软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(6):1472−1485 [doi: 10.3724/SP.J.1001.2012.04069]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

利用统计搭配模型改进基于实例的机器翻译*

刘占一[1,2+], 李 生[1], 刘 挺[1], 王海峰[2]

[1](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[2](百度公司,北京 100085)

Improving Example-Based Machine Translation with Statistical Collocation Model

LIU Zhan-Yi[1,2+], LI Sheng[1], LIU Ting[1], WANG Hai-Feng[2]

[1](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[2](Baidu.com, Inc, Beijing 100085, China)

+ Corresponding author: E-mail: zhanyiliu@gmail.com

Liu ZY, Li S, Liu T, Wang HF. Improving example-based machine translation with statistical collocation model. *Journal of Software*, 2012,23(6):1472−1485. http://www.jos.org.cn/1000-9825/4069.htm

Abstract: Example-Based machine translation (EBMT) uses a preprocessed bilingual corpus as a main translation knowledge. The final translation is generated by editing examples that match the input sentence. In the EBMT system, the performances of example selection and translation selection heavily influence the quality of the final translation. This paper proposes a method to improve the performance of the EBMT method by using statistical collocation model, which is estimated from monolingual corpora, in three aspects. First, the statistical collocation model is used to estimate the matching degree between the input sentence and examples to improve the performance of the example selection. Second, the performance of translation selection is improved by evaluating the collocation strength of the translation candidates and the context. Third, the collocated words of the translation candidates in the example are detected by the statistical collocation model and then the collocated words are corrected according to the context. In order to evaluate the proposed method, this study conducts a series of experiments. First, the study evaluates the proposed methods in a word-based EBMT system. As compared with the baseline, the methods achieves absolute improvements of 4.73~6.48 BLEU score on English-to-Chinese translation. Then, the study also applies the proposed translation selection method to a semi-structured EBMT system, and the translation qualities are further improved, with an improvement of 1.82 BLEU score. The results of human evaluation show that the translations generated by the improved semi-structured EBMT system can express the majority of the meaning of source sentences, and the fluency of theses translations can also be accepted.

Key words: statistical collocation model; example-based machine translation; example selection; translation selection

摘 要: 基于实例的机器翻译(example-based machine translation,简称 EBMT)使用预处理过的双语例句作为主要翻译资源,通过编辑与待翻译句子匹配的翻译实例来生成译文.在EBMT系统中,翻译实例选择及译文选择对系统性

* 基金项目: 核高基国家科技重大专项(2011ZX01042-001-001)
收稿时间: 2010-09-26; 定稿时间: 2011-05-25

能影响较大.提出利用统计搭配模型来增强 EBMT 系统中翻译实例选择及译文选择的能力,提高译文质量.首先,使用单语统计词对齐从单语语料中训练统计搭配模型.然后,利用该模型从 3 个方面提高 EBMT 的性能:(1) 利用统计搭配模型估计待翻译句子与翻译实例之间的匹配度,从而增强系统的翻译实例选择能力;(2) 通过引入候选译文与上下文之间搭配强度的估计来提高译文选择能力;(3) 使用统计搭配模型检测翻译实例中被替换词的搭配词,同时根据新的替换词及上下文对搭配词进行矫正,进一步提高EBMT系统的译文质量.为了验证所提出的方法,在基于词的EBMT系统上评价了英汉翻译的译文质量.与基线系统相比,所提出的方法使译文的BLEU得分提高了4.73~6.48个百分点.在半结构化的EBMT系统上进一步检验了基于统计搭配模型的译文选择方法,从实验结果来看,该方法使译文的 BLEU 得分提高了 1.82 个百分点.同时,人工评价结果显示,改进后的半结构化 EBMT 系统的译文能够表达原文的大部分信息,并且具有较高的流利度.

关键词: 统计搭配模型;基于实例的机器翻译;实例选择;译文选择

中图法分类号: TP18 **文献标识码**: A

1 引 言

基于实例的机器翻译(example-based machine translation,简称 EBMT)是一种自动翻译方法[1],该方法使用预处理过的双语例句作为主要翻译资源,通过编辑与待翻译句子匹配的翻译实例来生成最终译文.如果翻译实例库中存在与待翻译句子相似的翻译实例,那么 EBMT 方法就可以为待翻译句子生成高质量的译文.在某些特定领域的翻译中,例如口语翻译,EBMT 方法具有很强的实用性.

典型的 EBMT 模型通常包括 3 个翻译步骤:在翻译实例库中搜索与待翻译句子匹配的翻译实例;识别待翻译句子和翻译实例之间的差异词,并且为待翻译句子中的差异词构造候选译文片段,识别翻译实例中差异词的对应片段;重新组合翻译实例及候选译文片段从而得到最终翻译[2].在实际翻译的过程中,由于源语言和目标语言之间存在语义、句子结构等差异,EBMT 方法中通常存在如下几个问题.

1.1 翻译实例选择方面的问题

翻译实例选择是 EBMT 系统中的重要步骤,其性能直接关系到 EBMT 系统的译文质量,如下面的英汉翻译示例所示:

例(A).

待翻译句子:

Can I take a picture of the painting?

翻译实例(A.1):

Can I take a picture of the *car*?

我 能 为 这 辆 *汽车* 拍 张 照片 吗?

候选译文(A.1):

我 能 为 这 辆 *油画* 拍 张 照片 吗?

翻译实例(A.2):

Can *we* take a *photo* of the painting?

我们 能 为 这 幅 油画 拍 张 *照片* 吗?

候选译文(A.2):

我 能 为 这 幅 油画 拍 张 *相片* 吗?

在上例中,EBMT 系统为待翻译句子找到两个匹配的翻译实例,翻译实例(1)中仅有一个不一致的词“painting≠car”,而翻译实例(2)存在两个不一致的词“I≠we”和“picture≠photo”.由于翻译实例(A.1)具有更少的编辑次数,所以通常认为翻译实例(A.1)比翻译实例(A.2)更接近待翻译句子,从而优先被用来生成译文.但是,从生

成的两组候选译文来看,根据翻译实例(2)生成的译文要好于翻译实例(1)生成的译文.

在 EBMT 系统中,目前主要有基于词[3]和基于句法语义分析[4]两类方法来搜索匹配的翻译实例.两类方法各有优缺点,基于词的方法仅利用句子的表层信息,在判断句子整体结构相似方面有欠缺;基于句法语义分析的方法考虑了句子组成词汇的语义信息与整体框架结构信息.虽然该方法能部分地弥补基于词计算句子匹配度方法的不足,但是该方法不仅需要句法语义分析器,而且依赖分析器的性能.本文主要研究基于词的翻译实例选择算法.

在基于词的匹配方法中,通常采用编辑距离来评价翻译实例与待翻译句子之间的相似程度,编辑距离计算了从翻译实例转换到待翻译句子所需要的最少的插入、删除和替换的数目.翻译实例的编辑距离越小,那么该翻译实例就和待翻译句子越接近.由于纯粹的编辑距离仅考虑词汇是否一致,而忽略了差异词之间的匹配关系,这样,在 EBMT 系统中用来生成译文的例句可能不是匹配度最好的翻译实例.如例(A)中,翻译实例(A.1)和待翻译句子的编辑距离是 1,而翻译实例(A.2)和待翻译句子的编辑距离是 2,虽然翻译实例(A.2)和输入句子之间的匹配度更高,但是 EBMT 系统会采用翻译实例(A.1)来生成译文.为了解决这个问题,有些研究者提出在编辑距离的计算中使用同义词典来衡量不一致词之间的相似程度[5].然而,对于有些语言来说,可能缺少该类型的词典,所以该方法在实际应用中具有一定局限性.

1.2 译文选择方面的问题

待翻译句子与翻译实例进行比较以后,如果待翻译句子中的差异词具有多个候选译文,那么如何从中选择最恰当的译文?如下面的例子(B):

例(B).

待翻译句子:

That ship is anchored in the ***bank***.

翻译实例:

That ship is anchored in the ***dock***.

那 艘 船 停靠 在 *码头*.

候选译文(B.1):

那 艘 船 停靠 在 *银行*.

候选译文(B.2):

那 艘 船 停靠 在 *河岸*.

在例(B)中,待翻译句子和翻译实例之间不一致的单词是“bank≠dock”,“bank”的译文有“河岸、银行”等.EBMT 系统使用 bank 的译文来替换 dock 的译文,得到了两个最可能的候选译文(B.1)和(B.2),因为“河岸”和“银行”都是 bank 的常用译文,所以如果不借助更多的翻译知识,很难选择出合适的译文来.

为了选择正确的译文,早期人们常使用歧义消解方法,而此方法需要依靠语言学家手工编制的规则.手工编写规则费时费力,成为知识获取的一个瓶颈.后来,语言学家提供的各类词典成为人们获取词义消歧知识的一个重要来源[6−9].这些方法中使用的词典一般仅适用于通用领域,不能满足特定领域翻译的需要,而特定领域中存在大量的未登录词.

近年来,随着计算机存储容量和运算速度的飞速提高,通过使用各种机用资源和大规模语料库,计算机能够自动获得各种统计知识.因此,歧义消解研究中涌现出许多基于语料库的统计方法[10−13].如通常使用翻译概率和基于 *N*-gram 的语言模型概率来进行译文选择,但是在这些统计模型中不能很好地解决句子中词汇的长距离搭配问题,尤其是在机器翻译的译文生成过程中,基于 *N*-gram 的语言模型基本忽略了词汇的长距离搭配关系.

1.3 搭配词选择方面的问题

在根据待翻译句子编辑翻译实例的过程中,EBMT 系统如何保持最终译文中词语搭配的一致性,尤其是翻译实例中插入或替换的词和其他词之间的搭配一致性?如下面的例(C)所示.

例(C).

待翻译句子:

她们 正在 打 *篮球*.

翻译实例:

她们 正在 打 *毛衣*.

They are knitting a ***jumper***.

候选译文:

They are **knitting** a ***basketball***.

在上面的例(C)中,待翻译句子和翻译实例仅有一个差异词:“篮球≠毛衣”,由于这两个词本身并没有歧义,所以 EBMT 系统生成了译文“They are knitting a basketball”.然而在该译文中,“knitting”和“basketball”不是一个正确的搭配.为什么 EBMT 系统会生成这样的译文呢?比较待翻译句子和翻译实例我们发现,这两个句子里面各包含了一个搭配,即“打 篮球”和“打 毛衣”.这两个“打”字有不同的意义,他们在英语中对应不同的译文.到目前为止,在 EBMT 的方法中,还没有相关研究提出有效方法来解决该问题.

搭配通常指由两个或多个词组成的符合人们习惯的表达式.搭配是自然语言处理中一种重要资源,广泛应用于机器翻译和文本生成等研究中[14−16].针对上述 EBMT 中的问题,本文利用统计搭配模型来增强 EBMT 系统中翻译实例选择和候选译文选择的性能.首先,利用单语统计词对齐方法从单语语料中分别构造源语言和目标语言的统计搭配模型,然后,使用源语言统计搭配模型提高翻译实例选择的能力,使用目标语搭配模型提高译文选择能力.具体来讲就是:(1) 在编辑距离的基础上,使用源语言统计搭配模型来估计待翻译句子和翻译实例之间不一致词的相似度,同时利用词汇之间的搭配概率来考虑句子中词汇的翻译/替换风险(本文称其为“编辑风险”);(2) 使用目标语统计搭配模型,在译文选择过程中考虑词汇之间的搭配情况,尤其是不相邻单词之间的搭配情况;(3) 为了进一步提高译文质量,使用目标语统计搭配模型检测翻译实例中与被替换词构成较强搭配关系的搭配词,并且根据新替换的词及上下文,进行搭配词的矫正.

本文的贡献包含以下几点:

(1) 虽然搭配是机器翻译中的重要资源,但是前人的工作却很少涉及,尤其是对 EBMT 模型来说.本文提出在 EBMT 方法中全面融合统计搭配模型来提高其翻译实例选择、译文选择及搭配词选择的能力,并且从实验数据来看,取得了令人满意的结果.同时,本文中使用的统计搭配模型是通过基于双语语料的单语部分自动构造的,对已有 EBMT 系统来说,无需增加任何额外资源;

(2) 本文提出使用统计搭配模型来估计待翻译句子和翻译实例之间的匹配程度,并且首次提出利用词汇之间的搭配概率估计句子中词汇的编辑风险;

(3) 本文提出在 EBMT 系统中使用统计搭配模型来检查翻译实例中被替换词的搭配词,然后根据新的替换词及上下文对搭配词进行矫正.实验结果显示,译文质量得到了显著提高;

(4) 本文提出的方法具有普遍意义,从实验结果来看,不仅提高了基于词的 EBMT 系统的性能,同时在半结构化的 EBMT 系统中,通过在译文选择过程中引入统计搭配模型,也有效提高了译文质量;并且,对译文的人工评价结果显示,改进后的半结构化 EBMT 系统的译文能够表达原文的大部分信息,并且译文流利度也基本可以接受.

本文第 2 节介绍统计搭配模型.第 3 节描述使用统计搭配模型提高 EBMT 系统性能的方法.第 4 节介绍实验设置及实验结果.第 5 节给出结论.

2 统计搭配模型

搭配通常指由两个或多个词组成的符合人们习惯的表达式.根据不同的应用,搭配通常具有不同的侧重点.例如:从语言学家的观点来看,搭配是处在固定短语和可自由组合词语之间的一种语言现象;而统计语言学家则认为,搭配通常指非偶然地经常一起出现的词组[17].在本文中,搭配由习惯上一起使用的两个词组成,他们可以

是相邻的,也可以不相邻.这些搭配包括专有名词、习惯用语、连词和其他词的组合,例如动词+名词、形容词+名词、副词+动词、副词+形容词等.

本文使用了单语统计词对齐方法来构造统计搭配模型[18].首先抽取了单语句子中潜在的搭配词对,然后基于获取的搭配词对,估计了词对的搭配概率.该方法不需要额外的语言处理或资源,并且与传统的搭配抽取方法[19]相比,在相同的实验环境下,该方法获得的搭配具有更高的精确率和召回率.

2.1 单语统计词对齐模型

首先,对单语语料库中的每个句子进行复制,得到一个具有相同句子的句对,然后使用单语统计词对齐算法搜索句子中潜在的搭配词对.按照单语统计词对齐算法,给定一个单语句子 $S=w_1^l$,本文使用单语词对齐(monolingual word alignment model,简称 MWA3)算法得到其最优的对齐结果,如下所示:

$$A^*=\arg\max_A\{p_{\mathrm{MWA3}}(A\mid S)\} \tag{1}$$

$$p_{\mathrm{MWA3}}(A\mid S)\propto\prod_{i=1}^{l}n(\phi_i\mid w_i)\cdot\prod_{j=1}^{l}t(w_j\mid w_{a_j})\cdot d(j\mid a_j,l) \tag{2}$$

其中,A 代表词对齐序列,任何一个单词都不能和其本身对齐,因此,对齐集合表示为 $A=\{(i,a_i)|i\in[1,l]\ \&\ a_i\neq i\}$;$\phi_i$ 代表对应到 w_i 上的单词个数.公式(2)的模型中主要使用了 3 种概率模型:基于词的搭配概率模型 $t(w_j\mid w_{a_j})$、基于位置的搭配概率模型 $d(j|a_j,l)$和词的衍生度模型 $n(\phi_i|w_i)$.

图 1 显示了句子“团队 负责人 在 项目 进行 中 起 关键 作用.”的单语词对齐结果,其中,对齐的词对为潜在的搭配词对,如“关键 作用”、“起…作用”等.

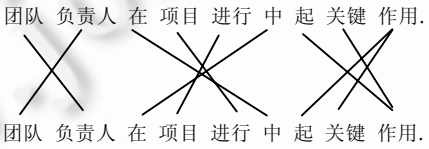


Fig.1 MWA example

图 1 单语统计词对齐示例

2.2 统计搭配模型

在经过单语统计词对齐处理后的语料中,计算对齐词对出现的频率.实验中,我们过滤掉了出现频率小于 2 的词对.基于对齐词对的频率,我们计算出了每个词对的对齐概率:

$$p(w_i\mid w_j)=\frac{freq(w_i,w_j)}{\sum_{w'}freq(w',w_j)} \tag{3}$$

$$p(w_j\mid w_i)=\frac{freq(w_i,w_j)}{\sum_{w'}freq(w_i,w')} \tag{4}$$

在本文中,搭配词对中的两个词是对等的,因此,词对的搭配概率采用了上述两个概率的平均值,见公式(5).

$$r(w_i,w_j)=\frac{p(w_i\mid w_j)+p(w_j\mid w_i)}{2} \tag{5}$$

公式(5)描述了两个词的搭配概率,可以看出:如果两个词之间的搭配概率较高,那么这两个词具有较强的搭配关系;反之,搭配关系较弱.

该方法有效描述了词汇之间的内在关系,统计搭配模型在其他 NLP 应用中取得了成功,如提高双语词对齐[20]、提高翻译系统的译文调序功能[21]等.

3 利用统计搭配模型提高 EBMT 性能

针对第 1 节中提到的 EBMT 系统中的 3 个问题,本节介绍了利用统计搭配模型提高 EBMT 性能的方法.首先,在利用编辑距离寻找匹配翻译实例的基础上,使用源语言统计搭配模型计算词语之间的匹配程度,并且估计句子中词汇的编辑风险;然后,在译文选择过程中,使用目标语统计搭配模型估计词汇之间的搭配情况;最后,为了进一步提高译文质量,使用目标语统计搭配模型检测翻译实例中与被替换词构成较强搭配关系的搭配词,同时根据上下文重新矫正了搭配词.

3.1 翻译实例选择算法

在目前的 EBMT 系统中,通常采用待翻译句子和翻译实例之间的编辑距离来计算他们之间的匹配情况,如

$$L(w_1^n,\tilde{w}_1^m)=\min\begin{Bmatrix}L(w_1^{n-1},\tilde{w}_1^m)+1\\L(w_1^n,\tilde{w}_1^{m-1})+1\\L(w_1^{n-1},\tilde{w}_1^{m-1})+\Delta(w_n,\tilde{w}_m)\end{Bmatrix} \tag{6}$$

其中, w_1^n 为待翻译句子, $\tilde{w}_1^m$ 为翻译实例中的源语言句子, $\Delta(w,\tilde{w})=\begin{cases}1, & w\neq\tilde{w}\\0, & w=\tilde{w}\end{cases}$.

如例(A)所示,虽然编辑距离小的翻译实例可以降低翻译例句的修改操作次数,但是该方法忽略了不一致词之间的匹配情况.为了进一步比较不同翻译实例与待翻译句子之间的匹配程度,本文提出借助于搭配概率来计算词语之间的语义相似距离.给定两个词(w_1,w_2),其相似距离计算为

$$dist(w_1,w_2)=1.0-\frac{F(w_1)\times F(w_2)}{\sqrt{F(w_1)\cdot F(w_2)}} \tag{7}$$

其中,$F(w)=\{w_i|r(w,w_i)>\delta\}$表示 w 的特征向量.在 w 的向量空间中,每个搭配词的权重就是该词和 w 的搭配概率.这里,δ=0.001.

句子中的词汇不是孤立的个体,每个词或多或少地与其他词形成一定的搭配关系.如果一个词与上下文具有较强的搭配关系,那么翻译该词或替换该词时,必须考虑上下文中与之搭配的词汇,见下例所示.

例(D).

待翻译句子:

May I take a *shower*?

翻译实例:

May I take a *picture*?

我 可以 拍 个 *照* 吗?

翻译实例中的“take”和“picture”构成了搭配,故“take”的含义受“picture”影响.虽然翻译实例匹配了待翻译句子,但“picture”的译文不能被“shower”的译文直接替换,否则,我们将得到不恰当的译文“我 可以 *拍* 个 *淋浴* 吗?”.也就是说,我们需要区别对待句子中每一个词,如果该词和其他词具有较强的搭配强度,那么这个词的译文就会受到上下文中搭配词的影响,修改这样的词会增加译文的错误风险(本文称为“编辑风险”).本文使用词与词之间的搭配概率来估计句子中词的编辑风险,如

$$r(w_i,w_1^m)=\frac{\sum_{i\neq j}r(w_i,w_j)}{m} \tag{8}$$

其中, w_1^m 表示句子,w_i 表示句子中的任意一个单词,$r(w_i,w_j)$表示两个词的搭配概率.

这样,在基于编辑距离计算两个句子的相似度时,除了不一致词之间的相似度,还需要考虑不一致词汇的编辑风险.因此,不一致词之间的相似度表示为

$$\Delta(w,\tilde{w})=\begin{cases}dist(w,\tilde{w})^{\beta_1}\times r(\tilde{w},\tilde{w}_1^m)^{\beta_2}\times r(w,w_1^n)^{\beta_3}, & w\neq\tilde{w}\\0, & w=\tilde{w}\end{cases} \tag{9}$$

其中,β_i是权重.

3.2 译文选择算法

当找到和待翻译句子匹配的翻译实例之后,EBMT 系统就要通过修改翻译实例得到最终译文.如果待翻译句子中与翻译实例不一致的单词的候选译文有多个,那么,EBMT 系统就要选择一个最好的译文,通常使用候选译文的翻译概率和语言模型概率来评价译文的质量,具体方法是:

1) 根据翻译实例的目标语句子和待翻译句子中替换词的候选译文,构造出一个翻译网格,网格的每个单元是一个候选译文,例如:

例(E).

待翻译句子:

Please develop these films.

翻译实例:

Please check these forms.

请 检查 一下 这些 表格.

由上面的待翻译句子与翻译实例构造出翻译网格,见表 1.

Table 1 Example of translation searching grid

表 1 翻译搜索网格示例

$e_{t=1}$	$e_{t=2}$ (develop 的译文)	$e_{t=3}$	$e_{t=4}$	$e_{t=5}$ (film 的译文)	$e_{t=6}$
	研制 0.435			归档 0.3385	
	开发 0.16			锉刀 0.2708	
	发展 0.155			案卷 0.1425	
请 1.0	发达 0.115	一下 1.0	这些 1.0	存档 0.1041	. 1.0
	冲洗 0.05			胶卷 0.0576	
	开展 0.04			文件 0.0065	
	…			…	

2) 使用 Viterbi 搜索算法在网格中搜索一条得分最高的路径,算法为

$$V(t,i)=\arg\max_j\{V(t-1,j)+\log(p_T(e_{t,i}\mid f_{a_t}))+\log(p_{LM}(e_{t,i}\mid e_{t-1,j}))\} \tag{10}$$

其中,$p_{LM}(w_{n,i}|w_{n-1,j})$是基于 N-gram 的目标语语言模型概率; $p_T(e_{t,i}\mid f_{a_t})$ 是候选译文的翻译概率,对于翻译实例中不需要替换的词,其翻译概率为 1.0.候选译文的翻译概率是从对齐的双语语料中估计出来的,如

$$p_T(e\mid f)=\frac{\#(f,e)}{\sum_{e'}\#(f,e')} \tag{11}$$

3) 基于找到的最优路径,输出该路径上的词汇,生成最终译文.

从上面的译文搜索过程可以看出,虽然候选译文的翻译概率和语言模型概率分别通过考虑候选译文对应的源语言词汇和候选译文的上下文信息来进行译文选择,但是这两个模型都没有考虑词汇的搭配情况,尤其是不连续词汇的搭配情况.例如上面的例子中,最终译文是“请 研制 一下 这些 胶卷.”但是在该译文中,“研制”是不恰当的译文.

在自然语言处理中,最早应用搭配资源的目的之一就是消除歧义,人们之所以能够在一定的上下文中理解多义词的不同意义,正是借助于这些彼此独立并且呈互补分布特征的搭配信息.认知语言学家的观察证明,人们通常仅仅利用上下文中的一个词或少数几个词就能够识别出多义词的词义.因此,可以根据词与词之间的搭配关系来正确翻译多义词.基于此理论,除了翻译概率和语言模型概率,我们在译文评价中加入了统计搭配概率.不过,由于互为搭配的词对可能是不相邻的,例如上例中的“冲洗”和“胶卷”,因此在上面的动态规划算法中无法直接引入搭配概率.不过在本文中,我们采取了近似方法,即在网格中的每个节点上,动态规划算法保留了得分最高的前 N 条回溯路径,该算法最后输出得分最高的前 N 个候选译文;然后,对于每个候选译文,我们使用单语词

对齐算法估计其中词汇之间的搭配概率.这样,在合并了翻译概率、语言模型概率和搭配概率的得分后,我们选出得分最高的译文 T^*作为最终译文.

$$T^* = \arg\max_T \{\log(p_{\text{MWA3}}(T)) + V(T)\} \tag{12}$$

其中,$V(T)$是译文 T 的 Viterbi 搜索得分, $p_{\text{MWA3}}(T)$ 是译文 T 的单语统计词对齐的最高对齐概率.

3.3 搭配词选择算法

从上面的介绍可以看出,EBMT 方法的特点之一是仅编辑翻译实例中与待翻译句子不一致的词汇,而避免修改其他匹配的部分.这样,通过减少编辑次数,有效降低了产生错误的可能性.但是,这可能会产生最终译文中词汇搭配不一致的问题,如第 1 节中的例(C).针对该问题,我们使用统计搭配模型检测与被替换词汇可能构成搭配的句子中的其他词汇,并且在译文生成过程中,通过利用上下文信息来进一步矫正这些词汇,具体算法如下.

首先,我们检测翻译实例中的哪些词和被替换词具有较强的搭配关系.使用统计搭配模型来计算被替换词和其他词之间的搭配强度,如果搭配强度大于句子的平均搭配强度,那么我们找到该词在待翻译句子中对应的源语言词汇;然后使用该源语言词汇的译文来替换该词,如果待翻译句子没有对应的源语言词汇,那么使用源语言到目标语词典和目标语到源语言的翻译词典来构造该词的同义词,如

$$Syn(w)=\{w'|p_T(w_s|w)>0 \ \& \ p_T(w'|w_s)>0\} \tag{13}$$

其中,候选译文 w'的翻译概率使用两个翻译概率的乘积 $p_T(w_s|w)\cdot p_T(w'|w_s)$来表示.

然后,在扩展后的搜索网格中,使用公式(12)描述的算法来搜索最佳译文.

3.4 译文排序

如果翻译系统从翻译实例库中找到多个与待翻译句子匹配的翻译实例,并且使用这些翻译实例生成了多个候选译文,我们使用下面的方法来综合评价译文的质量,包括待翻译句子和翻译实例之间的相似度、译文的生成得分及译文中词的搭配概率,如

$$T^* = \arg\max_T \{L(E,F)\times \exp(V(T;E))\times p_{\text{MWA3}}(T)\} \tag{14}$$

其中,$L(E,F)$表示待翻译句子和翻译实例之间的相似度,$p_{\text{MWA3}}(T)$表示译文 T 的单语统计词对齐的最高对齐概率,$V(T;E)$表示基于翻译概率和语言模型概率的生成得分.

最后,在候选译文中,选择得分最高的译文作为最终译文.

4 实验与分析

本节通过一系列实验对本文提出的方法进行了评测,实验数据来源于公开的英汉口语语料和口语翻译评测集.评价内容包括:(1) 在基于词的 EBMT 系统中,详细评价翻译实例选择、译文选择、搭配词选择以及译文排序的算法性能;(2) 在半结构化的 EBMT 系统中,验证译文选择算法的效果;(3) 基于相同的实验数据,对 EBMT 系统和基于短语的 SMT 系统的性能进行比较;(4) 对 EBMT 系统和 SMT 系统的译文进行人工评价,进一步考察本文提出方法的有效性.

4.1 实验设置

4.1.1 翻译实例库

翻译实例库采用了公开的 HIT 英汉双语语料库(http://mitlab.hit.edu.cn/index.php/resources/29-the-resource/111-share-bilingual-corpus.html),该语料库的基本情况见表 2.

Table 2 Statistics of translation examples

表 2 翻译实例库的基本情况

语言	句子数	词数
中文	130 393	1 299 881
英文	130 393	1 494 869

其中,对汉语句子进行分词处理(http://ir.hit.edu.cn/ demo/ltp/),对英语句子进行了 tokenization 处理.采用了 GIZA++(http://fjoch.com/GIZA++.html)自动获得了双语语料的词对齐信息.

4.1.2 测试集

基于上面的翻译实例库,本文在 IWSLT 的英汉测试集(IWSLT05 和 IWSLT08)上评测了中文的译文质量.由于 IWSLT08 的中文句子直接来源于口语语音识别模块的输出结果,所以没有标点信息.为了与测试语料保持一致,我们为其中的中文句子人工添加了标点信息,最终使用的测试集见表 3.

Table 3 Statistics of test sets

表 3 测试集的基本情况

测试集	句子数	词数	参考译文个数
IWSLT05	506	3 769	7
IWSLT08	251	1 320	7

其中,使用与双语例句相同的方法处理了测试语料的源语言部分.

4.1.3 译文评测标准

对译文质量的自动评价使用了 BLEU[22].BLEU 是通过 EBMT 系统生成的译文和参考译文之间 *N*-gram 的匹配情况来评价译文的质量,BLEU 得分越高,说明译文和参考译文越接近,译文的质量越好.对于汉语译文,本文使用了基于字的 4-gram BLEU 得分.

为了检查实验结果的有效性,我们对实验结果进行了统计有效性校验[23].

4.1.4 语言模型

使用 SRILM 语言模型工具[24]在双语语料的汉语部分上训练了 3-gram 语言模型.

4.1.5 统计搭配模型

统计搭配模型的训练语料来源于双语语料中的单语部分,由于标点很少参与构成搭配,所以在进行单语词对齐之前删除了单语语料中的标点.最终,中文和英文统计搭配模型分别得到 173 407 和 166 810 个潜在搭配词对.

4.1.6 翻译词典

基于词对齐的双语语料,我们构造了具有翻译概率的翻译词典,其中,英汉词典包括 49 250 个词条,该词典用于为待翻译句子中不一致的词汇提供译文.

4.2 基于词的EBMT系统

本文开发了一个基于词的英汉 EBMT 系统,其中,基线系统首先采用了公式(6)来计算翻译实例和待翻译句子的相似度,然后,使用公式(10)来搜索最佳译文.最后,待翻译句子的最终译文从与待翻译句子具有最高相似度的翻译实例中生成得到.基于该基线系统,我们详细调查了本文提出的翻译实例选择、译文选择、搭配词选择及译文排序算法的可行性.表 4 显示出了实验结果.

在基线系统的基础上,我们首先评测了翻译实例选择算法的性能.其中,翻译实例选择使用了公式(9)中的算法,β_i 分别设为 0.6,0.2 和 0.2.表 4 的"+翻译实例选择"中显示了使用该算法后的译文评价结果,从中可以看出,利用翻译实例选择算法,译文评价得分得到了显著提高,在两个测试集上,BLEU 得分提高了 0.91~0.95.我们比较了 IWSLT05 最终译文对应的翻译实例,在基线系统中,按照最小编辑距离来选择翻译实例,翻译实例的平均编辑次数为 1.46 次;应用翻译实例选择算法之后,平均编辑次数变为 1.64.虽然平均编辑次数上升了 12.3%,但是这些被编辑的词汇具有较低的编辑风险,因此,最终译文的平均质量得到提高.

Table 4 Experimental results of EBMT systems

表 4 EBMT 系统的翻译结果

翻译系统	IWSLT05	IWSLT08
基线系统(MT+LM)	32.82	32.59
+翻译实例选择	33.73	33.54
+译文选择	35.27	36.27
+搭配词选择	36.48	37.47
+译文排序	37.55	39.07

然后,我们在“基线系统+翻译实例选择”的基础上增加了公式(12)中的译文选择算法.在动态规划算法中,N 设置为 100.同样,待翻译句子的最优例句来源于与待翻译句子相似度最高的翻译实例.译文评价结果显示,在“+译文选择”中,译文搜索过程中加入搭配信息,大幅度提高了译文的 BLEU 得分.该结果说明,虽然系统已经使用了语言模型,但是考虑了词汇的搭配情况之后(搭配词汇不仅有相邻词汇,也有不相邻的词汇),有效增强了 EBMT 系统中译文选择的能力.

接下来,在译文搜索过程中又考虑了搭配词的选择算法.译文评价结果显示在“+搭配词选择”中,通过在翻译实例中对替换词的搭配词进行矫正,译文质量得到进一步提高.我们计算了测试集 IWSLT05 上被检测到的搭配词个数为 216,通过与参考译文进行比较,其中 53%的搭配词被修改正确,13%的搭配词被修改错误,剩下 34%的搭配词的修改并没有影响译文质量.所以,该算法对翻译实例中替换词的搭配词进行矫正是有效的.

最后,EBMT 系统综合使用了本文提出的方法,并且利用公式(14)从可能的候选译文中为待翻译句子选择最优译文.译文评价结果显示在“+译文排序”中,与基线系统相比,BLEU 得分在两个测试集上分别提高了 4.73 和 6.48.统计显著性校验结果显示,该结果的置信度 $p<0.01$.该结果说明,统计搭配模型应用到 EBMT 系统中,不仅有效提高了 EBMT 系统的性能,还提高了译文质量.

4.3 半结构化的EBMT系统

译文选择是很多机器翻译系统要解决的关键问题之一,从表 4 的实验结果可以看出,使用统计搭配模型增强的译文选择算法在基于词的 EBMT 系统中有效提高了译文质量.为了进一步验证本方法的有效性,我们在半结构化的 EBMT 模型[13]中也考察了该译文选择算法.

4.3.1 半结构化的 EBMT 翻译模型

半结构化的 EBMT 模型将双语翻译实例表示为树串映射(tree string correspondence,简称 TSC),一个树串映射由 3 部分组成:源语言分析树、目标语词串以及它们之间互译词的对应关系,如图 2 所示.其中,目标语中的〈**NPB**〉是替换符号,该符号被对应的子 TSC 的候选译文替换.

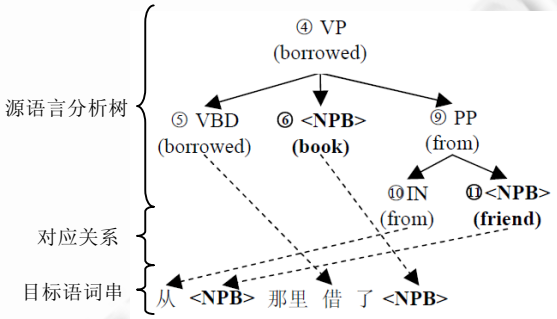


Fig.2 Translation example of semi-structured EBMT system

图 2 半结构化 EBMT 系统的翻译实例

对于给定的待翻译句子,该 EBMT 模型首先使用源语言分析器得到待翻译句子的分析树;然后,在翻译实例库中搜索与该分析树最匹配的树串映射森林;最后,使用线性对数生成模型产生最佳译文.与第 4.2 节中使用的

EBMT 模型相比,半结构化的 EBMT 系统在翻译过程中使用了更丰富的特征:待翻译句子与翻译实例之间的匹配得分和上下文相似度、翻译实例的翻译概率和目标语选择概率、译文的语言模型概率和长度概率.但是,这些特征都没有考虑不相邻词汇之间的搭配关系.在本实验中已有特征的基础上,我们引入了目标语统计搭配概率作为一个新特征,通过考察目标语字符串中的替换符号所对应的候选译文与上下文词汇之间的搭配关系,来帮助提高系统的译文生成能力.具体来说,就是对于 TSC 的每一个候选译文 T,使用其最大的单语统计词对齐概率作为该译文的搭配特征,如公式(15)所示.

$$f_{\text{collocation}}(T) = \arg\max_{A} p_{\text{MWA3}}(A \mid T) \tag{15}$$

本实验除了使用第 4.1 节介绍的资源之外,我们还使用 Collins 英语分析器分析了翻译实例中的英语句子和测试集中的英语句子[25].

在本实验中,使用 IWSLT05 作为开发集调整半结构化 EBMT 系统中特征的权重.

4.3.2 实验结果

表 5 显示了译文评价结果,从该结果可以看出,通过在系统中引入统计搭配模型估计译文中词与词之间的搭配概率,译文质量得到了显著提高(统计显著性校验结果显示,置信度 p<0.01).与基线系统相比,BLEU 得分提高了 1.82.

Table 5 Experimental results of semi-structured EBMT systems

表 5 半结构化 EBMT 系统的翻译结果

翻译系统	IWSLT08
半结构化的 EBMT 系统	40.44
+译文选择	42.26

通过基于词的 EBMT 系统和半结构化的 EBMT 系统上的实验结果可以看出,基于统计搭配模型的译文选择算法对于提高 EBMT 模型的性能具有普遍意义.

4.4 EBMT系统与SMT系统的比较

在第 4.1 节描述的数据上,我们比较了上述 EBMT 系统与基于短语的 SMT 系统的译文质量,以此来检验上述 EBMT 系统的性能.基于短语的 SMT 系统采用了 Moses 系统[26],该系统使用了和 EBMT 系统相同的语言模型,翻译模型使用了与 EBMT 系统相同的双语语料来训练.我们使用 IWSLT05 作为开发集来调模型参数的权重[27],IWSLT08 作为测试集来评价系统的性能.

表 6 显示了 EBMT 系统和 SMT 系统的译文评价结果.基于词的 EBMT 系统融合了统计搭配模型之后,取得了和 SMT 系统差不多的译文质量,半结构化的 EBMT 系统在引入基于统计搭配模型的译文选择算法之后,译文的 BLEU 得分明显高于 SMT 系统(统计显著性校验显示,置信度 p<0.01).

Table 6 Comparison between EBMT systems and phrase-based SMT system

表 6 基于短语的 SMT 和 EBMT 的比较结果

翻译系统	IWSLT08
Moses	40.21
基于词的 EBMT+译文排序	39.07
半结构化的 EBMT+译文选择	42.26

4.5 人工评价

为了进一步检验译文的质量,我们对 IWSLT08 的译文进行了人工评价,评价标准采用了流利度(fluency)和忠实度(adequacy)两个标准.忠实度反映的是机器翻译系统生成的译文在多大程度上忠实于原文所要表达的意思,流利度用于评价译文本身是否流畅、是否符合目标语言的表达习惯等,见表 7.

Table 7 Standard of fluency and adequacy

表 7 流利度和忠实度的评价标准

等级	流利度	忠实度
5	句子是流畅并且地道的句子	译文准确完整地表达了原文的信息
4	句子流畅,但是不够地道	译文表达了原文的大部分信息
3	句子基本流畅	译文基本表达了原文的意思
2	句子很不流畅	译文只有少数内容符合原文的意思
1	句子不可理解	译文基本没有表达原文的意思

为了客观公正地进行评价,我们开发了一个评价工具.在评价之前,该工具会把每个待翻译句子对应的翻译系统的译文顺序进行随机排列;当完成人工评价时,该工具在评价完的数据中找回译文的原始顺序,并且计算每个系统的平均得分.

在本实验中,我们评价了 5 个系统的译文,按照流利度、忠实度分别对这些译文进行人工评价.评价结果见表 8.

Table 8 Results of human evaluation

表 8 人工评价结果

翻译系统	流利度(aluency)	忠实度(adequacy)
Moses	3.90	3.70
基于词的 EBMT	3.57	3.47
基于词的 EBMT+译文排序	3.85	3.74
半结构化的 EBMT	4.04	3.83
半结构化的 EBMT+译文选择	4.14	3.96

人工评价的结果显示:本文提出的方法应用到基于词的 EBMT 系统和半结构化的 EBMT 系统中,译文质量都得到了提高;尤其是基于词的 EBMT 系统,融合了统计搭配模型以后,译文质量无论是流利度得分还是忠实度得分都得到了提高;尽管流利度略低于 SMT 系统,但是忠实度取得了和 SMT 系统差不多的得分.虽然半结构化的 EBMT 系统已经应用了多种特征来评价候选译文的质量,但该系统忽略了词汇之间的长距离搭配关系.从该结果还可以看出,应用了基于统计搭配模型的译文选择方法之后,半结构化的 EBMT 系统的译文质量得到了提高.按照表 7 中的评价标准,半结构化的 EBMT 系统产生的译文基本上表达了原文的大部分信息(*fluency*=4),并且译文流利度也基本可以接受(*adequacy*=4).

5 结 论

本文首次提出使用统计搭配模型提高 EBMT 系统的性能,具体来说,就是通过估计待翻译句子与翻译实例之间的匹配度来提高 EBMT 系统对翻译实例的选择能力;通过估计候选译文词汇与上下文之间的搭配强度,提高系统的译文选择能力;通过检测句子中的搭配词并且根据上下文对其进行矫正,进一步提高译文质量.我们的方法应用到基于词的 EBMT 系统,使得最终译文的 BLEU 得分显著提高了 4.73~6.48 个百分点.在半结构化的 EBMT 系统中引入基于统计搭配模型的译文选择方法来估计候选译文中词与词之间的搭配关系,进一步提高了译文的 BLEU 得分.本方法对于 EBMT 系统是普遍有效的,同时,人工评价的结果显示,半结构化的 EBMT 系统产生的译文能够表达原文的大部分信息,并且译文流利度也基本可以接受.

References:

[1] Nagao M. A framework of a mechanical translation between Japanese and English by analogy principle. In: Proc. of the Int'l NATO Symp. on Artificial and Human Intelligence. 1984. 173−180.

[2] Somers H. Review article: Example-Based machine translation. Machine Translation, 1999,14(2):113−157. [doi: 10.1023/A:1008109312730]

[3] Matsumoto Y, Ishimoto H, Utsuro T. Structural matching of parallel texts. In: Proc. of the 31st Annual Meeting of the Association for Computational Linguistics. 1993. 23−30. [doi: 10.3115/981574.981578]

[4] Al-Adhaileh MH, Tang EK. Example-Based machine translation based on the synchronous SSTC annotation schema. In: Proc. of the Machine Translation Summit VII. 1999. 244−249.

[5] Liu ZY, Wang HF, Wu H. Example-Based machine translation based on tree-string correspondence and statistical generation. Machine Translation, 2006,20(1):25−41. [doi: 10.1007/s10590-006-9016-4]

[6] Luk AK. Statistical sense disambiguation with relatively small corpora using dictionary definitions. In: Proc. of the 33rd Annual Meeting on Association for Computational Linguistics. 1995. 181−188. [doi: 10.3115/981658.981683]

[7] Gale W, Church K, Yarowsky D. A method for disambiguation word senses in a large corpus. Computer and Humanities, 1993,26: 415−439.

[8] Towell G, Voorhees EM. Disambiguating highly ambiguous words. Computational Linguistics, 1999,24(1):125−145.

[9] Yarowsky D. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In: Proc. of the 32nd Annual Meeting on Association for Computational Linguistics. 1994. 88−95. [doi: 10.3115/981732.981745]

[10] Akiba Y, Watanabe T, Sumita E. Using language and translation models to select the best among outputs from multiple MT systems. In: Proc. of the 19th Int'l Conf. on Computational Linguistics. 2002. 8−14. [doi: 10.3115/1072228.1072304]

[11] Imamura K, Okuma H, Watanabe T, Sumita E. Example-Based machine translation based on syntactic transfer with statistical models. In: Proc. of the 20th Int'l Conf. on Computational Linguistics. 2004. 99−105. [doi: 10.3115/1220355.1220370]

[12] Carl M, Schmidt P, Schutz J. Reversible template-based shake & bake generation. In: Proc. of the MT Summit X Workshop on Example-Based Machine Translation. 2005. 17−25.

[13] Liu ZY, Wang HF, Wu H. Log-Linear generation models for example-based machine translation. In: Proc. of the MT Summit XI. 2007. 305−312.

[14] Brown PF, Stephen DP, Vincent JDP, Robert LM. Word-Sense disambiguation using statistical methods. In: Proc. of the 29th Annual Meeting of the Association for Computational Linguistics. 1991. 264−270. [doi: 10.3115/981344.981378]

[15] Radev DR, McKeown KR. Generating natural language summaries from multiple on-line sources. Computational Linguistics, 1998, 24(3):470−500. [doi: 10022/AC:P:29341]

[16] Zhao SQ, Zhao L, Liu T, Li S. Paraphrase collocation extraction based on binary classification. Journal of Software, 2010,21(6): 1267−1276 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3586.htm [doi: 10.3724/SP.J.1001.2010.03586]

[17] McKeown KR, Radev DR. Collocations. In: Dale R, Moisl H, Somer H, eds. A Handbook of Natural Language Processing. New York: Marcel Dekker, 2000. 507−523.

[18] Liu ZY, Wang HF, Wu H, Li S. Collocation extraction using monolingual word alignment method. In: Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing. 2009. 487−495.

[19] Manning CD, Schütze H. Foundations of Statistical Natural Language Processing. Cambridge, London: Bradford Book and MIT Press, 1999.

[20] Liu ZY, Wang HF, Wu H, Li S. Improving statistical machine translation with monolingual collocation. In: Proc. of the 48th Annual Meeting of the Association for Computational Linguistics. 2010. 825−833.

[21] Liu ZY, Wang HF, Wu H, Li S. Reordering with source language collocations. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. 2011. 1036−1044.

[22] Papineni K, Roukos S, Ward T, Zhu W. BLEU: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. 2002. 311−318. [doi: 10.3115/1073083.1073135]

[23] Philipp K. Statistical significance tests for machine translation evaluation. In: Proc. of the 2004 Conf. on Empirical Methods in Natural Language Processing. 2004. 388−395.

[24] Stolcke A. SRILM—An extensible language modeling toolkit. In: Proc. of the 7th Int'l Conf. on Spoken Language Processing. 2002. 901−904.

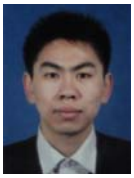
[25] Collins M. Head-Driven Statistical Models for Natural Language Parsing. University of Pennsylvania, 1999.

[26] Franz JO. Minimum error rate training in statistical machine translation. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. 2003. 160−167. [doi: 10.3115/1075096.1075117]

[27] Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E. Moses: Open source toolkit for statistical machine translation. In: Proc. of the 45th Annual Meeting of the ACL. Poster and Demonstration Sessions, 2007. 177−180.

附中文参考文献:

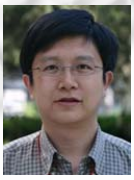
[16] 赵世奇,赵琳,刘挺,李生.基于二元分类的复述搭配抽取.软件学报,2010,21(6):1267−1276. http://www.jos.org.cn/1000-9825/3586.htm [doi: 10.3724/SP.J.1001.2010.03586]



刘占一(1978－),男,河北邢台人,博士生,主要研究领域为自然语言处理,机器翻译.

刘挺(1972－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,信息检索.

李生(1943－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理.



王海峰(1971－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理.