文。虽然用这种方法可以得到高质量的译文，但需要基于数量级巨大的数据库。

为了在保证译文质量的前提下缩小数据库规模，日本京都大学黑桥实验室提出 KyotoEBMT 系统[9]。该系统在依存树到依存树的双语实例中引入泛化的方法，抽取所有子树实例对，并对部分叶节点进行泛化处理。当输入一个待翻译句子时，首先将其完整的依存树分解为子树的形式，并在实例库中搜索相似的实例片段，重组生成的译文树片段并生成译文。Vandeghinste 等[10]将实例表示为短语结构树的形式，并将短语结构树分解成包含名词短语的子数组，作为名词短语转化规则存储起来。在进行相似实例检索时，将名词短语作为标记，并对其余非名词短语的单词进行完全的匹配，从而选取最大匹配的结构作为译文。

虽然基于结构树到结构树对齐的实例包含丰富的句法结构信息，可以大大提高译文的质量，但是实现过程中有两大难点：1) 对于很多语言，并不存在质量较高的句法分析器；2) 对于不同语系的语言，句法结构存在较大的差异，难以建立精确的对应关系[11]。

为解决上述问题，Liu 等[12]提出一种基于 TSC (tree-string correspondence)的半结构化实例的英汉机器翻译方法，源语言采用短语结构树的形式存储，目标语言采用词串的形式进行存储。对于输入的待翻译句子，先对句子进行句法分析，再在实例库中寻找与其匹配的树。采用基于树间语义相似度、双语实例的词汇翻译概率以及目标语语言模型 3 个特征的统计生成模型并获得译文。

以上基于结构化实例的 EBMT 方法在翻译过程中，相似实例检索、实例对齐以及译文生成均需围绕结构化实例进行。相对于非结构化实例，结构化的实例虽然融合了更多的语法信息，可以得到更高质量的译文，但需要以准确的句法分析结果为前提。在结构化 EBMT 的研究中，研究者们多将重点放在如何提升相似实例的匹配度上，忽略了生成结构化实例的准确性。由于结构化实例的获得一般需要对原始语料做分词、词性标注和句法分析的预处理，传统管道式的方法极易造成错误的迭代传递，进而影响系统的翻译性能。因此，如何提高结构化实例的准确率，进而提升 EBMT 系统的翻译性能具有重要的研究价值。

1 系统架构

本文的 EBMT 系统主要由两大模块构成：依存树到串实例库构建模块和翻译模块，系统架构如图 1 所示。在进行翻译前，实例库构建模块先分别对汉英双语语料进行预处理及对齐处理，然后抽取并泛化得到依存树到串实例库。在翻译模块中，对于输入的中文句子，先采用中文词法句法分析联合模型将其表示为依存树的形式，再在泛化的依存树到串实例库中检索相似实例，得到一系列候选译文，最后根据对数线性模型的得分，对生成的候选译文进行排序，选取得分最高的 1-best 译文作为系统的最终翻译结果。

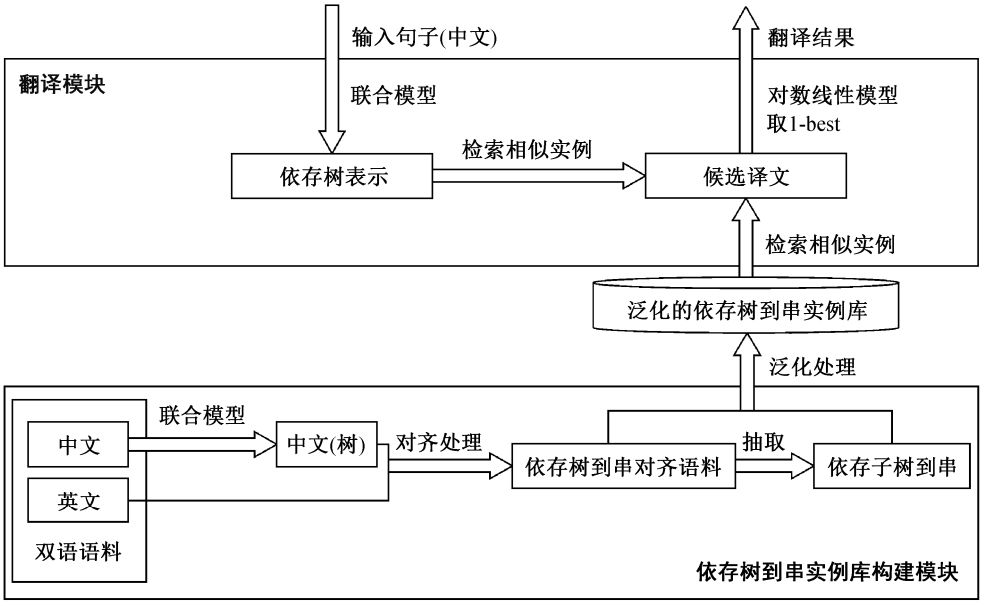

图 1　系统架构
Fig. 1　Framework of EBMT system

2　依存树到串泛化实例库构建

本文以汉英翻译为例, 采用结构化依存树到串的形式存储对齐的双语语料。在构建依存树到串汉英双语语料库时, 采用中文词法句法分析联合模型对汉语源语言进行预处理。

2.1　中文词法句法分析联合模型

词法分析、句法分析和语义分析是中文信息处理的主要任务。词法分析主要包括分词、词性标注和未登录词识别等子任务。分词、词性标注和句法分析是机器翻译的重要技术环节。传统机器翻译方法通常将分词、词性标注及句法分析看做一个管道, 进行分步骤、分层次的处理, 容易发生已有错误迭代传递的现象, 导致各个层次间的部分特征无法正确获取和利用, 影响翻译质量。

郭振等[13]利用词语内部结构, 将基于词语的依存句法树扩展为基于字符的依存句法树(如图 2 所示), 以解决不同任务间字串的粒度冲突问题。该模型结合了 N-gram 序列特征和依存子树特征, 对语料进行半监督训练, 获取半生语料, 通过其中蕴含的丰富信息提高联合模型的性能。

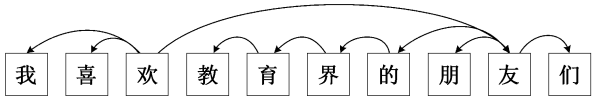

图 2　基于字符的依存句法树
Fig. 2　Dependency tree based on character

本文提出的 EMBT 方法引入郭振等[13]提出的联合分析方法, 旨在降低层次间的错误传递率, 提高翻译质量。一方面, 用于构建 EBMT 的依存树到串双语语料库过程中的各种词法、句法处理及其处理的环节, 以期提高 EBMT 实例库的质量; 另一方面, 在进行翻译处理过程中使用该联合模型, 对输入的汉语句子进行分词、词性标注和依存句法分析。

2.2　构建依存树到串泛化实例库

本文采用依存树到串的形式存储对齐的双语实例, 以汉英翻译为对象进行描述。汉英树到串对齐双语实例的构建过程如下。

1) 中文语料预处理。利用中文词法句法分析联合模型[13]对源语言中文语料进行分词、词性标注及句法分析。由于该联合模型是基于字符的, 因此本文在对中文语料预处理的过程中, 将原始语料中的每个句子按照字符进行分割, 即可得到以单词为单位、具有词性标注和词间依存关系的中文依存树实例。

2) 英文语料预处理。对英文语料进行小写化及词形还原处理, 在译文生成的最后一步, 相应地增加首字母大写处理的操作。

3) 中英文语料的词串对齐处理。由于单词对

齐的准确性直接影响到后续相似实例检索以及最终生成译文的质量, 因此在词对齐阶段要确保对齐的准确性。采用开源工具 GIZA++对汉英双语语料分别进行训练, 得到汉英和英汉的双向词汇对齐, 然后使用 Grow-Diag-Final 算法[14]对这两个方向分别求交集和并集, 并扩展做对称融合。

经过以上 3 个步骤, 得到句子级的汉英依存树到串对齐实例, 其结构举例如图 3(a)所示。

4) 抽取依存树到串实例(短语级)。考虑到实例库数据稀疏的问题, 本文在生成的依存树到串对齐的双语语料基础上进行片段化实例对的抽取, 并做进一步的泛化处理, 以减少由于数据稀疏无法找到合适实例, 进而出现翻译结果偏差过大的情况。

近年来, 短语翻译对的抽取技术大多基于词汇对齐的语料。在依赖词对齐的方法中, Och 等[15]提出的方法在机器翻译中应用最广泛。他们利用平行句对间的词对齐信息抽取短语翻译对, 首先枚举源语言句子中的所有短语片段, 根据词对齐关系寻找每个源语言实例中的单词在目标语实例中对应的最小位置和最大位置, 以此确定目标语区间。若目标语区间中的所有单词对应的源语言单词都在当前源语言实例中, 则抽取该短语翻译对。本文将 Och 等[15]的经典方法应用到本文短语级依存树到串实例对的抽取中, 抽取方法如下。

① 由于一个依存子树中的词汇能够表达一个完整的语义片段, 因此本文以依存树中的子树为抽取对象。在源语言依存树 D 中依次选取所有由连续单词组成的子树, 根据单词对齐关系 A 得到对应的目标语言单词集合。

② 根据目标语言词汇集合在原目标语词串中的最大位置 q 和最小位置 p 确定一个区间, 若区间 $[p, q]$内所有词汇对应的源语言词汇都在当前源语言依存树片段 $D[i, j]$中, 则抽取该短语级的依存树到串实例对。

例如, 对于实例对“建筑市场增大出口数量(construction market increased export volume)”, 将其进行依存树到串的对齐处理后, 结构如图 3(b)所示。为了方便描述实例对抽取的方法, 对每个单词编号。

从图 3(b)中词汇对齐关系, 可以得到所有由连续单词构成的源语言依存子树与其对应的目标语词串的对齐关系, 如表 1 所示。根据对齐关系, 表 1 中两对依存树到串实例对的英文单词对应的中文单词都在其相应的依存子树中。因此, 这两对依存树到串实例对是满足要求的, 将其抽取出来。

对于图 3(b)中的依存树到串实例对, 可以抽取出图 4 所示的两个实例对, 并将其加入到双语实例库中。

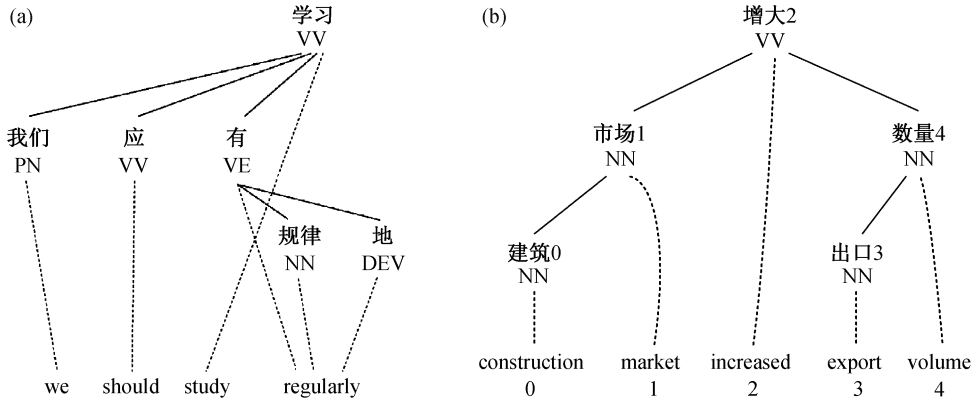

图 3 依存树到串对齐双语实例(句子级)

Fig. 3 Tree-to-string aligned bilingual example (sentence level)

表 1 依存子树到串对应关系

Table 1 Projection of dependency subtree-to-string

中文依存子树	英文词串	英文词串对应位置[p, q]
建筑(0) NN — 市场(1) NN	construction (0) market (1)	[0, 1]
出口(3) NN — 数量(4) NN	export (3) volume (4)	[3, 4]

5) 泛化依存树到串实例。过长的实例对不仅占用内存空间，而且在实际的机器翻译过程中很少使用。为了解决这个问题，本文对得到的依存树到串对齐实例库做进一步的泛化处理。

CMU 大学研发的 PANGLOSS 系统中的 EBMT 引擎[16]，通过识别句子实例中的命名实体，将实例库句子中的对应部分进行泛化。例如，对实例句子“John Hancock was in Philadelphia on July 4th.”进行命名实体识别，得到以下泛化结果:

〈PERSON〉was in〈PLACE〉on〈DATE〉。

本文借鉴文献[16]中泛化的思想，将已抽取出的短语级实例片段对视为对应 PANGLOSS 中命名实体的部分，作为泛化的变量。若某个句子级实例对中包含已抽取的短语级实例对，则将句子级实例对中包含的短语级实例对的部分用变量替换掉，可以有效地提高过长实例对的使用频率。例如，图 3(b)所示的句子级实例对中包含刚刚抽取出的两个短语级实例片段对。对其进行泛化处理，得到实例“XX 增大 XX” (XX 代表泛化的部分)，可提高检索率。将这两部分及其对应部分进行泛化处理，可以得到图 5 所示的泛化实例对。

得到泛化的依存树到串实例对后，将原来的句子级实例对替换为其对应的泛化实例对。那么，由图 3(b)中原始的句子级依存树到串实例对，可以得到图 4 和 5 所示的所有实例对。

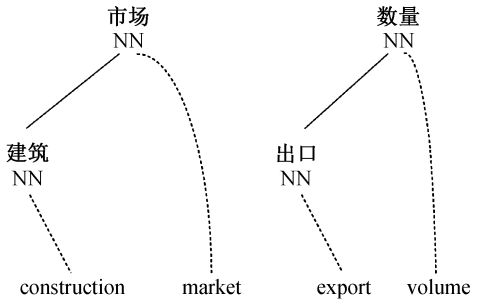

图 4　抽取的实例片段对(短语级)

Fig. 4　Extracted example fragment pairs (phrase level)

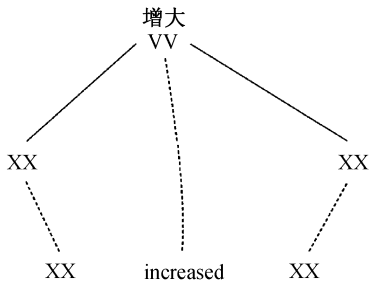

图 5　泛化的依存树到串实例对

Fig. 5　Generalized dependency tree-to-string aligned example

3　翻译模块

本文翻译模块主要分为相似实例检索子模块和译文生成子模块。

3.1　相似实例检索子模块

由于本文的双语对齐实例库以汉英依存树到串的形式存储，因此本文的相似实例检索算法针对中文依存树间的相似度进行计算。在基于依存树的方法中，Sato 等[17]提出的基于匹配表达式进行相似实例检索的方法是经典的方法，本文采用这种方法。该方法采用依存树到依存树的句子级英日双语实例，无需对实例进行片段抽取，而是将每个输入句子表示为一个或多个匹配表达式。每个匹配表达式表示在实例库中找到的某个依存子树的特定节点上进行的某种操作。匹配表达式有以下 3 种形式。

1) 替换: [r, 〈ID〉, 〈ME〉]，表示用〈ME〉节点替换〈ID〉节点。

2) 删除: [d, 〈ID〉]，表示删除〈ID〉节点。

3) 添加: [a, 〈ID〉, 〈ME〉]，表示将〈ID〉视为根节点，添加〈ME〉为其子节点。

利用这些操作，通过在数据库中找到的实例片段组合得到输入句子，即可完成相似实例检索。借鉴文献[17]中对依存结构的表示方法，将本文的依存树到串对齐实例中的源语言依存树进行形式化表示。下面为两个源语言依存实例 D1 和 D2。

D1: [c1, [购入, VV]
　　[c2, [他, PN]]
　　[c3, [杂志, NN]
　　[c4, [本, DT]
　　[c5, [一, CD]]]]]
%%　他购入一本杂志

D2: [c11, [读, VV],
　　[c12, [我, PN]]
　　[c13, [书, NN]
　　[c14, [一, CD]
　　[c15, [政治, NN]
　　[c16, [本, DT]
　　[c17, [有关, JJ]]
　　[c18, [的, DEG]]]]]]]
%%　我读一本有关政治的书

其中, 带有“c”前缀的标号为中文依存树中子树的标号, 如 c1, c5, c13 等。依存树中的每个节点包含两个属性: 词语及其对应词性。

在进行检索的过程中, 从输入句子依存树的根节点开始, 结合替换、删除和添加操作, 自上而下依次检索相同的实例片段, 直到将输入句子的所有节点完全覆盖为止。例如, 对于输入语句“他购入一本有关政治的书”, 经过中文词法句法分析联合模型处理后, 得到依存树的形式化表示如下。

[[购入, VV]
[[他, PN]]
[[书, NN]
[[一, CD]
[[政治, NN]
[[本, DT]
[[有关, JJ]],
[[的, DEG]]]]]]]

实例库中的源语言依存实例 D1 和 D2 分别包含该输入句子的一个片段, 将源语言依存树 D1 中的 c3 根节点子树用 D2 中的 c13 根节点子树替换, 即可拼接得到输入句子的结构。因此, 其对应的匹配表达式为[c1, [r, c3, [c13]]]。这是一个替换型的检索过程, 可以用依存树的树形结构来形象地表示(图 6)。

上述示例中的 D1 和 D2 源语言依存实例为短语级的依存树到串实例。对于泛化的依存树到串实例, 其中的泛化节点只有泛化变量一个属性, 用 XX 表示, 其他与短语级实例表示方法相同。下面为一个泛化实例源语言依存树的形式化表示。

D3: [c21, [购入, VV]
[c22, [我, PN]]
[c23, [XX]]]

对于以上两个检索示例, 其用词为常见词汇, 均可在实例库中检索到与每个节点完全相同的实例片段进行覆盖。然而, 在实际的检索过程中, 有时会出现某一子树根节点无法在实例库中检索到以该节点为根的源语言依存实例的情况。此时需要根据词语的语义相似度, 在实例库中搜索在语义上最为相似的源语言依存实例。通过计算两棵子树中对应节点语义相似度的总和, 选取与输入语句子树各节点语义相似度总和最高的实例片段作为相似实例片段。这里的语义相似度采用基于知网的词汇语义相似度计算方法[18]进行计算。

例如, 对于输入语句“江苏依托大运河建厂”, 其依存树形式化表示如下。

[[依托, VV]
[[江苏, NR]]
[[运河, NN]
[[大, JJ]]
[[建厂, VV]]]]

对于以节点“运河”为根节点的子树, 假定实例库中不存在以“运河”为根节点的源语言依存树实例, 则需用基于知网的词汇语义相似度计算方法, 检索与以“运河”为根节点的子树对应节点语义相似

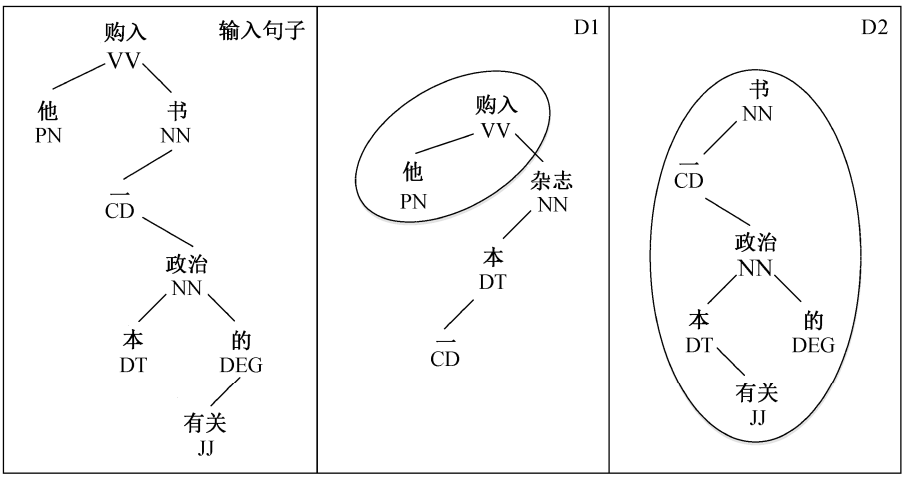

图 6 相似实例检索示例

Fig. 6 Alternative similar case retrieval

度总和最高的源语言依存树实例。对于上面的例子，检索到其语义最相似的源语言依存树实例如下。

```
[[河流, NN]
    [[大, JJ]]
    [[建厂, VV]]]
```

由于依存树中上部分的节点多为常用的主干性词汇(如动词等)，在实例库中存在的几率很大，因此利用语义进行相似度计算的方法多用在较下层且深度较浅的子树中，故基于语义进行计算的方法不会造成太大的误差。

对于每个输入的待翻译句子，进行相似实例检索后均可得到一些对应的匹配表达式，每个匹配表达式对应着一个完整的源语言句子。在译文生成阶段，利用双语实例间的对应关系，将这些源语言句子生成对应的译文，再从中选出最佳译文进行输出。

3.2　译文生成子模块

在相似实例检索模块得到某输入语句的大量匹配表达式后，需要结合其依存树到串的对应关系，得到对应的拼接而成的译文。然后，采用对数线性模型，对所有候选译文进行打分，选出得分最高的译文作为系统最终的输出结果。

采用 Sato 等[17]的方法，依据双语词对齐关系，将源语言匹配表达式转化为目标语匹配表达式，从而得到对应译文。在依存树到串实例中，同样借鉴 Sato 等[17]的方法，对目标语和对应关系进行形式化表示。其中，目标语部分不进行层次化表示，对应的译文匹配表达式的生成方法也做相应的改变。在实例库中完整的结构举例如下。

```
D3: [c21, [购入, VV],
          [c22, [我, PN]],
          [c23, [XX]]]
    %% 我购入 XX
S3: [e21, i]
    [e22, buy]
    [e23, XX]
    %% i buy XX
A3([c21, e22], [c22, e21], [c23, e23])
    %% c21↔e22, c22↔e21, c23↔e23
D5: [c51, [书, NN],
        [c52, [一, CD],
            [c53, [政治, NN],
                [c54, [本, DT]]
    %% 一本政治书
S5: [e51, a]
    [e52, politics]
    [e53, book]
    %% a politics book
    A5([c51, e53], [c52, e51], [c53, e52], [c54, e51])
        %% c51↔e53, c52↔e51, c53↔e52, c54↔e51
```

其中，对目标语词串 S 中的每个单词采用“e+目标语词串序号+单词序号”的形式进行标号，单词序号即为单词在目标语词串中的顺序号，如 e51 代表第 5 个目标语词串中的第一个单词，即“a”。其对应关系 A 为汉英实例间的词汇对齐关系，如 c51↔e53 等。

对于输入语句“我购入一本政治书”，结合源语言实例 D3 和 D5，检索得到的匹配表达式结果之一为[c21, [r, c23, [c51]]]。对于得到的源语言匹配表达式，结合词对齐关系，将源语言中的节点标号对应地替换为目标语节点标号。目标语译文如下。

```
I buy a politics book.
```

每个输入的待翻译语句，均可得到若干源语言匹配表达式。将其转换成对应的目标语匹配表达式，得到若干目标语译文。对于目标译文，采用对数线性模型进行打分，将得分最高的 1-best 译文作为最终的输出结果。在本文的对数线性模型中，使用了以下几个特征函数。

1) 相同单词数。

2) 翻译概率。指由实例的源语言短语翻译到目标短语的概率。实例的翻译概率越大，译文质量越高。翻译概率通过计算某特定短语级实例对数量在句子级实例对库中的比例来衡量，比例越大代表翻译概率越高。句子级实例对库中，某特定短语级实例对的数量用 count(e, f)表示，句子级实例对的数量用 count(e, f_i)表示。翻译概率的计算公式[19]如下:

$$\varphi(f \mid e)=\frac{\text{count}(e, f)}{\sum_{f_i}\text{count}(e, f_i)},$$

其中，e 表示某个特定的短语级实例对，f 表示包含短语级实例对 e 的句子级实例对数量，i 表示句子级实例对的个数，$\sum_{f_i}\text{count}(e, f_i)$ 表示所有句子级实例对的数量。

3) 目标语实例片段的平均长度。由平均长度越长的目标语实例片段组成的译文质量越高。本文根据目标语匹配表达式, 计算目标语实例片段的平均长度。

4) 语言模型。本研究采用目标语的五元语言模型。

4 实验与分析

采用 CWMT2015 官方评测用汉英新闻语料作为实验语料, 其中单语的语言模型使用双语语料中的英文语料进行训练, 不添加额外的资源。采用 CWMT2015 官方评测用测试集。为验证本文提出的 EBMT 方法的有效性, 我们基于同样的实验语料, 做了 4 组对比实验, 采用的翻译系统分别为: 日本京都大学黑桥实验室的 KyotoEBMT 系统、北京交通大学自然语言处理实验室研发的依存树到串统计机器翻译模型(Tree-to-string, SMT)、用 Stanford parser 构筑的基于依存树到串的 EBMT 方法(Stan-Tree-to-string)以及本文提出的 EBMT 方法。实验结果如表 2 所示。

从表 2 的实验结果可以看到, 在基于相同实验语料的情况下, 本文 EBMT 方法的系统性能略优于日本京都大学黑桥实验室的开源 KyotoEBMT 系统, BLEU5 值和 NIST 值分别高出 0.1 和 0.0017。同时, 本文的 EBMT 方法性能也优于既有的依存树到串统计机器翻译系统模型, BLEU5 值和 NIST 值分别高出 0.34 和 0.0083。从实验结果还可以看出, 本文提出的方法由于采用汉语分词与句法分析的联合模型, 性能明显优于未采用联合模型的方法。

表 2 翻译系统对比实验

Table 2 Contrast experiments of machine translation systems

系统	BLEU5	NIST
KyotoEBMT	24.31	5.6563
Stan-Tree-to-string	23.96	5.5873
Tree-to-string (SMT)	24.07	5.6497
本文方法	24.41	5.6580

KyotoEBMT 采用依存树到依存树的形式存储双语实例, 在进行相似实例检索时将源语言依存树划分为片段, 对每个片段寻找相似的目标语依存片段。然而, 中文和英文属于两个不同的语系, 在句法结构上存在较大差异, 因此在相似实例查找的过程中, 会在语义上出现一定的偏差, 从而影响最终输出译文的质量。本文方法采用依存树到串的形式存储双语实例, 并在抽取短语级实例对时, 完全依据中文源语言句子的句法结构对所对应的英文目标语译文进行选择, 可以在一定程度上保证译文语法的正确性。

从 4 个系统的译文结果中抽取部分译文进行分析, 表 3 列出测试集中的语句“建筑业对外开放呈现新格局”翻译结果的译文。

表 3 中 KyotoEBMT 与本文系统的译文的主要差别在于对“对外开放”一词的翻译。KyotoEBMT 系统的翻译结果为“The opening up outside of ...”, 与参考译文“The opening of ... to the outside”在含义上有较大差别。分析其原因, 推断是 KyotoEBMT 采用依存树到依存树的双语实例存储形式, 由于中文和英文在句法结构上存在较大的差异, 导致英文部分的对齐结果不一定符合英文的语法和语义, 因而得到不通顺的语句。

本文方法对于“对外开放”的翻译结果为“The opening to the outside of ...”, 虽然与参考译文在语序上有一些差别, 但是在语义上是一个符合语义的完整片段, 不会产生歧义。这种结果的出现得益于本文方法中双语实例的存储形式以及对应的短语级

表 3 不同翻译系统的译文对比

Table 3 Translation results comparison of different translation systems

项目	内容
原文	建筑对外开放呈现新格局
参考译文	The opening of construction industry to the outside present a new structure.
KyotoEBMT	The opening up outside of construction industry to present a new structure.
Stan-Tree-to-string	The opening of construction industry to the outside show a new pattern.
Tree-to-string (SMT)	Opening of construction industry to the outside show a new pattern.
本文方法	The opening to the outside of construction industry present a new structure.

表 4　本文方法不理想译文

Table 4　Unsatisfactory translation of the proposed method

项目	内容
原文	城建是外商投资新热点。
参考译文	Urban construction is a new hot spot of foreign business to invest.
KyotoEBMT	Urban construction is a new hot spot of foreign business investment.
Stan-Tree-to-string	Urban construction is a new hot spot of foreign investment.
Tree-to-string (SMT)	Urban construction is new hotspot of foreign investment.
本文方法	Urban construction is foreign business investment hotspot.

实例的抽取及相应的泛化方法。首先，本文采用依存树到串的形式存储不同语系的中英双语实例，不用过多地考虑双语实例在句法结构上的对应关系，只需在双语词对齐的基础上抽取符合条件的实例片段。这种方法的优势在于，在相似实例检索的过程中，每次检索到的片段都是符合语义的，即使最后生成的句子在语义片段的顺序上与参考译文有所出入，仍可以保证其语义上的通顺度。

另外，对依存树到串的统计机器翻译系统的翻译结果与未使用联合模型的依存树到串的 EBMT 系统的翻译结果进行人工比较，结果显示本文系统的译文在正确率和语义流畅度上整体表现较好。

然而，用本文方法得到的结果中也有一些不理想的译文。表 4 列出对于测试集中的语句“城建是外商投资新热点”在 KoyotoEBMT 和本文的 EBMT 系统的译文对比，主要差别在于对“新热点”一词的翻译。KyotoEBMT 的翻译结果与参考译文相同，而本文系统的翻译结果虽然有同样的含义，但由于没有译成一个短语结构，导致整句的流利度与通顺度受到影响。原因在于，本文在相似实例检索的模块采用依据依存树对应节点的相似度计算实例的相似度，虽然基于依存结构相对简单的依存树实例能够取得很好的实例检索效果，但仍有部分实例无法准确地检索到。KyotoEBMT 系统采用一种基于复杂树匹配的相似实例检索算法，利用依存结构的特点对相似实例进行详细分析，可以得到更高精度的相似实例检索结果，从而得到更流利的译文。因此，本文相似实例检索模块的算法有待完善。

由以上实验结果及分析可知，虽然本文 EBMT 系统在相似实例检索模块的算法有待完善，但从整体上看，本文提出的 EBMT 方法可以有效地获取正确率较高的译文。在基于相同语料的情况下，与 3 种基线系统相比，本文 EBMT 系统显示出更好的翻译性能。

5　结论

本文首次将中文词法句法分析联合模型融入基于实例的机器翻译系统中，以构建高质量的汉英依存树到串实例库，并完整实现一个基于实例的机器翻译系统。机器翻译系统性能的比较实验证明了本文方法的有效性，可以有效地获取正确率较高的译文。融入联合模型的对比实验结果表明，本文将中文词法句法分析联合模型融入实例机器翻译系统的方法可以有效提升译文质量，改善系统性能。但是，相似实例检索的算法有待完善。

参考文献

[1] Nagao M. A framework of a mechanical translation between Japanese and English by analogy principle. Artificial & Human Intelligence, 1984, 25(3): 351–354

[2] Dandapat S, Morrissey S, Way A, et al. Combining EBMT, SMT, TM and IR technologies for quality and scale // ESIRMT and HyTra. Avignon, 2012: 48–58

[3] Xuan H W, Li W, Tang G Y. An advanced review of Hybrid Machine Translation (HMT). Procedia Engineering, 2012, 29: 3017–3022

[4] Somers H, McLean I, Jones D. Experiments in multilingual example based generation // Proceedings of the 3rd Conference on the Cognitive Science of Natural Language Processing (CSNLP 1994). Tokyo, 1994: 149–164

[5] Phillips A B, Brown R D. Cunei machine translation platform: system description // Proceedings of the 3rd Workshop on Example-Based Machine Translation. Santiago, 2009: 29–36

[6] Och F J, Tillmann C, Ney H, et al. Improved

alignment models for statistical machine translation // Proceedings of EMNLP. Aachen, 1999: 20−28

[7] Sato S. MBT1: example-based word selection. Journal of Japanese Society for Artificial Intelligence, 1991, 6(4): 592−600

[8] Sato S. MBT2: a method for combining fragments of examples in example-based translation. Artificial Intelligence, 1995, 75(1): 31−49

[9] Nakazawa T, Kurohashi S. EBMT system of Kyoto team in PatentMT task at NTCIR-9 // Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-9). Tokyo, 2011: 657−660

[10] Vandeghinste V, Martens S. Top-down transfer in example-based MT // Proceedings of the 3rd International Workshop on Example-Based Machine Translation. Dublin, 2009: 69−76

[11] Al-Adhaileh M H, Kong T E, Yusoff Z. A synchronization structure of SSTC and its applications in machine translation // Proceedings of the 2002 COLING workshop on Machine translation in Asia — Volume 16. Taipei: Association for Computational Linguistics, 2002, 38(5): 1−8

[12] Liu Z, Wang H, Wu H. Example-based machine translation based on tree-string correspondence and statistical generation. Machine Translation, 2006, 20 (1): 25−41

[13] 郭振, 张玉洁, 苏晨, 等. 基于字符的中文分词、词性标注和依存句法分析联合模型. 中文信息学报, 2014, 28(6): 1−8

[14] Och F J, Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003, 29(1): 19−51

[15] Och F J, Ney H. The Alignment template approach to statistical machine translation. Computational Linguistics, 2004, 30(4): 417−449

[16] Brown R D. Example-based machine translation in the pangloss system // Proceedings of the 16th conference on Computational linguistics — Volume 1. Pennsylvania: Association for Computational Linguistics, 1996: 169−174

[17] Sato S, Nagao M. Toward memory-based translation // Proceedings of the 13th conference on Computational linguistics-Volume 3. Helsinki: Association for Computational Linguistics, 1990: 247−252

[18] 刘群, 李素建. 基于《知网》的词汇语义相似度计算. 中文计算语言学, 2002, 7(2): 59−76

[19] 殷乐. EBMT 中基于依存结构的翻译知识获取和翻译系统的实现[D]. 北京: 北京交通大学, 2014