中 文 信 息 学 报
第 20 卷 增刊 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 SUPPLEMENT

文章编号:1003 -0077(2006)增刊 -0047 -06

大规模句子相似度计算方法

黄河燕[1],陈肇雄[1],张孝飞[1],张克亮[1,2]

(1. 中国科学院计算机语言信息工程研究中心,北京 100083; 2. 南京理工大学,江苏 南京 210094)

摘要:如何根据源语言文本从大规模语料库中找出其最相近的翻译实例,即句子相似度计算,是基于实例翻译方法的关键问题之一。本文提出一种多层次句子相似度计算方法:首先基于句子的词表层特征和信息熵从大规模语料库中选择出少量候选实例,然后针对这些候选实例进行泛化匹配,从而计算出相似句子。在多策略机器翻译系统 IHSMTS 中的实验表明,当语料规模为 20 万英汉句对时,系统提取相似句子的召回率达 96%,准确率达 90%,充分说明了本文算法的有效性。

关键词:句子相似度;基于实例的机器翻译; 多策略机器翻译;泛化匹配

中图分类号:TP391 **文献标识码**:A

Approach of Large-Scale Sentence Similarity Computation

HUANG He-yan, CHEN Zhao-xiong, ZHANG Xiao-fe, ZHANG Ke-Liang[1,2]

(1. Research Center of Computer & Language Information Engineering, Chinese Academy of Science, Beijing 100083, China;
2. Nanjing University of science & Technology, Nanjing, Jiangsu 210094, China)

Abstract: The retrieval of the similar translation examples corresponding to the SL sentence from the large-scale corpora, or the computation of sentence similarity, is one of the key problems of EBMT. A new multi-layer sentence similarity computation approach is proposed in this paper. First, a few candidate translation examples are selected form a large-scale corpus on the basis of the surface features and entropies of the given words. Second, the degree of generalization match between the input sentence and each of those candidate translation examples is computed respectively. Finally, the sentence similarity is computed according to the outcomes of the previous two steps. Experimental results from tests on IHSMTS show that this approach has a recall rate of 96% and a precision rate of 90% when applied to a corpus of 200,000 English-Chinese sentence pairs.

Key words: sentence similarity; example-based machine translation; hybrid-strategy machine translation; generalization matching

1 引言

基于实例的机器翻译 EBMT(Example-based machine translation)的基本思路是:预先构造由双语对照的翻译单元对组成的语料库,然后翻译过程选择一个搜索和匹配算法,在语料库中寻找最优匹配单元对,最后根据例句的译文构造出当前所翻译单元的译文[1]。如何根据源语言文本找出其最相近的翻译实例,是基于实例翻译方法的关键问题之一。尤其是实用的 EB-

收稿日期:2005 -11 -05 定稿日期:2006 -01 -10

基金项目:国家自然科学基金资助项目(60502048,60272088);国家 863 计划资助项目(2002AA117010 -02)

作者简介:黄河燕(1963—),女,研究员,博士生导师,主要研究方向为自然语言处理与机器翻译、大型智能应用系统.

MT 系统所需要的翻译实例库都非常大,一般在百万级乃至千万级双语句对以上[2]。因此,如何从这么大的一个语料库中高效地计算出相似的翻译实例,提供给后面的双语词对齐、类比翻译处理等模块,是影响 EBMT 系统翻译能否成功的关键因素之一。因为得不到有效的相似实例,其结果只有一个:导致 EBMT 翻译失败(或生成的译文质量很差)。

目前计算句子相似度的方法主要有:基于 N 元模型的方法[3,4]和基于编辑距离的方法[5]等,并且在这些方面的研究也取得了许多进展。但是,这些方法主要是针对机器翻译系统的评测,一是评测时要求处理的语料都比较小,而进行 EBMT 翻译时需要处理大规模语料,这些方法难以胜任。二是这些方法几乎没有使用任何语法、语义知识,不能有效地融合翻译系统其他模块相关的处理结果和处理方法,最终效果难以提升。本文针对这些问题,提出一种多层次句子相似度计算的新方法:首先基于句子的词表层特征和信息熵从大规模语料库中选择出少量候选实例,然后针对这些候选实例进行泛化匹配,从而计算出相似句子。

论文其余部分安排如下:第二部分将详细讨论本文提出的多层次句子相似度计算方法;第三部分给出本文算法在多策略机器翻译系统[6] IHSMTS 上的实验结果及数据分析;第四部分是对本文算法的一个简短总结和下一步研究的设想。

2 多层次句子相似度计算

2.1 基于词表层特征和信息熵的候选实例检索

候选实例检索要解决的问题是,如何高效快速地从大规模语料库中选出少量句子以进行精确地句子相似度计算。因此,候选实例的检索需要考虑以下一些方面:

1、候选实例检索算法的设计,首先也是最重要的应该是能把最相似的、最有利于类比翻译的实例检索出来。因为如果检索不到相似实例或检索出来的实例相似性过低,都会导致类比翻译的失败。

2、检索出来的候选实例数量要适当。候选实例太少很容易遗漏最相似的翻译实例,导致翻译失败或译文质量不高;候选实例太多,则会占用过多的计算资源,导致系统性能严重下降。根据我们的经验,理想的候选实例应该在 5 个左右。

3、到目前为止,研究人员还没有找到一种简单通用的方法来计算句子之间的相似度。因此,候选实例的检索策略还需要与具体系统的后续处理方法和处理过程通盘考虑,以取得整个系统的最优化。

在处理过程中,我们把句子表示成单词的集合。

定义 1:句子的词集合表示为

$$\pi(S) = \{W_1, W_2, \ldots W_n\} \tag{1}$$

其中 S 表示句子,W_i 为句子中的单词。对于英语,词 W_i 需要事先进行形态还原;对于中文,句子 S 需要事先进行词切分处理。

定义 2:句子 S_1 和 句子 S_2 的表层相似度:

$$Sim_s(S_1, S_2) = 2 * \Gamma(\pi(S_1) I \pi(S_2)) / (\mathrm{Len}(S_1) + \mathrm{Len}(S_2)) \tag{2}$$

其中 I 表示集合的求交运算。Γ 运算符表示求集合中的元素个数,Len 表示句子的长度,即句子中含有的单词数。

两个句子的表层相似度越大,则输入的待翻译句子与翻译实例相同的单词就越多,后续类比译文构造过程对翻译实例所要做的修改量也就越少。这说明表层相似度的计算方法从总体上是符合 EBMT 系统的要求,即有利于最终生成高质量的译文。

定义 3:词信息熵

$$H(w) = \lg(M/m) \tag{3}$$

其中 w 表示词,M 表示语料库中的句子总数,m 表示出现了词的 w 句子数。词的信息熵值越大,说明该词在语料库中的出现频度越低,对区分句子的作用也就越大。

定义 4:句子 S_1 和句子 S_2 的信息熵相似度:

$$Sim_H = \sum H(w_i) \tag{4}$$

其中 $w_i \in \{\pi(S_1) I \pi(S_2)\}$,运算符 π 和 I 的含义参见前面定义 1 和定义 2。

两个句子的信息熵相似度越大,则从概率上来讲,输入的待翻译句子与翻译实例在语义上更相似。同时通过信息熵的计算方法,对一些特别常用的词比如{the、a、and、of}等起到了抑制作用。

进行候选实例检索时,首先根据(2)式的表层相似度计算方法,从大规模语料库中选出一定数量的句子,比如 m 个句子,然后根据(4)式的信息熵相似度计算方法,再从这 m 个句子中选出 n 个句子。实验中我们设定 $m=20$,$n=5$。要说明的是,我们没有在整个语料库中直接利用(4)式信息熵的大小来筛选候选实例,这是因为如果在整个语料库中直接利用信息熵的大小来筛选候选模式,则会给一些非常用词以过大的比重,比如在我们的统计中仅出现一次的词(比如单词 borax)其信息熵是最常用词{the}的 16.8 倍,结果会导致选出来的翻译实例在句子结构上与输入的待翻译句子相差很大,不利于后续的类比译文构造。

2.2 基于泛化的匹配度计算

基于泛化的匹配度计算,指的是在泛化的基础上计算候选实例与输入的待翻译句子间的模糊匹配度。

2.2.1 泛化

我们看下面这个例子:

翻译实例:I'll look in my diary to see if I'm free next Wednesday.

输入句子:I'll look in my diary to see if I'm free next Friday.

输入句子中是"Friday",而翻译实例中是"Wednesday",如果基于实例模式精确匹配,则输入句子没法翻译。但如果把翻译实例经词法分析泛化成下例模式:

I < NP > will < AUX > look < VP > in < PROP > my < T > diary < NP > to < PROP > see < VP > if < WH > I < NP > am < BE > free < AP > next < AP > X < TIM >.

其中 X 表示短语变量,X < TIM > 表示一个类别属性为 TIM 的短语变量,它可以跟任何一个类别属性为 TIM 的短语匹配。为了方便说明泛化的意义,这里属性只使用了词类这一个参量。在实际系统中除了词类属性外,还应该使用词形、词汇等价类、词的同义、反义、上下位、蕴含以及语境语用(上下文)等信息。

对输入句子"I'll look in my diary to see if I'm free next Friday."做同样的词法分析处理,结果如下:

I < NP > will < AUX > look < VP > in < PROP > my < T > diary < NP > to < PROP > see < VP > if < WH > I < NP > am < BE > free < AP > next < AP > Friday < TIM >.

比较泛化实例和输入句子的词法分析结果,Friday < TIM > 与 X < TIM > 能匹配上,从而可以根据该翻译实例类比推理构造出输入句子的译文,即在本例中,将 Wednesday < Tran:星期三 > 换成 Friday < Tran:星期五 > 即可。

由此可见,首先根据待翻译的输入句子对翻译实例的相关语法单位进行泛化,即形成具有

一定复杂特征的变量;再根据泛化实例类比推理构造出输入句子的译文。在这里,类比推理实际上就是一个变量属性约束匹配的过程;而译文构造主要就是通过对泛化实例进行替换、复制、删除和插入等操作来完成。

2.2.2 泛化匹配度

进行泛化匹配时,我们综合考虑了词形、词类、词汇等价类、词的同义、反义、上下位、蕴含以及语境语用(上下文)等信息。

定义 5:词汇泛化匹配度 LGD(Lexical Generalization Matching Degree):表示输入句子中的某个词汇与翻译实例中的某个词汇可以相互替换的可能性。它实际上跟词汇相似度、词汇在句子中的词性以及词汇的上下文信息有关。*LGMD* 由下式计算:

$$LGMD(w_1,w_2) = f(Sim_{Lex},Sim_{Pos},Sim_{Con})$$

$$= \begin{cases} \alpha \cdot Sim_{Lex} \cdot Sim_{Pos} \cdot Sim_{Con}, & if\ Sim_{Lex} \neq 0 \& Sim_{Pos} \neq 0 \\ \beta \cdot Sim_{Lex} \cdot Sim_{Con}, & if\ Sim_{Lex} \neq 0 \& Sim_{Pos} = 0 \\ \gamma \cdot Sim_{Pos} \cdot Sim_{Con}, & if\ Sim_{Lex} = 0 \& Sim_{Pos} \neq 0 \\ 0, & if\ Sim_{Lex} = 0 \& Sim_{Pos} = 0 \end{cases} \tag{5}$$

上式中 α,β,γ 是三个系数,表示各种情况的可信度权值。实验中我们设 $\alpha=1,\beta=0.9,\gamma=0.8$。$Sim_{Lex}$ 表示词汇相似度,Sim_{Pos} 表示词性相似度,Sim_{Con} 表示语境相似度。其中 Sim_{Lex} 的计算式如下:

$$Sim_{Lex}(w_1,w_2) = \begin{cases} 1, if\ w_1 = w_2 \text{ 或 } w_1,w_2 \in \text{同一词汇等价类} \\ \dfrac{\alpha}{dis\_sem(w_1,w_2)}, if\ w_1 \neq w_2 \end{cases} \tag{6}$$

上式中 $dis\_sem(w_1,w_2)$ 表示词汇 w_1,w_2 的语义距离,α 是权值系数,实验中我们没 $\alpha=0.8$。语义距离的计算我们采用了基于 HowNet 的方法。HowNet 提供的义原分类树把各个义原以及它们之间的关系以树的形式组织在一起,树中父节点和子节点的义原具有上下位关系。因此,可以利用义原分类树来计算两个词汇之间的语义距离。计算时,我们把语义距离定义为两个词对应的义原在义原分类树中与最近邻共同祖先节点间距离的平均值。

Sim_{Pos} 的计算式如下:

$$Sim_{Pos}(w_1,w_2) = \begin{cases} 1, if\ Pos(w_1) = Pos(w_2) \\ 0, if\ Pos(w_1) \neq Pos(w_2) \end{cases} \tag{7}$$

上式中 $Pos(w)$ 表示词汇 w 在句子中的词类标注属性。

Sim_{Con} 的计算式如下:

$$Sim_{Con}(w_1,w_2) = \begin{cases} 1, if\ dis\_con(w_1,w_2) = 0 \\ \dfrac{\omega}{\sqrt{dis\_con(w_1,w_2)}} \end{cases} \tag{8}$$

上式中 ω 也是一个权值系数,实验中我们取值 $\omega=0.8$。$dis\_con(w_1,w_2)$ 表示词汇 w_1,w_2 的上下文偏移距离,它的值主要是在同语词对齐的基础上通过观察当前词左右 $N/2$ 个词(即宽度为 N 的词汇窗口)来决定。

定义 6:句子泛化匹配度 SGMD(Sentence Generalization Matching Degree):表示该翻译实例作为范例对输入句子进行类比翻译的可信度。*SGMD* 由下式计算:

$$SGMD(s_1,s_2) = \frac{2 \cdot \sum_{i=0}^{i<Len(s_1)} \underset{0 \leqslant j < Len(s_2)}{\operatorname{argmax}} LGMD(w_i,w_j)\}}{Len(s_1) + Len(s_2)} \tag{9}$$

上式分母中 $Len(s_1)$,$Len(s_2)$分别表示输入句子和翻译实例的句子长度。

2.3 句子相似度计算

最后的句子相似度由下式计算:

$$Similarity(s_1,s_2) = \alpha \cdot SGMD(s_1,s_2) + \beta \cdot Sim_s(s_1,s_2) + \gamma \cdot Sim_H \tag{10}$$

其中 α,β,γ 是三个权值系数,分别表示泛化匹配度、表层相似度和信息熵相似度的权值,并且 $\alpha+\beta+\gamma=1$。实际设置时,由于泛化匹配度能比较全面的从词汇、语法、语义和上下文等多方面考察句子的相似性,所以 α 设置的相对大一些,又因为信息熵相似度是在表层相似度基础之上计算出来的,所以 γ 设置的相对小一些。

3 实验设计与结果分析

EBMT 系统的句子相似度计算,目的就是为了能从大规模语料库中选择出最相似的翻译实例,供后续模块进行类比译文构造。为了比较全面地评估本文算法,我们使用了准确率、召回率和 F 值等三个指标,它们的计算式分别定义如下:

$$\text{准确率} = \frac{\text{能正确找出相似句子的总数}}{\text{测试句子的总数}} \times 100\% \tag{11}$$

$$\text{召回率} = \frac{\text{能正确找出相似句子的总数}}{\text{存在标准相似句的总数}} \times 100\% \tag{12}$$

$$F\_score = \frac{2(\text{准确率} \times \text{召回率})}{\text{准确率} + \text{回率}} \tag{13}$$

开始测试前,我们首先向系统中导入 20 万英汉句对;中文和英文大约各有 200 多万词,其中英文平均句长为 12.5 个词左右,中文平均句长为 11 个词左右。然后从这 20 万英汉句对随机挑选出 100 个中文句子,再对这 100 个中文句子分别进行人工修改,最终形成 400 个不同的句子,作为测试集。

测试时,我们逐句地把测试集中的句子输入系统,系统返回相似度大于 0.75 的所有翻译实例,并且返回的每个翻译实例都附有机器自动计算出来的相似度。然后对返回的每个翻译实例进行人工判别,人工判别的依据是:①能否把相似句子提取出来,这一项主要反映在召回率上;②机器自动计算出来的相似度与人工主观判断的拟合性,这一项主要反映在准确率上。试验结果如表 1 所示。

从表 1 中的试验结果可以看出,准确率达 90% 说明机器自动计算出来的相似度与人工的主观判断是很接近的;召回率达 96% 说明算法能够从大规模语料库中比较有效地检索出相似实例。

表 1 相似度计算测试结果

语料库规模(句)	测试集规模(句)	准确率(%)	召回率(%)	F
200,000	400	90	96	93

另外,我们重点分析了影响准确率和召回率的因素:

最主要因素之一就是分词标准的不一致,比如在我们的分词系统中把“看电影”切分成一个词,而把“看电视”切分成“看 电视”两个词。而且这种现象比较多、比较严重,如果能解决这类问题,试验效果应该要更好的多。

另外一个主要因素就是词类标注不准确。词类标注错误也会影响相似度计算的准确性，不过由于词类信息在相似度计算中不是特别重要，所以相当分词来说影响较小。

4 结束语和进一步的研究

首先基于句子的词表层特征和信息熵从大规模语料库中选择出少量候选实例，然后针对这些候选实例进行泛化匹配，从而计算出相似句子。在多策略机器翻译系统 IHSMTS 中的实验表明，语料规模为 20 万英汉句对的情况下，系统提取相似句子的召回率达 96%，准确率达 90%，充分说明了本文所提出的这种多层次句子相似度计算方法的有效性。

此外，在 EBMT 系统中，句子相似度计算的最终目的就是为了能够使后续模块有效的进行类比译文构造，产生高质量的译文。但在本文的算法中，计算句子相似度时没有考虑翻译实例的译文部分。下一步的研究中，我们准备在计算句子相似度时，把翻译实例的对译信息（如双语词对齐结果）考虑进来，以便更全面的考察待翻译句子与翻译实例的相似性。

参 考 文 献：

[1] H. Maruyama and H. Watanabe. Tree Cover Search Algorithm for Example-Based Translation[A]. In: Proceeding of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation [C](TMI-92). Montreal, 1992, 173 - 184.

[2] Ralf D. Brown, Example-Based Machine Translation in the Pangloss System[A]. In: Proceedings of the 16th International Conference on Computational Linguistics[C] (COLING - 96).. Copenhagen, Denmark, August 5 - 9, 1996, 169 - 174.

[3] Keiji Yasuda, Fumiali Suagya, etc, An Automatic Evaluation Method of Translation Quality Using Translation Answer Candidates Queried from a Paralledl Corpus[A]. In: Proceeding of MT Summit's conference[C]. Santiago de Compostela, 2001.

[4] Jianmin Yao, Ming Zhou etc, An Automatic Evaluation Method for Localization Oriented Lexicalised EBMT System[A]. In:Proceeding of the 19th International Confernce on Computational Linguistics[C](COLING2002). Taipei, 2002.

[5] Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita, Using Multiple Edit Distances to Automatically Rank Machine Translation Output[A]. In:Proceeding of MT Summit's conference[C]. Santiago de Compostela, 2001.

[6] 黄河燕，陈肇雄. 基于多策略的交互式智能辅助翻译平台总体设计[A]. 见：黄河燕主编，机器翻译研究进展(2002 年全国机器翻译研讨会论文集)[C]. 北京：电子工业出版社，2002 年 11 月，137 - 146.