基于句对比较的自动获取翻译模板方法改进

方 淼,关小薇,高庆狮
FANG Miao,GUAN Xiao-wei,GAO Qing-shi

大连理工大学 计算机科学与工程系,辽宁 大连 116024
Department of Computer Science and Engineering,Dalian University of Technology,Dalian,Liaoning 116024,China

FANG Miao,GUAN Xiao-wei,GAO Qing-shi.Improvement of automatic acquisition of translation template based on sentence pairs comparison.Computer Engineering and Applications,2007,43(34):16-18.

Abstract: Translation template is not only vital resource for machine translation,but also useful linguistic knowledge.Improves the analogy technique to induce translation template from given bilingual examples by filtering with the word alignment of the examples,and presents Translation Template with Functional Relationship(TTFR) to capture dependencies such as "he...his" pairs and similar linguistic phenomena,which is a more general translation template and can be used in bi-direction machine translation directly.Experimental results show that the improved method breaks the limitation of the old learner and increases the accuracy of automatic inducing the translation templates.

Key words: translation template;functional relationship;machine translation;word alignment

摘 要: 翻译模板不仅是机器翻译的重要资源,而且是有用的语言学知识。使用词对齐结果改进了从实例中类比学习翻译模板的方法,并提出了带有函数关系的模板(TTFR)以获取语言之间的依赖关系,如"he…his"对。带函数关系的模板是一个更一般化的模板并能直接用于双向的机器翻译。实验结果表明改进的方法有效地克服类比方法的问题并且提高了自动获取模板的准确率。

关键词: 机器翻译;函数关系;翻译模板;词语对齐
文章编号: 1002-8331(2007)34-0016-03 **文献标识码:** A **中图分类号:** TP18

1 引言

作为基于实例机器翻译的一个重要资源,翻译模板在不影响翻译质量的情况下,增加了实例的覆盖率并且提高了效率。模板学习方法大致可以分为如下几类:(1)对特定词语的聚类泛化形成翻译模板[1]。该方法能提高实例翻译的覆盖率但变量仅限于有限的几类词语。(2)利用双语的句法分析和词语匹配结果在各个句法层次上抽取翻译模板[2]。此类方法需要健壮的双语句法分析器和准确的双语对应方法。(3)类比学习[3]。在这些方法中,类比学习是最简单和有效的一种方法,但是它本身也存在一些问题。

统计机器翻译中也采用一种翻译模板—对齐模板—保持词语的对应关系。Och[4]首先双语词语对齐,然后获取短语,最后通过词类泛化对齐的短语以得到对齐模板。这里短语只是一个连续的词串,不一定有语言学上的意义。

模板的定义和抽取方法各不相同,但是模板的定义和质量在很大的程度上决定着机器翻译的性能。本文使用双语词对齐结果改进了类比学习方法并且提出了一种带有函数关系的翻译模板,该模板能够描述语言之间的相互依赖关系,如英语中的人称和数的一致,汉语中的形式量词对名词的依赖等等。它同时也是一个更一般的翻译模板。第2章将对翻译模板类比学习方法进行介绍,然后第3章分析类比学习产生的错误并提出改进方法。第4章提出带函数关系的翻译模板,第5章是实验分析讨论,第6章是结论。

2 翻译模板和类比学习方法

一种语言可以看作由该语言上的字所组成的字符串的集合,并且该语言上的字是一个有限集。例如,汉语或英语中的字符串是一个字或词的序列,它可以是词、短语或句子,也可以是一组词、短语或者句子。通过变量替换字符串中的词、短语或句子来泛化一个字符串。这个泛化的字符串就是一个翻译模板,它保持了字符串之间的对应关系。不带变量的字符串可以看作是翻译模板的一个特例。

一个通用的翻译模板可以定义为:$T_C \leftrightarrow T_E$(这里以汉语和英语为例)。T_C表示翻译模板的汉语部分,T_E表示翻译模板的英语部分。

$T_C=<C_0,X_i,C_1,X_j,\cdots,C_n,X_k>$

$T_E=<E_0,Y_i,E_1,Y_j,\cdots,E_n,Y_k>$

这里,$C_i(1\leqslant i\leqslant n)$是一个汉语的片断,由汉字所组成的字符串。而$X_i$是汉语模板中的一个变量,它可以被其他的模板或词语替换。同样地,$E_i(1\leqslant i\leqslant n)$是一个英语单词所组成的片

基金项目: 国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2006AA012140);大连理工大学创新基金。

作者简介: 方淼(1980-),男,博士生,主要研究方向:自然语言处理、机器翻译;关小薇(1979-),女,博士生;高庆狮(1934-),男,院士,教授。

段,而 Y_i 是英语模板中的变量。任何一个 C_i,E_i,X_i 和 Y_i 都可以为空(ε),但是不可以模板的一边为空。模板中变量之间有一一对应关系,如 $X_i \leftrightarrow Y_i$,但是片断(C_i 和 E_i)之间不一定有对应关系。例如,下面是 4 个翻译模板:

(1)This morning↔今天早上

(2)I↔我

(3)Y_1 have had enough↔X_1 吃饱了　($X_1 \leftrightarrow Y_1$)

(4)Y_1 in $Y_2 \leftrightarrow X_2$,X_1　($X_1 \leftrightarrow Y_1$,and $X_2 \leftrightarrow Y_2$)

前两个不带变量,而后面两个则带有一个或多个变量。如果使用第二个模板替换第三个模板中的变量,就可以得到互译的双语句对:"I have had enough↔我吃饱了"。同样,如果用这个新句对和第一个模板去替换最后一个模板,可以得到一个更大的互译句对:"I have had enough in this morning↔今天早晨,我吃饱了"。

使用同样的方法,机器翻译可以匹配翻译模板的源语言模板,然后展开其对应的目的语言的部分,同时根据变量进行替换操作。

类比学习器[4](Learner)是一个经典有效的学习方法。它基于一个假设:给定两个翻译实例,源语言中相同的部分应该对应于目标语言中相同的部分,同样源语言中不同部分也应该对应于目标语言中不同的部分。在比较两个实例之后,产生一个双语的匹配序列,然后使用变量替换不同的部分就得到一个翻译候选模板。如果仅有一对不同的部分或者除了至多一个之外其余的不同部分均能相互对应,这个翻译候选模板和仅有的双语不同部分均可以作为新的翻译模板。

例如:通过比较两个句对 "She has had enough↔她吃饱了" 和 "He has had enough↔他吃饱了"可以得到一个匹配序列:"(She,He)have had enough↔(她,他)吃饱了"。类比学习就可以得到三个翻译模板:"Y_1 have had enough↔X_1 吃(eat)饱了(enough)","她↔she"和"他↔he"。

同样,在比较两个句对 "她喜欢吃苹果。↔She loves to eat apple."和"他喜欢吃香蕉。↔He loves to eat banana."之后,类比学习器 Learner 可以得到匹配序列"(她,他)喜欢吃(苹果,香蕉)?(She,He)loves to eat(apple,banana)"。由于已经从上面学习到汉语的不同部分"(她,他)"可以与英语的不同部分"(She,He)"对应,类比学习就可以推断新模板为"X_1 喜欢吃 $X_2 \leftrightarrow Y_1$ loves to eat Y_2","苹果↔apple"和"香蕉↔banana"。

3　类比学习的错误及改进

在真实的语料库中常常包含很多不满足类比假设的双语句对。它们会导致 Learner 失败。下面实例就是这些例外(每组的前两行是待比较的实例,后面是 Learner 从这些实例中抽取出来的翻译模板)。

(1)他我行我素。↔He emerged as his own man.
他是一位老人。↔He is an old man.
他 X_1。 He X_1 man.
我行我素　emerged as his own
是一位老人　is an old

(2)他力大如牛。↔He is as strong as an ox.
他是教师。↔He is a teacher.
他 X_1。He is X_1
是教师　a teacher

(3)他是个自相矛盾的人。↔He was a paradox.
他是一位老人。↔He is an old man.
他是 X_1。He X_1.
个自相矛盾的人　was a paradox
一位老人 is an old man

(4)你给她回信了吗?↔Have you answered her letter?
你下定决心了吗?↔Have you made up your mind?
你 X_1 吗?　Have X_1?
下定决心 made up your mind

从这 4 组翻译实例中可以看出,两个英语句子的相同部分并不能对应汉语句子中的相同部分,不同的部分之间也没有正确的对应关系。这是因为一种语言中的字或词和另一种语言中的字或词之间有着复杂的对应关系,如一对一(1:1),一对多(1:m 和 m:1),以及多对多(m:n),有时甚至有一对空(1:0)的情况出现,在一种语言中两个句子的相似部分可能不等于另一种语言中的相似部分。如例 2 中"he is"与"他"就不相等。而且单语中存在的语言单位之间的依赖关系也会造成双语中的相似部分不对等,如例 4 的第二个句子中英语单词"you"和"your"之间有相互依赖关系,Learner 无法捕获这种依赖关系,最终导致错误的结果。

类比学习器难以正确处理上面的两类情况。在对统计词对齐结果进行研究之后,发现借助正确的词对齐结果可以消除第一类情况的影响,即由于 Learner 在复杂对应情况下造成错误。

开源词对齐工具 GIZA++[1] 针对第三个实例的两个句子产生对齐结果如 "1:1 2:2 3:3 4:3 4:4 4:6 5:7"and "1:1 2:2 3:3 3:4 4:4 4:5 5:5 6:6"。其中,数字表示双语中的位置,用冒号":"隔开,前面表示英语词的位置,后面汉语词的位置。

虽然现有的词对齐程序会有一些错误,不能做到理想的对齐。但是其结果仍可以用来过滤类比学习器的错误。提出一个算法过滤错误的结果,如图 1 所示。

```
输入:两个双语句对的匹配序列,这两个句对的词对齐结果
输出:TRUE/FALSE
1. 根据第一个句对的词对齐结果获取匹配序列中的英文单
   词的所有中文词语对应,组成集合 SC
2. for 每一个匹配序列中的中文词语
     If 该词语不在 SC 中 then
       Return FALSE;
     取其所有英文对应单词加入到集合 SE 中
3. for 匹配序列中的每一个英语单词
     If 该单词不在 SE 中 then
       Return FALSE;
4. 对第二个句对重复上述过程 2-3
5. Return TRUE;
```

图 1　过滤算法

过滤算法分别在检查每一个句对中是否存在交叉现象。交叉现象就是词对齐结果中相互对应的两个词却分别出现在相同部分和差异部分中。如果有交叉现象存在,过滤算法返回"FALSE",忽略其结果。通过此过滤算法,大部分复杂对应中的部分对应带来的错误被排除掉。对于语言单位之间的依赖关系导致的错误,提出一种带有函数关系的模板进行校正。

[1] www.fjoch.com/GIZA++.html

4 带函数关系的翻译模板

在一种语言中,某一个语言单位(如词)的选择和使用需要依赖于其它相关的语言单位或受其他语言单位的影响的这种关系称为函数关系[5]。例如,“She shakes her head”中的“her”需要与“She”保持一致。“university president”和“U.S.A president”译成中文就分别为“大学校长”和“美国总统”。“校长”和“总统”都是最高行政长官的意思,受另外两个词“大学”、“美国”的影响。

为了形式化地表示函数关系,引入一个记号:F形式,$F(arg_0, arg_1, \cdots, arg_n)$表示一个函数关系。其中$arg_0$表示该函数的名称,$arg_1 \sim arg_n$表示相关的参数(变量)。函数关系是与具体语言紧密相关,本文仅以汉语和英语为例列举一些典型的函数关系。

4.1 汉英语言中的函数关系

4.1.1 英语中的代词一致

“She shakes her head?她摇摇头。”中的“her”指向“She”,这是一种函数关系。因为词语“her”依赖词语“she”,或者说“shake one's head”是一个固定的表达,“one's”只能是物主代词的形式出现,并且要和主语保持一致。函数关系表示为“She shakes F(PPA,she') head”,PPA 表示这个依赖关系是代词依赖关系。

4.1.2 英语中的其他函数关系

“浓”表示“密度大”,然而在具体表达“浓”这个概念的选词上依赖于其它的单词。“浓茶↔strong tea;浓墨↔thick ink;浓烟↔dense smoke”。它们的函数关系分别表示成“$F(C$,浓',tea)tea”,“$F(C$,浓',ink)ink”,$F(C$,浓',smoke)smoke”。其中,C 表示上下文关系,在该处出现的词的意义取决于浓(arg_1),同时表达又依赖于 arg_2。该函数关系可以作为翻译时的选词模板来使用,达到语义消歧目的。

4.1.3 汉语中的形式量词

汉语里的量词和其后的名词有着密切的关系,随着名词的变化而变化。可以用公式 $F(L,N')$来表示。如 “一支笔”、“一本书”、“三个苹果”它们分别表示成“一 $F(L$,笔')笔”,“一 $F(L$,书')书”,和“一 $F(L$,树')树”。其中,笔'代表“笔”一类意义的词,其它类似。

4.1.4 汉语中的其他函数关系

“good”的意思是“美好的,良好的,令人满意的”,但是在汉语中的表示不仅受本身意义的影响, 而且也受到其他词的影响。如:“good soil↔肥沃的土壤”,“good oil↔提纯了的油”,“good money↔真的货币”。函数关系分别为:$F(C$,good',土壤)土壤,$F(C$,good',油)油,$F(C$,good',货币)货币。

4.2 带函数关系的翻译模板

根据函数关系,扩展翻译模板($T_C \leftrightarrow T_E$)以包含函数关系的 F 形式。

$$T_C=<C_0, X_i, F_1, C_1, X_j, F_2, \cdots, C_n, X_k, F_n>$$

$$T_E=<E_0, Y_i, F'_1, E_1, Y_j, F'_2, \cdots, E_n, Y_k, F'_n>$$

其中,C_i,E_i,X_i和 Y_i的定义不变。F_i和 F'_i分别表示双语中的函数关系,F_i中含有变量,既可以是其它变量,也可以有独立的变量。其它约束条件也不变

例 1 模板 “X_1<$N_{人}$>摇头表示 X_2<$N_{态度}$>↔Y_1<$N_{人}$>shook $F(PPA, Y_1<N_{人}>)$head in Y_2<$N_{态度}$>”[2]的汉语部分有一个不变量和两个变量;而英语部分包含两个不变量,两个变量和一个函数,函数包含的变量是第一个变量“Y_1<$N_{人}$>”;汉英的两个变量分别一一对应。在翻译的时候,如果输入汉语句子“他摇头表示拒绝”,当前模板与其匹配,得到第一个变量“他”和第二个变量“拒绝”,分别取两个变量的译文“He”和“refusal”,在转换到英语的时候发现还有一个函数“$F(PPA, Y_1<N_{人}>)$”,根据函数关系函数可以得到其英语表示“his”。因此,英语的译文为:“he shook his head in refusal.”反过来,如果输入句子“She shook her head in disapproval.”,在匹配翻译模板的时候发现两个变量“She”和“disapproval”,并且“her”是第一个变量“She”的函数关系表示。那么直接应用模板进行翻译,可得到如下汉语句子:“她摇头表示不赞成。”

例 2 模板“一 $F(L, X_1<N>)X_1<N>\leftrightarrow F(Article, Y_1<N>)Y_1<N>$”表示了英语里一个冠词函数和名词变量构成模板的英语结构,在汉语中是一个数字“一”和形式量词函数以及名词变量结构,这两个结构中名词变量相互对应互译构成了翻译模板。当输入一个英语短语“a doctor”,与模板匹配时发现与当前模板的英语部分匹配,“doctor”对应于名词变量,“a”为冠词函数,那么可以应用当前模板进行翻译,得到汉语译文:“一名医生”。

同样,当输入是一个汉语短语“一架飞机”时,发现“架”为形式量词,“飞机” 为名词变量, 应用模板翻译得到英语译文:“an airplane”。

从上面的例子可以看出,带函数关系的翻译模板能够处理同一种语言中语言单位(词语)之间的依赖关系,从而在翻译的过程中准确地把一种语言中翻译成另一种语言。

5 实验分析

本文实现了汉英双语的类比学习器,并对其进行改进。然后分别进行翻译测试。

5.1 语料和预处理

选择 1 000 个通用文本双语句对进行训练。首先,使用工具 tokenizeE.perl.tmpl[3] 对英语进行断词(tokenization)和词形还原并使用工具 ICTCLAS[4] 对汉语进行分词。然后,使用 Learner 从语料中抽取翻译模板。接下来,是用一个树剪枝的翻译程序根据学习到的模板翻译测试句子。然后是用词对齐工具 GIZA++对齐训练的双语句对并利用其结果过滤学习到的错误翻译模板。并对测试句对进行翻译。最后,使用带有函数关系的模板进一步改正翻译模板的错误。

5.2 实验结果

类比学习方法(Learner)以及改进方法(A-Learner)的性能如表 1 所示。其中“TT-SUM”表示获取的翻译模板总数,“Acc.”表示正确的翻译模板占所获取的模板的比例,“Error”表示错误模板占所获取的模板的比例。从表 1 中可以看出,A-Learner 的准确率大大提高,说明本方法是有效的。

表 1 学习器性能比较

	TT-SUM	Acc./%	Error/%
Learner	8 301	32.54	67.46
A-Learner	1 628	85.67	14.33

不必对每一个句子都进行翻译, 随机从训练集中选择了 100 个句子作为测试数据。分别对 Learner,A-Learner 以及 A-

[2] 其中<N_{PERSON}>表示变量 X_1的类型,表示人。 [3] http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/ [4] http://www.nlp.org.cn

(下转 28 页)