基于汉英双语语料库的翻译等价单位自动获取研究

常宝宝(北京大学)

关键词:双语语料库;翻译等价单位;翻译等价单位抽取

摘　要:双语语料库在机器翻译或机器辅助翻译研究中的重要作用已经越来越多地得到研究人员的认可。本文探讨了如何利用汉英双语语料进行汉英翻译等价单位的抽取,提出了基于词语关联度进行多词组合单位的识别方法,并利用假设-检验的方法,在汉英双语语料库中抽取翻译等价单位。本文还对不同的关联度量方法进行了对比,并提出利用范畴假设改进抽取算法的效率。

Extraction of Translation Equivalent Pairs from Chinese-English Parallel Corpus

CHANG Baobao

Keywords: bilingual corpus, translation equivalent pair, automatic extraction of TEPs

Abstract: More and more researchers have recognized the potential value of the parallel corpus in the research on Machine Translation and Machine Aided Translation. This paper examines how the translation equivalent pairs could be extracted from parallel corpus. An iterative algorithm based on degree of word association is proposed to identify the multiword units for Chinese and English. Then a hypothesis-testing approach is used to extract the Chinese-English Translation Equivalent Pairs. We also made comparison between different statistical association measurement and proposed to use categorical hypothesis to improve the performance of extraction.

一、引　言

双语词典是基于规则机器翻译系统的一个关键部件。为了能够合理覆盖真实语言文本,一个实用的机器翻译系统,动辄需要数十万双语词条,编纂、维护双语词典的任务十分艰巨。大规模双语词典的构建始终是限制一些实验系统走向实用的瓶颈性问题。因而,从长远看,双语词典的自动或半自动构建是一条必经之路,从大规模双语真实语料中获取双语词典的研究是十分有意义的,大规模双语语料库不但为双语词典自动或半自动构建提供了一个可靠的基础,还在以下两点上有助于词典质量的提高:

(1)利用双语语料库提取双语词典,有利于保证所提取的源语言单词及其目标译词具有真正的翻译等价关系。长期以来,许多实用机器翻译系统,为了迅速扩充词典,大多尽可能利用已经存在的机器可读词典。然而,通常这

些语文性或百科性词典的预期用户是人，功能定位一般是辅助语文教学和外语理解。在这种定位下，双语词典中更多关注的是如何用目标语言定义一个源语言单词，而不是提供源语言单词的等价目标语译词。然而对于机器翻译而言，把一个源语言单词译作其目标语言定义往往是不合适的。

(2) 利用双语语料库提取双语词典，有利于扩大提取翻译单位。长期以来，机器翻译默认词为语言翻译的基本单位。但是，从翻译人员的翻译实践来看，仅仅把词作为翻译单位并不合适，一些多词组合或搭配在特定的语言环境中，往往需要作为一个整体来进行翻译。然而，人工编纂多词组合及其目标译语往往是不现实的，需要经年累月的艰苦努力。利用双语语料库进行自动获取有助于加快这一进程。

正是由于基于双语语料库词典编纂的种种优势，目前国内外相关研究十分活跃，应用目标涵盖了辅助传统双语词典编纂以及信息处理用双语词典编纂。本文基于这些工作，探索从汉英双语语料中获取翻译等价单位，在提取双语词表的同时，探索了双语多词单位及其目标译语提取方法以及利用词性、句法模式信息改善双语翻译等价单位的抽取性能。

二、基本定义

本文中使用了下面三个基本术语，其含义分别如下：

多词组合单位(MWU)：多词组合单位指源语言或目标语言文本中，稳定共现、并且具有合理句法结构的多个单词的组合。

翻译单位(TU)：一个翻译单位是源语言或目标语言中的一个词或一个多词组合单位。

翻译等价单位(TEP)：一个翻译等价单位是一个二元组(STU, TTU)，其中 STU 是一个源语言翻译单位，TTU 是一个目标语言翻译单位，STU 和 TTU 之间具有翻译等价关系，也就是说 STU 以及 TTU 之间具有互译关系。

三、基于双语语料库的翻译等价单位获取流程

从双语语料库中获取翻译等价单位的工作可以进一步分作三个步骤，如图 1 所示。

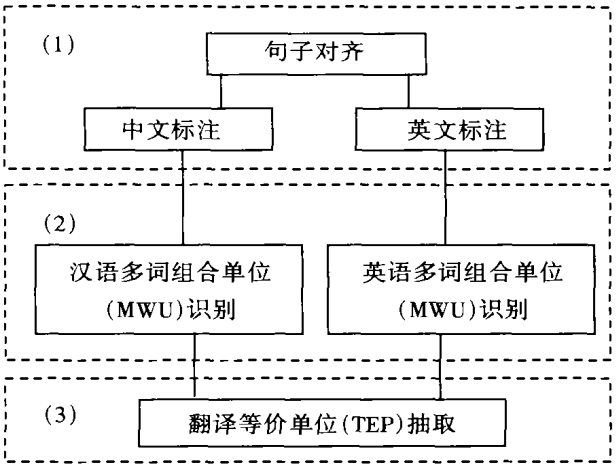

图 1　基于双语语料库的翻译等价单位自动获取模型

(1) 双语语料的预处理，为了获取汉英翻译等价单位，首先要对双语语料库进行必要的预处理，主要包括：对语料库的汉语部分进行切分和词性标注；英语部分进行断词(Tokenization)、形态分析以及词性标注处理；汉语文本和英文文本在句子一级进行对齐。

(2) 在预处理的基础上，分别识别标记汉语部分和英语部分中的多词组合单位。

(3) 翻译等价单位的自动抽取。

四、双语语料的预处理

本文工作使用了香港法律文献语料库，该语料库主要收录了香港特别行政区出版和使用的法律条文以及修正条款，由于特殊的历史背景，香港的所有法律条文都同时具有中英两个版本，这为双语技术研究提供了一个重要的研究基础。

由于所有法律文献中文英文逐条对应，所以基本上不再需要进行句子一级的对齐工作。为了探索词性、多词固定搭配的影响，

对双语语料库主要进行了下列预处理。对其中的中文部分进行了中文切词和词性标注工作，英文部分进行了断词(tokenization)、形态分析(lemmatization)和词性标注工作。中文切词和标注是按照北京大学计算语言学研究所有关规范进行的。英文词性标注使用了宾州树库的词性标注集。图2给出了经过预处理加工后的语料的一个片段。

Chinese texts		English Texts		
...		...		
<s id = 5>		<s_id = 5>		
本	r	This	DT	this
条例	n	Ordinance	NN	ordinance
可	d	May	MD	may
...		...		
通则	n	General	JJ	general
条例	n	Clauses	NNS	clause
》	w	Ordinance	NN	ordinance
。	w	.	.	.
<s id = 6>		<s_id = 6>		
附注	n	Remarks	NNS	remark
：	w	：	：	：
...				

图2 经过预处理的汉英双语语料

在图2中，XML标记分别标明了句子(法律条款)的编号，编号相同的两个句对即可视做一个对齐的句对。中文部分第一列是切词结果，第二列是词性信息。英文部分第一列是断词结果，第二列是词性信息，第三列是单词经过形态分析得到的单词的词典形式。

五、 事件关联度度量

为了识别单语多词单位组合以及抽取翻译等价单位，我们共使用了四种随机事件的关联度度量方法。基本原理如下：

令X和Y为两个随机事件，这两个随机事件的分布情况可以用下面的联立表来描述：

	Y	$-Y$
X	A	b
$-X$	C	d

图3 2×2的联立表

表中a、b、c、d的含义如下：

a：所有试验中，事件X和Y共现次数

b：所有试验中，事件X出现但事件Y不出现的次数

c：所有试验中，事件X不出现但事件Y出现的次数

d：所有试验中，事件X和Y均不出现的次数

基于上述联立表，目前有多种关联度度量方法，我们使用了点式互信息、DICE系数、统计值以及对数可能性分值计算两个事件的关联度，计算方式如下：

(1) 点式互信息

$$MI(X,Y)=\log_2\frac{n\times a}{(a+b)\times(a+c)}$$

(2) DICE系数

$$DICE(X,Y)=\frac{2a}{(a+b)\times(a+c)}$$

(3) χ^2 统计值

$$\chi\chi^2(X,Y)=\frac{n\times(a\times d-b\times c)}{(a+b)\times(a+c)\times(b+d)\times(c+d)}$$

(4) 对数可能性分值

$$LL(X,Y)=2\times\left(a\times\log\frac{a\times n}{(a+b)\times(a+c)}+b\times\log\frac{b\times n}{(a+b)\times(b+d)}+c\times\log\frac{c\times n}{(c+d)\times(a+c)}+d\times\log\frac{d\times n}{(c+d)\times(b+d)}\right)$$

六、 多词组合单位的自动识别

我们认为，多词组合单位是单词的一个扩展，多词组合单位既不是单词也不是完整的短语，而是介于二者之间的一种语言单位。多词组合单位应该具有下面的属性：

1) 组成多词组合单位的各个词应该频繁共现，从统计学的角度看，多词组合单位应该是以高于期望值的方式共现的多个词的组合。

2) 多词组合单位不是任意词的任意组合，从语言学角度看，多词组合单位应具有合

理的内部句法结构。

基于上述的两个属性，我们使用了一种结合统计方法以及语言学方法的递归识别汉语、英语多词组合单位的算法，其工作过程如下：首先利用第五节中所提出的关联度度量办法，计算汉语文本以及英语文本中所有词对之间的关联度，并标记出所有关联度大于某预设域值的词对。然后，递归调用算法，进一步标记出长度大于2词的具有较强关联度的多词组合。然而经过以上统计手段得到的多词组合很多并不具有合理的句法结构，为此，我们开发了一个过滤器，该过滤器使用一个句法模式表，检查统计出来的多词组合单位所对应的词性序列是否构成一个合法的句法模式，如果不构成合法句法模式，则剔除。如果构成合法句法模式，则作为一个多词组合单位。图4是过滤器中使用的句法模式，表中左部为汉语句法模式，右部是英语句法模式。

"a + n",	"NN + NN",
"b + n",	"NN + NNS",
"n + n",	.....
...	"NN + IN <of> "
"MWU + n",	"JJ + NN",
"n + MWU",	...
"MWU + MWU"	"MWU + MWU"

图4　预定义的多词组合单位的句法模式

七、双语翻译等价单位的自动抽取

关于双语词对的提取，目前提出的方法主要有两种，一种基于假设－检验的办法，另一种基于统计翻译模型。我们主要采用了基于假设－检验的办法，该方法主要基于下面的观察：互为翻译的一对单词要比相互不为翻译的一对单词更有可能出现在同一个对齐的句子对中。

由于已经对语料库进行过多词组合单位的识别，所以算法不但可以抽取单词及其译词，也可以抽取出多词翻译单位及其目标翻译，即翻译等价单位。整个抽取过程可以分作两个阶段，第一个阶段为生成阶段，从对齐的语料库中列出所有可能的翻译等价单位。第二个阶段，可视作“检验阶段”，该阶段选择那些关联度高于期望值的翻译等价单位输出。在这里，关联度度量再次使用了第五节中介绍的DICE系数、点式互信息、对数可能性分值以及统计分值。

由于在生成阶段，上述算法会穷尽生成所有可能的翻译等价单位，这使得搜索空间很大，算法效率较低，尤其是处理较大规模的语料时更是如此。为了提高算法的效率，我们作了如下的假设：即假设一个源翻译单位一般会被翻译成为一个句法类别相同的目标语言翻译单位。例如，英语的名词一般翻译作汉语的名词，英语结构（"JJ + NN"）一般翻译作汉语结构（"a + n" /"b + n"）。这个假设称为范畴假设。很明显，对于汉英互译或英汉互译，这个假设并不总是成立。但确实使得算法在准确率下降较小的情形下而极大地提高了算法的效率。

八、实验以及结果分析

基于上述的方法，我们初步实现了一个基于汉英双语语料库抽取翻译等价单位的原型系统，并针对不同的关联度度量方法，初步进行了一些实验，测试不同统计度量方法在效果上的差异、范畴假设对抽取结果的影响。

目前，实验使用了一个小规模的语料库，作为测试语料，我们从香港法律文献语料库中选择了500个句对（约40000汉字，25000英语单词）进行。对这500个句对进行预处理，预处理不进行校对，分别采用点式互信息(MI)、DICE系数(DICE)、对数可能性(LL)以及统计值(CHI)作为关联度量办法进行翻译等价单位的提取工作，并选择提取结果的前100个翻译等价单位进行评价，计算正确或部分正确的词对，结果如图5所示：

	MI	DICE	LL	χ^2
正确	39	5	70	75
部分正确	5	1	10	6
准确率	44%	6%	80%	81%

图 5　不同关联度度量的性能差异

从图 5 的结果可以看出，对数可能性分值以及统计值的抽取效果优于互信息和 DICE 系数。其原因似乎可以从对数可能性分值以及统计值在计算时均考虑了随机事件均不出现的情形，即联立表中的元素 d。

我们也针对范畴假设进行了实验，实验显示，使用范畴假设，抽取准确率会有下降，但抽取速度得到明显改善。对上述 500 个句对，在一台 CPU 为 Pentium III 800MHz，内存配置 128M 的机器进行名词性翻译等价单位的提取，耗时越 4 秒钟，提取准确率为 79%，而不使用范畴假设，抽取耗时 90 秒钟。可见，使用范畴假设，准确率下降约 4%，但运行效率提高了 200%。

图 6 是针对上述 500 个句对进行抽取得到的部分结果，在识别多词单位组合单位以及翻译等价单位抽取时，选用统计值作为关联度度量办法。在图中，位于双语翻译等价单位右部的/＊＊/之间的数值是翻译单位之间的统计分值。

九、结束语

本文主要探索了利用汉英双语语料库进行翻译等价单位的抽取，提出了一种基于关联度的递归多词组合单位识别算法，并利用假设－检验技术进行汉英翻译等价词对的抽取工作，并尝试利用范畴假设改善翻译等价单位的抽取算法的效率。本文还考察了不同的关联度量在汉英翻译等价单位抽取工作中的表现，并利用实验进行了分析和解释。这些小规模的实验表明基于双语语料库提取翻译等价单位在很大程度上是可行的，对机器翻译系统的词典构建工作可以起到一个很好的辅助作用。

参 考 文 献

〔1〕 Simard, M. et al., Bilingual text alignment: where do we draw the line, in Multilingual Corpora in Teaching and Research, Rodopi publisher, 2000, pp. 38－64.

〔2〕 Gale, W., Identifying words correspondences in parallel Texts, DARPA speech and Natural language

```
1. 见 see　/* CHI2 score = 496.471 */
2. 追溯力 _的　see　/* CHI2 score = 496.471 */
3. 款　subsection /* CHI2 score = 496.237 */
4. 废除　repeal　/* CHI2 score = 495.814 */
5. 命令　order /* CHI2 score = 493.195 */
7. 豁免　exemption　/* CHI2 score = 490.829 */
25. 附属 _法例　subsidiary_legislation　/* CHI2 score = 477.173 */
26. 公共 _机构　public_body /* CHI2 score = 475.711 */
28. 财政司 _司长　Financial_Secretary　/* CHI2 score = 475.711 */
31. 条例　ordinance　/* CHI2 score = 470.081 */
34. 基本 _文书　primary_instrument　/* CHI2 score = 468.068 */
41. 卫生 _主任　health_officer　/* CHI2 score = 468.068 */
42. 裁判官　magistrate　/* CHI2 score = 468.068 */
43. 担当　discharge　/* CHI2 score = 468.068 */
45. 合约　contract　/* CHI2 score = 468.068 */
46. 终审 _法院 _首席 _法官 Chief_Justice_of_Final Appeal /* CHI2 score = 468.068 */
53. 香港 _特别 _行政区　Hong_Kong_Special_Administrative_region　/* CHI2 score = 448.576 */
63. 审裁处　tribunal　/* CHI2 score = 420.579 */
64. 宣布　declare　/* CHI2 score = 420.579 */
```

图 6　抽取结果样例

workshop. Asilomar, CA. , 1991

〔3〕Fung, P., K－vec: A new approach for aligning parallel texts, 15th international conference on computational linguistics. Kyoto, Japan., 1994, 1096－1102.

〔4〕Brown, P., The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, Vol 19, No. 2, 1993, 263－311.

〔5〕北京大学计算语言学研究所．词语切分与词性标注——规范与加工手册，见 http://www.icl.pku.edu.cn/research/corpus/coprus－annotation.htm

〔6〕宾州树库词性标柱集及手册，见 http://www.cis.upenn.edu/～treebank/

〔7〕Chang Baobao et al., Chinese－English Translation Database: Extracting units of translation from parallel texts, in proceedings of 6th TELRI Seminar, Bulgaria, 2001.

〔8〕Jörg Tiedemann, Automatic Lexicon Extraction from Aligned Bilingual Corpora, Ph. D. Thesis in Otto－von－Guericke－Universität Magdeburg.

〔9〕Dekai Wu and Xuanyuin Xia. Learning an English－Chinese Lexicon from a Parallel Corpus. In Proceedings of the 1st Conference of the AMTA, Columbia/Maryland, 1994. Association for Machine Translation in the Americas.

〔10〕Tufis, D., Computational bilingual lexicography: automatic extraction of translation dictionaries, In Journal of Information Science and Technology, Romanian Academy, Vol. 4, No. 3, 2001

第五届东亚术语论坛将在海口市召开

第五届东亚术语论坛将于2002年12月1日至8日在海口市召开。本届东亚术语论坛将继续举办术语学与术语标准化学术会议。前四届座谈会分别于中国(1997)，韩国(1998)，蒙古(1999)和日本(2001)召开。

东亚术语论坛的目标是促进东亚地区术语学的研究和发展，为东亚术语研究中心的创立做准备，其秘书处及日常工作设在中国标准研究中心。

论坛的主要活动：论坛各成员之间术语研究的成果与发展的信息交流；全球术语学文献的交流与收集并建立档案；贯彻执行各领域与术语学相关的活动；为创建多语言和多功能的术语库做准备；识别术语学领域的需求并执行联合研究项目；促进和协调术语学工作的知识与技术的传播；促进东亚术语论坛成员间在术语学及其文献上的合作；促进术语学研究包括相关的教育和培训；翻译和出版术语学著作并促进研究成果的应用。

论文征集：本届术语论坛的论题包括：

●术语学同其他学科间的关系(术语学同哲学，术语学同语言学，术语学同心理学，术语学同知识工程，术语学同知识工程，术语学同情报与文献，术语学同图书馆学，术语学同百科全书)

●社会科学和人文学科中的术语工作(社会科学和人文学科术语学的特点，综述：既有工作的回顾，主要研究方法：问题和解决办法)

●术语标准化(术语标准化的经济效益和社会效益，术语标准化同经济发展：宏观和微观，术语标准化的原则与方法，概念和术语的国际协调，术语集的系统分类)

●描述性术语学(描述性术语学和规范性术语学的比较，描述性术语学：作为术语标准化的前期工作，描述性术语学：作为必然的选择，描述性术语学：作为向拥护提供及时服务的独立活动)

●术语学同词典编纂(词典编纂用符号，词汇的设计，单语词典：特别是分类释义词典和专业词典，多语词典)

●基于汉字符的信息处理和术语(汉字符的标准化，基于汉字符的术语间的协调，汉字符的切分)

●计算机辅助(机助)术语工作(术语学同知识数据管理，字词和术语的多语种概念表示问题，术语的动态管理，术语提取，术语数据库的开发，电子词典)

●词典和主题词表的机助编纂(电子百科全书，术语学同多媒体/超媒体出版，机器翻译，多语种处理技术，信息检索技术，术语库间的数据交换和信息共享)

★ 术语学同因特网(网上术语工作：术语学的新时代，术语论坛，用户参与的术语编纂工作，术语学同信息高速公路)

●术语学同知识传播(教育工作中的术语学，术语学同翻译，术语学同科技培训，科技领域中的术语学、通讯和发展，发展中国家的术语工作)

●术语的实际问题(消费者需求和术语学产品，术语学的经济学，术语学和全球经济，术语学和世界贸易组织，术语和产权)

稿件请用电子邮件寄至：yuxl@cnis.gov.cn

或邮寄至下面地址：北京市朝阳区育慧南路3号 100029 于欣丽收

稿件格式要求：用10磅字型，单栏，两倍行距；调整页边距，使文字可容纳于14.5×23厘米(5.5×9英寸)的矩形范围内。在第一页标题和文本主体之间，要填入(所有)作者的姓名、地址和电子信箱地址。投稿截止日期：2002年8月25日。