面向口语翻译的双语语块自动识别

程 葳[1),2)] 赵 军[1)] 刘非凡[1)] 徐 波[1)]

[1)]中国科学院自动化研究所模式识别国家重点实验室 北京 100080)

[2)]北京城市学院人工智能研究所 北京 100083)

摘 要 语块识别是实现"基于语块处理方法"的基础.目前,针对单语语块的研究成果已有很多,但机器翻译更需要双语相关的语块分析.该文根据口语翻译的实际需要,提出了"双语语块"的概念.并在此基础上,实现了一种针对并行语料库进行双语语块自动识别的新方法.该方法将统计和规则相结合,可同时保证双语语块的语义特性和句法规范.通过在一个 6 万句的旅馆预定领域口语语料库中的实验可以看出,该方法对汉英并行语料的双语语块识别正确率可达到 80%左右.

关键词 语块;语块分析;语料库;口语翻译

中图法分类号 TP391

Automatic Identification of Co-Chunks for Spoken-language Translation

CHENG Wei[1),2)] ZHAO Jun[1)] LIU Fei-Fan[1)] XU Bo[1)]

[1)]*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080)

[2)]*Institute of Artificial Intelligence, Beijing City University, Beijing* 100083)

Abstract Chunk parsing is a basic step for the chunk-based processing. There have been many chunk parsing methods for single languages. However chunk parsing for bilingual language is specially needed in the machine translation. The paper presents the idea of co-chunks which are defined according to the characteristics of both Chinese and English. A new algorithm is also proposed to automatically identify the co-chunks in the sentence-aligned bilingual corpus. It combines rules into statistical model, which assure that the co-chunks identified have both legal syntactical structure and semantical explanation. The algorithm is trained in a sentence-aligned Chinese-English bilingual corpus with the size of about sixty thousand sentence pairs. This corpus consists of spontaneous utterances from hotel reservation dialogs. The experiments show that the accuracy of the method is above 80%.

Keywords chunk; chunk parsing; corpora; spoken-language translation

1 引 言

语块分析(chunk parsing),又称为浅层句法分析或部分句法分析,是近年来自然语言处理领域出现的一种新的语言处理策略.它与完全句法分析相对,不要求得到完整的句法分析树,只需要识别其中某些结构相对简单的成分——语块(chunk).该过程包括:(1)语块的识别和分析;(2)语块之间的依附关系.其中对语块的识别是语块分析的主要任务[1].

收稿日期:2003-04-16;修改稿收到日期:2004-04-13.本课题得到国家自然科学基金(60272041,60121302)资助.程 葳,女,1973 年生,博士研究生,主要研究领域为口语机器翻译和自然语言处理.E-mail:wcheng@nlpr.ia.ac.cn.赵 军,男,1966 年生,博士,副研究员,主要研究方向为自然语言理解、智能信息处理和知识工程.刘非凡,男,1979 年生,博士研究生,主要研究方向为计算语言学和信息提取.徐 波,男,1966 年生,研究员,博士生导师,主要研究方向为语音识别和语音翻译.

目前, 针对单语的语块分析工作已取得大量成果[2~4]. 不过, 在翻译当中, 我们更希望得到双语对应的语块识别. 有关这方面的工作, 按输出结果可分为两类:

(1)确定双语语料库中每个句对内语块和语块间的对齐关系.

(2)建立双语语块库. 其所用到的方法根据研究对象的不同, 又可分为: (i)从两个不同语言的单语语块库中抽取语义对应的双语语块[5]; (ii)从双语并行语料库中抽取翻译短语对[6].

根据笔者查阅的文献资料, 当前绝大多数研究成果, 主要是集中在"语块库建立"的问题上. 针对对齐的"双语语块识别"则研究较少. 因此, 探索合理、高效的双语语块自动识别方法, 对语料库语言学和机器翻译的发展都具有重要意义.

综上所述, 本文提出了一种用于语块对齐的双语语块自动识别新方法. 该方法以句子级对齐的并行语料为研究对象, 根据口语翻译的实际需要, 将统计与规则相结合, 实现了双语同步的语块分析. 以下是文章的主要内容: 第 2 节着重介绍了面向口语翻译的双语语块的基本概念和分析时用到的一些假设; 第 3 节则对自动识别的原理和打分算法进行了详细论述, 并给出了系统的结构框图和公式推导; 第 4 节应用一个限定领域内的汉英口语语料库对该方法进行实验, 其结果表明, 自动识别的正确率已达到 80%. 最后, 本文对方法中一些尚待研究的问题给予了讨论.

2　双语语块的基本概念

2.1　双语语块的定义

语块概念的最经典描述是在文献[7]中提出的. 他根据心理语言学理论, 将具有单一语义核心和严格非递归句法结构的单词或连续词串定义为 chunk (语块). 之后, 学者们将这一概念引入到中文, 提出了各种有关中文的语块定义. 文献[5]将单语语块概念进行了拓展, 提出一种具有语义自足性和转换充分性的无歧义、可嵌套翻译单元——双语 E-Chunk (Extended-Chunk). 根据上述总结, 我们对双语语块(Co-Chunk)的概念进行了如下定义[8]:

(1)结构. 双语语块包括源语子块和目标语子块两部分, 其形式化表示为

$$BC=\{\langle bs, bt\rangle \mid bs=ws_0, \cdots, ws_l, bt=wt_0, \cdots, wt_m;\ bs \leftrightarrow bt;\ ws_i \in ws_0^l,\ wt_j \in wt_0^m;\ l \in [0, NS],\ m \in [0, NT]\} \quad (1)$$

式中, BC 为双语语块集合; bs 和 bt 分别为源语子块和目标语子块; ws 和 wt 分别为源语言单词和目标语言单词. l 和 m 为源语子块和目标语子块的长度; NS 和 NT 为源语句子和目标语句子的长度. 从中可以看出, 空集合、单词、连续词串和句子都有可能成为双语语块中的子块.

(2)句法. 为了有利于继承单语语块已有的研究成果, 我们定义双语语块中的每个子块都不相交、非嵌套, 且满足正则短语结构文法, 其句法解析树(parse-tree)是句子解析树的某一连通子图, 树根(root)代表了子块的结构类型. 应用有限状态自动机就可对其进行分析.

(3)语义. 子块中的语义核心(s-head)一般由一个或多个实词组成, 它们之间满足语义关联的局部性假设[9], 因此缓解了词汇的歧义性; 同时, 两子块间还具有翻译上的转换充分性. 也就是说, 在并行语句对中, 源语子块和目标语子块是一一对整的, 且句子完全由双语语块构成, 如图 1 所示.

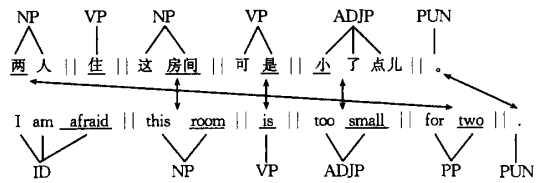

图 1　双语语块举例(图中: 双线隔开的词串构成了双语语块的子块; *NP*, *VP*, *ADJP* 等是各子块的句法标记, 其具体含义请参见文献[8]; 带下划线的单词是子块的语义核心; 带箭头的连线将单语子块对应为双语语块, 没有被连接的子块表示对空.)

2.2　双语语块识别中的假设

根据翻译的特点, 在对并行语料进行双语语块识别时, 除了要满足其基本定义以外, 我们还进行了以下假设:

(1)扩展优先假设. 为了尽可能减少双语语块的歧义性, 在进行语块分析时我们遵循扩展优先的假设. 即, 在需要对单元进行合并或拆分时, 应优先考虑合并的方式(即扩大语块).

(2)以源语为基准假设. 由于翻译系统的输入只有源语, 因此当定义的源语和目标语在句法上发生冲突时, 应尽量保证源语句法结构的合理性. 这主要包括两种情况: 一是在需要拆分或合并的情况下, 应尽量保证源语子块句法结构的完整性; 另一是尽量依据源语子块对目标语进行语块分析.

3 双语语块的自动识别方法

3.1 自动识别的统计模型

基于统计的双语语块识别过程,可被描述为:

已知:源语言句子 S: ws_1, ws_2, …, ws_{NS},目标语言句子 T: wt_1, wt_2, …, wt_{NT}. NS 和 NT 分别为 S 和 T 的长度.

求取:双语语块序列集合 $Q=\{\langle bs_k, bt_k\rangle \mid bs_k \in S, bt_k \in T, k \in [0, K]$ 中的 Q^*,使得概率 $P(Q^*)$最小,即

$$Q^* = \arg\min_Q\{-\log P(Q)\} = \arg\min_Q\left\{\sum_{k=1}^{K} -\log p(bt_k \mid bs_k)\right\} \quad (2)$$

式中, bs_k为源语子块, $S = bs_1 \cdots bs_K$; bt_k为目标语子块, $T = bt_1 \cdots bt_K$; K 为双语语块的个数.由贝叶斯公式可知,

$$p(bt_k \mid bs_k) = \frac{p(bt_k) \cdot p(bs_k \mid bt_k)}{p(bs_k)} \quad (3)$$

其中, $p(bs_k)$和 $p(bt_k)$分别为源语子块和目标语子块的语言模型,可由二元模型(bigram)参数估计.

$$p(bs_k) = \prod_{j=1}^{m_k} p(ws_j \mid ws_{j-1}) \quad (4)$$

$$p(bt_k) = \prod_{i=1}^{l_k} p(wt_i \mid wt_{i-1}) \quad (5)$$

$p(bs_k \mid bt_k)$为从源语子块到目标语子块的翻译转换模型,可由 IBM 的模型估计[10].设 m_k和 l_k分别为子块 bs_k和 bt_k的长度,则

$$p(bs_k \mid bt_k) = p(l_k \mid m_k) \cdot \prod_{j=1}^{m_k} \sum_{i=1}^{l_k} p(ws_j \mid wt_i) \quad (6)$$

参数 $p(ws_j \mid wt_i)$是从目标语言到源语言的单词直译概率,可根据 EM 算法进行估计[10];参数 $p(l_k \mid m_k)$是子块间的长度概率,可根据泊松分布进行估计.

将式(3)~(6)代入式(2)中,可得

$$\log P(Q) = \sum_k \Big[\sum_{i=1}^{l_k} \log p(wt_i \mid wt_{i-1}) - \sum_{j=1}^{m_k} \log p(ws_j \mid ws_{j-1}) + \log p(l_k \mid m_k) + \sum_{j=1}^{m_k} \log \sum_{i=1}^{l_k} p(ws_j \mid wt_i)\Big] \quad (7)$$

基于统计的双语语块识别方法的实质就是:通过对输入并行语句对的搜索,找到使打分 $\log P(Q)$最小的双语语块组合.

3.2 识别步骤

根据以上数学描述和假设,我们设计了一个双语语块的自动识别系统.该系统主要包括 3 个部分,如图 2 所示.

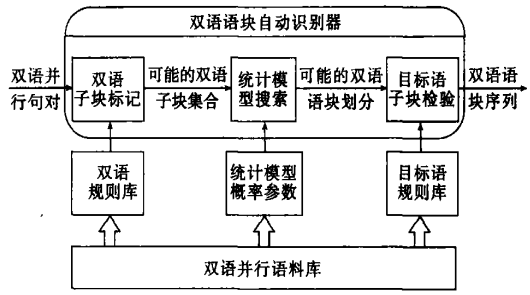

图 2 双语语块自动识别系统的结构示意图

(1)双语子块标记.根据规则库,分别对源语言句子和目标语言句子中所有符合规则的子块进行标记,并将结果存储在如图 3 所示的数组结构中.其中,数组的行和列分别表示句子中的位置 i 和 j;数组的项表示从第 i 个单词开始到第 j 个单词为止所组成的词串是否符合子块的句法规则,1 表示符合,即该词串可能是双语语块的子块,0 表示不符合,即词串不可能是子块.

子块起始位置

子块终止位置	1	2	3	…	m
1	1	0	0	…	0
		…	…	…	
m	0	0	1	…	1

图 3 子块标记数组的结构

(2)基于统计的双语语块搜索.根据统计模型,对所有可能的源语子块在可能的目标语子块集合中搜索其最佳的对应,最终形成双语句对的可能的双语语块划分.其中,空子块将被作为一个特殊的子块加入到目标语子块集合中.

(3)目标语子块检验.目标语子块进行句法检查,其中不能被分析成子块序列的划分结果将被剔除.

上述过程中,双语子块标记和目标语子块检验可由有限状态自动机完成,而双语语块搜索采用的是带启发函数的动态规划策略.

4 实验结果

4.1 实验过程

由于大多数口语翻译研究是在限定领域中进行的,因此其语料库收集也多是针对某些特殊领域.本实验采用的就是旅馆预定领域中的汉英口语并行语料库,其具体测试过程为:

(1)参数训练. 首先, 从语料库中抽取中英文对应语句 66061 对, 作为训练语料, 分别训练中文语言模型概率 $p(ws_j \mid ws_{j-1})$、英文语言模型概率 $p(wt_i \mid wt_{i-1})$、从中文到英文的直译概率 $p(ws_j \mid wt_i)$和长度概率 $p(l \mid m)$, 同时, 对于可能出现集外现象的情况, 在概率模型中加入 back-off 平滑算法; 然后, 分别从中、英文树库中提取部分短语规则, 包括中文规则 296 条和英文规则 344 条.

(2)整理参考答案. 从语料库中分别选取集内测试语料 2487 句对和集外测试语料 845 句对; 并由人工根据双语语块的定义对其进行分析; 然后将结果作为测试参考的标准答案(以下简称为答案).

(3)测试和评价. 应用双语语块自动识别系统分别对集内和集外的测试语料进行分析, 并将结果与答案比较, 计算其精确度(precision)和召回率(recall), 具体公式为

$$precision = \frac{N_r}{N_p} \times 100\%, \quad recall = \frac{N_r}{N_a} \times 100\% \tag{8}$$

其中, N_r是识别正确的双语语块个数; N_p是系统自动识别出来的双语语块个数; N_a是答案中包含的双语语块个数.

4.2　实验结果

表 1 给出了实验的结果. 其中, I 为集内测试, II 为集外测试; time 为处理每一句对的平均耗时; N-best为识别结果中前 N 个打分较高的输出. 从中可以看出, N-best 的取值对实验结果影响不大.

表 1　实验结果

测试集	N-best	Na	Np	Nr	精确度(%)	召回率(%)	时间(s/pair)
I	Top 1	25146	25305	21236	83.92	84.45	0.61
	Top 10		25358	21266	83.86	84.57	0.65
	Top 20		25384	21286	83.86	84.65	0.77
II	Top 1	8249	8083	6559	81.15	79.51	0.58
	Top 10		8218	6680	81.29	80.98	0.56
	Top 20		8248	6697	81.20	81.19	0.59

表 2 是实验中的一些实例. 其中, 双竖线为语块的边界; 斜体字为识别错误的语块; 括号内的数字为语块编号, 具有相同数字的中、英文子块共同构成一个双语语块, 编号"－1"表示该语块对应空子块.

表 2　双语语块自动识别的实例

例子	分析
麻烦 您 (4)‖ 把 预约 (3)‖ 推迟 (2)‖ 到 三 天 后 (1)‖ . please (4)‖ postpone (2)‖ my reservation (3)‖ for three days (1)‖ .	完全正确
预定 (10)‖ 是 (9)‖ 住 (8)‖ 两 个 晚上 (7)‖ , (6)‖ 但 (5)‖ 想 (4)‖ 改为 (3)‖ 住 (2)‖ 三 个 晚上 (1)‖ . ***I*** (4)‖ ***had*** (8)‖ a reservation (10)‖ ***for*** (2)‖ two nights (7)‖ , (6)‖ but (5)‖ please (－1)‖ change (3)‖ ***it*** (9)‖ to three nights (1)‖ .	词义对错
我 (7)‖ 今天 (6)‖ 订 了 房间 (5)‖ 但是 (4)‖ 突然 (3)‖ 有 了 (2)‖ 急事 (1)‖ . *I* (7)‖ have a reservation (5)‖ for tonight (6)‖ but (4)‖ due to (2)‖ urgent business (1)‖ ***I am unable*** (3)‖ to make it (－1)‖ .	对空情况判断有误

根据以上结果, 我们可以看出:

(1)本方法可以有效地对并行语料进行双语语块分析. 其集内测试的精确度和召回率已接近 85%, 集外测试的精确度和召回率也可达到 80% 左右.

(2)双语语块识别错误中的大部分是中英文子块的对应错误, 这其中主要包括语义对错和对空判断错误两种类型. 前者可能是由于算法中的概率参数不够准确, 扩大训练语料规模和加入词典信息将会有利于这方面的改进; 后者形成的原因则略微复杂, 不过, 如果能对语料库进行预处理, 提前识别出部分习语或常用语, 将会有利于减少该问题.

5　结论和将来的工作

本文提出了一种统计和规则相结合的双语语块自动识别算法, 实现了对并行语料库的双语语块自动分析. 应用该方法, 再辅以人工校对, 可以方便地获取包含双语语块信息的并行语料库和双语语块库, 这些都是自然语言处理领域中的重要资源. 目前, 该工作尚处于起步阶段, 还有很多问题需要深入探讨, 主要包括: (1)本文中的实验是在特定领域中

进行的,但方法本身实际上并没有领域限制,因此如何在非特定领域内有效应用该方法将是下一步研究的方向之一;(2)本文的测试语料十分有限,还需通过在大规模语料库中的进一步实践来发现问题;(3)对算法的改进将从加入词典等更多语言学信息和采用更为有效的句法分析策略两方面进行.

参 考 文 献

1 孙宏林, 俞士汶. 浅层句法分析方法概述. 当代语言学, 2000, 2(2): 74~83

2 Zhou Qiang, Sun Mao-Song, Huang Chang-Ning. Chunk parsing scheme for Chinese sentences. Chinese Journal of Computers, 1999, 22(11): 1158~1165(in Chinese)
(周 强, 孙茂松, 黄昌宁. 汉语句子的组块分析体系. 计算机学报, 1999, 22(11): 1158~1165)

3 Argamon S., Dagan I., Krymolowski Y.. A memory-based approach to learning shallow natural language patterns. In: Proceedings of COLING-ACL'98, Montreal, Canada, 1998, 63~73

4 Erik F., Tjong Kim Sang, Sabine Buchholz. Introduction to CoNLL-2000 shared task: Chunking. In: Proceedings of CoNLL-2000, Lisbon, Portugal, 2000, 127~132

5 Li Mu, Lu Xue-Qiang, Yao Tian-Shun. A machine translation model based on E-Chunk. Journal of Software, 2002, 13(4): 669~675(in Chinese)
(李 沐, 吕学强, 姚天顺. 一种基于 E-Chunk 的机器翻译模型. 软件学报, 2002, 13(4): 669~675)

6 Fung P., Church K. W.. K-vec: A new approach for aligning parallel texts. In: Proceedings of the 15th International Conference on Computational Linguistics (COLING'94), Tokyo, Japan, 1994, 1096~1102

7 Abney S.. Parsing by chunks. In: Berwick R., Abney S., Tenny C. eds.. Principle-Based Parsing. Kluwer Academic Publishers, 1991

8 Cheng Wei, Zhao Jun, Xu Bo, Liu Fei-Fan. Bilingual chunking for Chinese-English spoken-language translation. Journal of Chinese Information Processing, 2003, 17(2): 21~27(in Chinese)
(程 葳, 赵 军, 徐 波, 刘非凡. 一种面向汉英口语翻译的双语语块处理方法. 中文信息学报, 2003, 17(2): 21~27)

9 Yael K., Edelman S.. Learning similarity-based word sense disambiguation. Computational Linguistics, 1998, 24(1): 41~60

10 Brown P. F., Stephen A. D., Vincent J. D., Robert L. M.. The mathematics of statistical machine translation: Paramater estimation. Computational Linguistics, 1993, 19(2): 263~309

CHENG Wei, born in 1973, Ph. D. Her research interests include spoken-language translation and natural language processing.

ZHAO Jun, born in 1966, Ph. D., associate professor. His research interests include natural language understanding, intelligent information processing and knowledge engineering.

LIU Fei-Fan, born in 1979, Ph. D. candidate. His research interests include computational linguistics and information extraction.

XU Bo, born in 1966, Ph. D. professor. His research interests include speech recognition and speech translation.

Background

This work is supported by the Natural Sciences Foundation of China under grant No. 60272041, 60121302. The project is titled as "Chunk and Sentence Pattern Alignment Based Statistical Machine Translation". The research on machine translation (MT) has both great application and academic value, and the effective combination of statistical and rule-based methods is one of the most promising directions in the field of MT. The project will emphasize its studies mainly on: statistical translation model based on chunk alignment and sentence pattern alignment; the preprocessing algorithm; the approach of chunk alignment and sentence alignment for translation; automatic and rough evaluation for MT systems, etc. The objective is to find a new practical and effective statistical MT approach based on chunk and sentence pattern alignment. The approach will incorporate the analysis strategies in rule-based method and the ideas of chunking and similarity measurement in example-based method into the basic statistical MT model. In the approach, the chunks are the basic analysis element to compensate for the problems resulted from the too small analysis elements in the word-based model, while sentence pattern is contributed to resolve the problem with the absence of linguistic knowledge in the pure statistical model. The above two ideas can improve the performance of statistical translation system in its robustness and quality. The research contents of the project include:

1. Chunk-based statistical machine translation (SMT) framework;
2. The approaches of preprocessing in SMT;
3. Bilingual chunk alignment;
4. Study on sentence pattern for SMT;
5. Automatic evaluation of SMT.

At present, we have had some progress in 1, 2, 3 and 5. The paper mainly introduces our work on 3-bilingual chunk alignment, which plays an important role in the whole framework.