基于改进互信息的译文选择技术研究

林晓庆,徐惠红
(辽东学院 信息技术学院,辽宁 丹东 118003)

摘 要:提出了一种改进互信息的译文选择方法,认为词语的译文的选择不是孤立进行的,上下文对译文的选择有着重要的意义,通过对已有的互信息公式加入翻译模型特征进行改进,结合翻译模型与互信息来选择最佳译文,经过 BLEU(BiLingual Evaluation Understudy)作为机器评价准则的实验结果表明,该方法优于传统的互信息词语译文选择的方法。

关键词:互信息;译文选择;翻译模型;译文选择模型
中图分类号:TP391 **文献标识码**:A

Research of translation selection based on improved mutual information

LIN Xiao Qing, XU Hui Hong
(Institute of Information Technology, Eastern Liaoning University, Dandong 118003, China)

Abstract: A method of translation selection based on improved mutual information is proposed to select the best word translation. The selection of words in the translation is not independently carried out. The context is helpful for correctly translating words in the context. On the basis of the improvement of the characteristics of the existing mutual information formula, the best translation can be selected by combining the translation model with mutual information. Experimental results show that our method outperforms a baseline pharaoh by using BLEU evaluation system.

Key words: mutual information; translation selection; translation model; translation selection model

译文选择是指根据从语料库中学习翻译知识,为源语言词选择对应的目标语言词。词译文选择的好坏决定了机器翻译系统的质量。Gale 等人[1]应用基于大型英法对齐语料库的统计方法,对 6 个常见的歧义词的消歧正确率在 82%~86%。刘小虎建立多上下文特征的词义消歧统计模型,对歧义词“interest”消歧测试的正确率达到 80%[2];而通过在英汉机译系统的译文选择中引入改进的 ID3 机器学习方法[3],歧义词“interest”消歧测试的正确率可达到 91%,荀恩东[4]在译文选择中使用以消歧矩阵为计算背景的贪心算法。Dagan[5]等人提出利用目标语同现统计消除源语言歧义的思想。哈尔滨工业大学 BT863-2 英汉机译系统继承 Dagan 的思想,译文选择的正确率为 75%。术语相关性计算的研究比较典型,有 EMMI weighting measure[6]、Term Similarity[7-9],本文方法与参考文献[10]中提出的查询翻译中用到的方法有些相似。

1 译文选择模型

Ballesteros 和 Croft[8]认为对语料库进行共现频率的统计有助于消除翻译的歧义问题。他们假定正确的翻译更可能在同一个目标句子中共现,否则相反。参考文献[7-9]也使用相类似的方法选择最佳的词语翻译。

正是因为各个词之间的关系不是相互独立的,本文提出词语相关性和翻译概率相结合的方法来选择相应的词语翻译,而不是逐词孤立地翻译。当翻译一个词语时,其他待翻译词的候选翻译会成为它的上下文信息,这是本文进行翻译选择的原则。给定一个待翻译的英文词语的集合,通过贪心算法和下文中的公式(5)找到每个词的正确译文。

例如,输入 NP (Noun Phrase):IC card intelligent door lock。

在本文的双语词典中,“intelligent” 对应的翻译候选有:(1) 智能国;(2) 智力。依次类推本例中的目标集合 T 为{“IC”,“卡”,“门”,“通道”,“锁”,“锁头”}。目标集合的获得是通过在双语词典中查找每个源语言词对应的汉语翻译候选组成的集合。通过公式(1)~(3)[11]计算,找到最可能的目标翻译,上例计算得到的翻译结果为“IC 卡 智能 门 锁”。

$$Correlation(x,y)=p(x,y)\times\log_2\left(\frac{p(x,y)}{p(x)\times p(y)}\right)-K\times\log_2 dis(x,y) \quad (1)$$

$$p(x,y)=\frac{c(x,y)}{c(x)}+\frac{c(x,y)}{c(y)} \quad (2)$$

$$p(x)=\frac{c(x)}{\sum_x c(x)} \quad (3)$$

$$P(c/e)=\frac{count(e,c)}{\sum_e count(e,c)} \quad (4)$$

式中 $c(x,y)$是术语 x 和 y 在双语例句中出现在目标 NP 中的次数，$c(x)$是 x 在双语例句库中出现的次数，$dis(x,y)$是 x，y 的平均距离，x 和 y 的距离定义为 x，y 之间的字符数。例如："IPv6 多域 分类 处理 方法"。其中"IPv6"与"处理"的距离为 2。(K 的值在 0.7~0.9 之间，在实验中当取 K=0.8 时为最佳的变量参数)。

$P(c/e)$是翻译概率，$count(e,c)$是双语例句库中 e 翻译成 c 的次数，$\sum_e count(e,c)$是 e 在例句库中出现的次数。翻译概率通过面向专利文献的双语 NP 语料训练得到，该语料包含 400 000 英汉 NP 对。词语 x 与其他词语构成的集合 X 的相关性定义为这个词语同集合中其他词语相关性最大值与该词语 x 的翻译概率的乘积。如公式(5)所示。

$$Cohesion(x,X)=P(x/e)\times Max_{x\in y}Correlation(x,y) \quad (5)$$

具体算法如图 1 所示。

```
For 每个源词 Si (i=1 to n, n 为 NP 的长度)，从双语词表中找到目标词的集合 Ti ( =1, to n) ;
    For 每个目标词 tij in Ti, do
        计算 Cij =Cohesion( tij  Tk) (k=1 to n && k!=i);
    Endfor
        选择在 Ti 中拥有 Cohesion(tij, Tk)最高得分术语词 tij，并且把该选择的词放到目标集合 T 中
Endfor
```

图 1 贪心算法实现词语翻译的查找

2 实验结果及分析

本文将翻译概率加入到公式(1)中，结合翻译概率与互信息来进行译文的选择，对比实验结果可知，翻译概率对翻译结果有较大的提高。

为了充分证明该结果，从英汉术语实例库中，随机挑选 500 个实例进行对比测试，采用 NIST 发布的最新版本 mteval-v11b.pl 作为自动翻译结果的评测工具，实验结果的曲线图如图 2 所示。

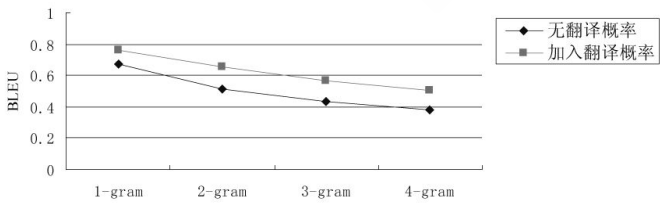

图 2 翻译概率对实验结果影响的对比

从表 1 中可以看出，加入翻译概率后，从 1-gram 到 4-gram 的 BLEU 值都有所提高。为了更加清楚地显示其对比效果，可以参见图 2。

表 1 对比实验结果

	1-gram	2-gram	3-gram	4-gram
无翻译概率	0.672 7	0.512 9	0.435 0	0.381 3
加入翻译概率	0.762 4	0.657 4	0.570 3	0.500 1

举一具体实例来说明上面原因。例如：输入 NP：Safety non-tipping mosquito incense device，在不加入翻译概率时，只通过公式(1)计算得出翻译结果为："安全不倒蚊蚊扣掣座"。

分析其原因，从表 2 可知，在没有加入翻译概率之前，通过公式(2)计算，"incense"选择了"蚊"这个译文，因为"蚊"的值最大，如表 3 所示。在加入翻译概率改进之后，通过公式(5)计算，结果如表 2 所示，由于其翻译概率很小，因此就会选择到更合适的译文"香"。("#"表示选择的译文)根据表 4，正确的译文为："安全 不倒 蚊 香 器"。

表 2 没加入翻译概率的译文选择表

# 安全 29.574 6	# 不倒 34.005 8	# 蚊 34.005 8	燃香 29.574	装置 1.615 92
安全型 1.583 99	–	蝇 6.204 86	# 蚊 34.005 8	器 26.847 9
安全性 1.829 8	–	–	香 33.009 9	设备 0.417 73
方便 0.997 499	–	–	蚊香 2.935 5	机 1.211 4
–	–	–	–	器件 0.896 067
–	–	–	–	具 0.621 699
–	–	–	–	器具 0.492 077
–	–	–	–	# 扣掣座 29.574 6

表 3 翻译概率表

Safety	non-tipping	mosquito	incense	device
安全 0.976 9	不倒 1	蚊 0.875	燃香 0.071 4	装置 0.609 8
安全型 0.002 42	–	蝇 0.125	蚊 0.071 4	器 0.281 843
安全性 0.002 42	–	–	香 0.571 4	设备 0.019 0
方便 0.001 21	–	–	蚊香 0.214 3	机 0.013 5
–	–	–	–	器件 0.010 8
–	–	–	–	具 0.008 13
–	–	–	–	器具 0.005 42
–	–	–	–	扣掣座 0.002 71

表 4 加入翻译概率后的译文选择表

# 安全 28.892 6	# 不倒 34.005 8	# 蚊 29.755 1	燃香 2.112 47	装置 0.001 961 1
安全型 0.003 844 7	–	蝇 1.748 79	蚊 2.428 99	器件 0.009 713 47
安全性 0.004 441 3	–	–	# 香 18.862 8	机 0.016 414 6
方便 0.281 138	–	–	蚊香 0.629 035	设备 0.007 924 41
–	–	–	–	扣掣座 0.080 147 9
–	–	–	–	器具 0.002 667 09
–	–	–	–	具 0.005 054 46
–	–	–	–	# 器 7.566 89

译文选择的好坏是机器翻译质量提高的关键。本文提出的改进互信息的译文选择方法,其中对互信息的理论作了简单介绍,对译文选择的相关研究也进行了简单描述。通过对比实验分析证明了该方法在已有的互信息方法上加入翻译模型特征后,翻译效果得到显著地提高,BLEU 值提高了 0.1 左右。

参考文献

[1] WILLIAM G, KENNETH C, DAVID Y. Using bilingual materials to develop word sense disambiguation methods[C]. The 4th Int'l Conf on Theoretical and Methodological Issues in Machine Translation, Montreal, Canada, 1992.

[2] LIU Xiao Hu, Li Sheng , Zhao Tie Jun . Statistical model selection for word sense disambiguation (in Chinse) [J]. Communications of Chinese and Oriental Languages Information Processing Society, 1997, 7(2): 69–75.

[3] 刘小虎. 英汉机器翻译中词义消歧的研究[M]. 哈尔滨:哈尔滨工业大学, 1997.

[4] 荀恩东, 李生, 赵铁军. 基于汉语二元同现的统计词义消歧方法研究[J].高技术通讯, 1998, 10(8): 21–25.

[5] DAGAN, LILLIAN L, FERNANDO P. Similarity–based models of cooccurrence probabilities [J]. Machine Learning, Special Issue on Natural Language Learning, 1999, 34(1–3): 43–69.

[6] RIJSBERGEN V . Information retrieval[J]. 2nd ed. Butterworths, London, 1979.

[7] ADRIANI M. Using statistical term similarity for sense disambiguation in cross –language information Retrieval [C]. Information Retrieval, 2000, 2: 69–80.

[8] BALLESTEROS L, CROFT W B Resolving ambiguity for cross–language retrieval [C]. In Proceedings of the 21st International Conference on Research and Development in Information Retrieval,1998.

[9] BALLESTEROS L , CROFT W B. Phrasal translation and query expansion techniques for cross–language information retrieval[C]. In: Proceedings of the 20th International Conference on Research and Development in Information Retrieval, 1997: 84–91.

[10] GAO J F , NIE J Y. A study of statistical models for query translation:finding a good unit of translation[C]. In SIGIR, 2006.

[11] GAO Jian Feng, NIE Jian Yun, ZHANG Jian, et al. Improving query translation for cross–language information retrieval using statistical models [C]. In SIGIR'01, NewOrleans, Louisiana, 2001: 96–104.

(收稿日期:2009–10–20)

作者简介

林晓庆,女,1979 年生,硕士,讲师,主要研究方向:机器翻译、术语翻译;

徐惠红,女,1974 年生,硕士,讲师,主要研究方向:模式识别。

(上接第 67 页)

[4] 古天龙.软件开发的形式化方法[M].北京:高等教育出版社, 2005.

[5] SANUEL I. C# LEX Manual.[2009–10–16].http://www.seclab.tuwien.ac.at/projects/cuplex//lex.htm,2003.

[6] SAMUEL IMRISKA. C# CUP Manual.[2009–10–16].http://www.seclab.tuwien.ac.at/projects/cuplex//cup.htm,2005.

(收稿日期:2009–10–16)

作者简介

王甲,男,1984 年生,硕士研究生,研究方向:网络软件与数据库。

康慕宁,男,1955 年生,教授,研究方向:网络软件与数据库、软件形式化。