构建和剖析中英三元组可比语料库

胡小鹏,袁 琦,耿鑫辉,朱 姝
HU Xiaopeng, YUAN Qi, GENG Xinhui, ZHU Shu

中国电子信息产业发展研究院,北京 100044
China Center for Information Industry Development(CCID), Beijing 100044, China

HU Xiaopeng, YUAN Qi, GENG Xinhui, et al. Building and profiling Chinese-English 3-tuple comparable corpora. Computer Engineering and Applications, 2014, 50(13):153-157.

Abstract: There exists inherent skewed language model in Chinese-English parallel corpus due to the influence of translationese. Obviously, natural language processing systems trained with these corpora, including machine translation and cross-language information retrieval, will inherit the skewed language model, thus seriously degrading the performance of applications. To fix the inherent defaults in parallel corpus, this paper proposes a technical research on building and profiling Chinese-English 3-tuple comparable corpora. The study adopts comparable corpora and automatic language profiling technologies and applies a combined method of statistics and rules for statistical analysis on native English and Chinglish in 3-tuple comparable corpora that consists of native English, Chinglish and standard Chinese. Based on this, automatic extraction technologies, such as *n*-grams and key clusters, are used in the mining of native-language-based bilingual resources to improve and develop natural language processing applications such as machine translation.

Key words: 3-tuple comparable corpora; language transfer; automatic language profiling; *n*-grams

摘 要: 由于受到翻译腔的影响,中英平行语料库存在固有的扭斜的语言模型。显然,用这样的语料库训练的机器翻译、跨语言检索等自然语言处理系统也承袭了扭斜的语言模型,严重影响到应用系统的性能。为了克服平行语料库固有的缺陷,提出构建和剖析中英三元组可比语料库的技术研究。这项研究采用可比语料库和语言自动剖析技术,使用统计和规则相结合的方法,对由本族英语、中式英语和标准中文三元素所组成的三元组可比语料库中的本族英语和中式英语进行统计分析。在此基础上,利用*n*-元词串、关键词簇等自动抽取技术挖掘基于本族语言模型的双语资源,实现改进和发展机器翻译等自然语言的处理应用。

关键词: 三元组可比语料库;语言迁移;自动语言剖析;*n*-元词串
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.1310-0106

1 引言

自1995年Rapp提出基于矩阵相似度计算的可比语料库双语词汇提取算法研究以来,构建和使用可比语料库的研究得到不断发展。特别是近十几年,随着网络跨语言资源和跨语言处理需求的剧增,基于可比语料库的双语资源提取研究已从最初的双语词汇提取发展到双语句对提取,双语片断提取,基于本族语言模型的双语资源提取,语义知识库建造,以及利用人机语言特征对比改进机器翻译系统等一系列可比语料库的数据挖掘研究。到目前为止,除本文发表的研究成果外,国内外尚未发现基于本族语言模型的可比语料库双语资源提取报道。随着可比语料库研究与应用的不断发展,它已成为自然语言处理各种相关学术会议的一个中心话题。从2008年起,ACL为该领域的研究创建了专门的学术交流平台,每年设定中心议题,召开"构建和使用可比语料库(BUCC)"专题研讨会。2013年8月召开的第6次研讨会的中心议题,是改进和发展可比语料库经典的词汇挖掘技术,提高数据挖掘准确度,扩展应用覆盖面。

本文中,构成三元组可比语料库的中式英语又称Chinglish,它有悖于本族英语规则和英语国家文化习

基金项目: 国家自然科学基金(No.61172101,No.61172102)。
作者简介: 胡小鹏,男,博士,主要研究方向:自然语言处理,机器翻译;袁琦,男,研究员;耿鑫辉,男,高级工程师;朱姝,女,硕士。
E-mail:huxp@ccidtrans.com
收稿日期: 2013-10-14 **修回日期:** 2013-12-23 **文章编号:** 1002-8331(2014)13-0153-05

惯。根据拉多(R.Lado)在《跨文化的语言学》中提出的“语言迁移(language transfer)”理论,中式英语充分表征了中国人在英语写作中母语的负迁移现象。由于受到汉语语言、文化、思维习惯等各方面的影响和干扰,中国人按照自己母语的习惯,主观编造、生搬硬套构造了中式英语,其中在词汇层面表现出的负迁移现象尤为严重。人们往往不顾两种语言的本质差异,直接把母语的表达方式生搬硬套到英语词汇中去。用包含着词汇层面负迁移现象的译文构建的平行语料库显然存在着扭斜的语言模型。图1中marketizaton reform是国内学术期刊上出现的词汇层面的中式英语典型例子,正确的本族英语表达是market-oriented reform。

图1　词汇层面的中式英语

由于从平行语料库提取的双语数据受到中式英语扭斜的语言模型影响,严重影响到跨语言处理应用。以Google在线跨语言检索为例,当检索“英国电子信息产品”时,Google的输出结果主要是涉及“图书馆服务和图书”文献(见图2的屏幕截图)。其原因是,根据平行语料库训练出的应用系统包括有扭斜的语言模型,在输入“电子信息产品”后,系统无法优先生成“electronics and IT products”,而是扭斜的表示电子图书类的“electronic information products”。

图2　“英国电子信息产品”Google跨语言信息检索结果

平行语料库是跨语言处理的重要资源。为克服平行语料库固有的缺陷,本文提出了构建和剖析中英三元组可比语料库的技术研究。这项研究使用统计和规则相结合的方法,对由本族英语、中式英语和标准中文三元素所组成的三元组可比语料库中的本族英语和中式英语进行统计分析。在此基础上,利用*n*-元词串、关键词簇等自动抽取技术挖掘基于本族语言模型的双语资源,改进和发展机器翻译等自然语言处理应用。本文提出的研究内容不仅对改进和发展跨语言处理应用具有实用价值,而且对外语教学、词典编纂、对外交流与合作也具有重要意义。

2　相关研究

2.1　国外相关研究

近年来,国外基于可比语料库的数据挖掘研究发展极其迅速。尤其是,基于可比语料库的双语术语提取成为国外可比语料库研究最为活跃的领域。对于科技领域,尤其是对于新兴领域,术语资源往往是短缺的或不是最新的。为了应对新兴和迅速发展的科技领域词汇短缺和陈旧的瓶颈,以及平行语料库固有的时间滞后和文本稀缺问题,在欧盟第7框架计划2010年—2012年期间,英、法、德等国通过实施基于可比语料库的术语提取(TTC)项目,实现了从特定领域(如再生能源)可比语料库提取中英、中法等12部词库的研发计划。TTC项目开发环境的数据工作流如图3所示,包括文本预处理、单语术语提取和双语术语对齐3个层面的开发工具模块。文本预处理模块包括词性还原、词性标注、词干提取和词形还原。单语术语提取模块用于处理单语语料库文件并提取术语,其处理流程包括识别并建立单字词和多字词的索引,计算词语的相对频率和领域特殊性,检测单个词术语构成的新古典复合词,以及采用相对频率或领域特殊性设定阈值过滤候选项。双语术语对齐模块可以根据术语不同的性质,采用不同的策略。对于单个词的术语采用基于上下文的预测方法,对于新古典复合词和多词术语采用基于语意合成性(compositionality)的方法。通过评估验证,该项目所产生的双语术语库有效地改进了面向特定领域的机器翻译性能[1-2]。

图3　TTC开发环境的数据工作流

2013年Dhouha等人在深入研究经典的可比语料库双语词汇提取技术基础上，观察到翻译上下文词向量中多义词的语义歧义问题，提出了基于WordNet的语义相似度度量的词义消歧处理的可比语料库双语词汇提取方法。

实验中，在经典的双语词汇提取3步骤，即建立上下文向量、翻译上下文向量、比较源语和目标语向量中加入了对上下文向量翻译的语义消歧步骤（见图4），使用单义词作为消除歧义的种子集来推断多义词的翻译意思，以减少上下文向量中的干扰噪音，提高双语词汇提取性能。

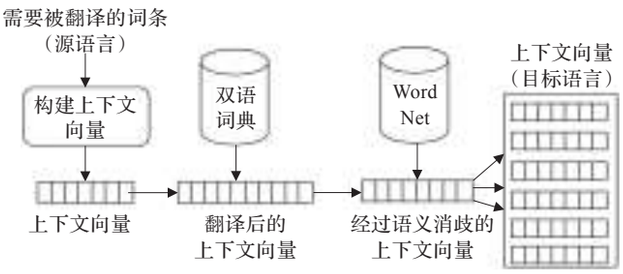

图4　基于WordNet语义相似度的可比语料库双语术语提取方法架构图

首先，利用双语词典中只含有一个义项的词条来构造单义词种子词典，在WordNet的检测中，这种方法的准确率可以达到95%。其次，通过基于路径长度的语义相似度的WUP算法[3]，在WordNet基础上，计算目标术语的上下文词向量中各单义词的义项与多义词的各个义项之间的语义相似度值。WUP算法利用两个词的同义词集 (s_1, s_2) 在WordNet中的深度和它们的最小公共包含（LCS），计算出两个词之间的相似度值，公式如下：

$$WupSim(s_1, s_2) = \frac{2 \times depth(LCS)}{depth(s_1) + depth(s_2)}$$

实际中，由于一个词可能会同时属多个同义词集，最终选取两个词的所有可能的相似度值中的最大值，作为两个词的相似度值，公式如下：

$$SemSim(w_1, w_2) = \max\{WupSim(s_1, s_2), (s_1, s_2) \in synsets(w_1) \times synsets(w_2)\}$$

最后，利用上下文词向量中，多义词各个义项与各个单义词义项的平均相似度值，为多义词的每个义项打分（公式如下），并选取分值最高的义项作为多义词的最终词义，以此达到语义消歧的目的。

$$Ave\_Sim(w_p^j) = \frac{\sum_{i=1}^{N} SemSim(w_i, w_p^j)}{N}$$

实证实验结果表明，该方法明显优于经典的方法[4]。

在可比语料库双语句对提取方面，经典的方法是使用信息检索（IR）技术，在文档对齐的基础上，使用句子层面模型来提取平行句对（或片断）。IBM Watson实验室的Tillmann等人提出了一种新的从可比数据中提取句对的算法，使用这种算法可以直接在句子层面打分候选句对集。基于该算法的句对提取，是通过有效执行基于IBM模型1翻译概率的对称打分函数实现的。该方法适用于无文档层面对齐信息的可比语料库句对提取[5]。在可比语料库双语片断提取方面，Munteanu等人受信号处理的启发，提出了在句子级别无法对齐的可比语料库中提取双语片断的算法。以词对齐概率（使用GIZA++获得）和对数似然比为统计量，来描述词汇间的相关性，在这些统计数据基础上，用过滤器模型从可比语料库中提取双语片断。他们把从可比语料库提取结果应用于统计机器翻译系统，BLEU测评值得到显著提升[6]。在基于可比语料库的语义知识库建造方面，Genc等人利用基于Wikipedia的多语可比语料库，通过候选实体匹配标题的算法和多条件对比抽取算法，构建中-英对照知识本体并发展了知识本体的可视化技术[7]。2013年，Ekaterina等人发表了“用可比语料库分析翻译变异“的成果，使用相同文本的不同翻译变体即专业人工翻译，基于规则机器翻译（Systran和Linguatec）和基于统计机器翻译（Google和Moses）构建可比语料库，从人机语言特征对比角度，开展单语可比语料库的翻译对比研究，改善机器翻译性能[8]。

2.2　国内相关研究

在可比语料库双语词汇提取方面，张永臣等提出了一种从可比语料库中抽取特定领域双语词典的算法，给出了利用词间关系矩阵法从特定领域可比语料库中抽取双语词典的过程，通过大量实验分析了种子词选择对词典抽取结果的影响，其实验结果表明种子词的数量和频率对词典抽取结果有积极作用[9]。孙广范等采用双向等价对获取计算然后求交集等方法提高翻译等价对提取正确率[10]。徐会芳等使用基于相似度计算和多特征融合的方法以及最小化样本风险算法调节特征权重，来提高从可比语料库中抽取双语术语互译对的准确率[11]。在可比语料库双语句对提取方面，Fung等人提出利用通用网络爬虫持续抓取网络资源来构建面向多领域的超大规模可比语料库，从中提取平行句对改善机器翻译性能。项目中使用面向招回和面向精度的算法，基于信息检索技术处理网页，匹配文档并提取平行句对。通过对网络资源的深入挖掘，来获取更多的语言资源[12]。胡弘思等在Wikipedia基础上，统计词汇数据、构建命名实体词典，并通过其本身的对齐机制构建了双语可比语料，从中抽取对齐句子[13]。基于本族语言模型的双语资源提取方面，肖健等人通过构建三元组可比语料库，解决了由中式英语导致的语言模型“扭斜”问题，进一步提高了MWE的自动抽取准确率，改善机器翻译效果[14]。另外双语资源提取方面，张桂萍等提出了面向单一双语网页的双语资源挖掘方法[15]。该方法重点采用了以频繁序列模式为特征的SVM分类方法，实现了包含双语资源的单一双语网页的筛选与识别，并以此为基础构建可比语料库，挖掘具有对译的双语资源。

3　研究框架

本文提出的研究框架包括三元组可比语料库建设，关键词簇自动剖析，语义多词表达提取，以及翻译模板自动提取4个模块。这4个模块紧密衔接，三元组可比语料库是本项研究的基础设施，通过建设三元组可比语料库的研究，将为整个项目实施提供数据资源。在此基础上，通过对三元组可比语料库的关键词簇自动剖析的研究，可以发现和比较本族英语与中式英语语言模型的区别特征，改进和验证所采用的自动剖析算法。在对关键词簇统计研究的基础上，将进一步研究从三元组可比语料库提取本族英语的语义多词表达和翻译模板的算法与模型，以期实现改进和发展机器翻译等自然语言处理系统性能的研究目标。

3.1　三元组可比语料库建设

三元组可比语料库是开展本项研究的基础资源，到目前为止，已经累计构建了百万句对级的三元组可比语料库。构建语料库的原始语料主要来自我院每年都要发布的几十种，总字数超过200万英语词语的ICT领域研究报告。为确保研究报告译文的准确度和可读性，所有报告的英文译文，需经本族英语的语言专家严格修改和编辑。每年积累的中式英语和修改后的本族英语文本经过图5所示的流程处理；通过语料库比较分析工具，构建满足可比语料库取样框架（sampling frame）要求的三元组可比语料库。为了保证定量比较分析的准确度，利用工具过滤掉中式英语文本和本族英语文本之间差异在10行以上或者每行差异大于10%的句对。然后，使用我院的句法分析工具（CCID-CESAT）、语料库标注分析工具（CCID-CTAT）以及英国Lancaster大学Wmatrix和USAS语义分析工具，对三元组可比语料库进行句法分析、词性和语义标注。通过对语料库所做的这些训练，为后续的关键词簇自动剖析、语义多词表达和翻译模板自动提取的研究奠定了基础。

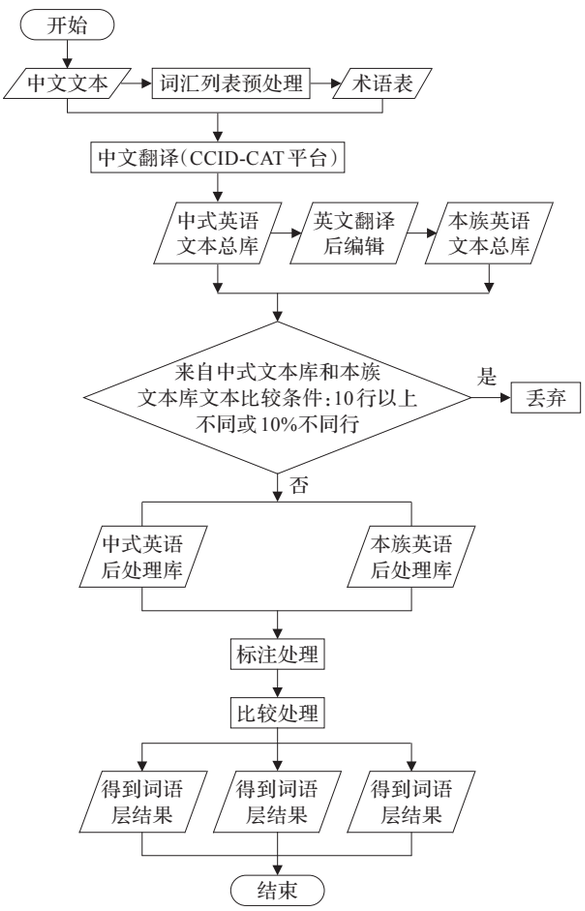

图5　三元组可比语料库的构建流程图

3.2　关键词簇的自动剖析

在建立三元组可比语料库的基础上，利用统计方法研究关键词簇在词语、词性和语义3个层面上的过使用和欠使用的语言现象，使用对数似然值（*LL*）定量分析关键词簇的差异显著性（keyness）。对数似然值计算方式如下：

假设 X 为要考察的关键词簇，a 为中式英语语料库中出现 X 的次数，b 为本族英语语料库中出现 X 的次数，c 为中式英语语料库中所有关键词簇的数目，d 为本族英语语料库中所有关键词簇的数目，其关系如表1的词频列联表所示。

表1　词频列联表

	中式英语语料库	本族英语语料库	总计
关键词簇X频度	a	b	$a+b$
其他关键词簇频度	$c-a$	$d-b$	$c+d-a-b$
总计	c	d	$c+d$

那么对数似然值（log-likelihood）计算方法[16]如下：

$$LL=-2\ln\lambda=2\sum_i O_i\ln\left(\frac{O_i}{E_i}\right)$$

O_i 为观察值，即表中的 a、b 值。E_i 为期望值，其计算方法如下：

$$E_i=\frac{N_i\sum_i O_i}{\sum_i N_i}$$

中式英语语料库中所有关键词簇的数目为 $N_1=c$，本族英语语料库中所有关键词簇的数目为 $N_2=d$，那么中式英语和本族英语中关键词簇的期望为：

$$E_1=c\times(a+b)/(c+d)$$

$$E_2=d\times(a+b)/(c+d)$$

对上述公式进一步解释如下：先求某个词 X 在整个语料库（指两个语料库：（1）中式英语；（2）本族语）中出现的概率（根据大数定理，用频率近似表示概率）。算法步骤是：（1）将 X 在两个语料库中的频次之和 $(a+b)$ 除以语料库中词的总量 $(c+d)$，也就是 E_i 等式右边除了 N_i 以外的那个分式。（2）再根据这个概率分别计算在中式英语中该词的期望出现次数，即为中式英语总词量乘以该词出现的概率。同理计算 E_2。

依据上述公式得到的 E_1 和 E_2，可以求得 LL 值：

$$LL=2\times((a\times\ln(a/E_1))+(b\times\ln(b/E_2)))$$

表2　词语表达层差异显著性剖析结果

关键词	归一化频率		过使用(+)/欠使用(-)	对数似然值	中文表达
	中式英语语料库	本族英语语料库			
electronic information products	471	5	+	604.37	电子信息产品
electronics and IT products	8	467	-	577.28	
network bubble	515	36	+	497.82	网络泡沫
Dot-com bubble	16	372	-	404.52	
e-government construction	126	3	+	150.34	电子政务建设
e-government development	9	120	-	113.55	
second-hand data	62	0	+	85.95	二手资料
indirect data	2	58	-	65.64	
sub-contract	36	0	+	49.90	子合同
subcontract	4	30	-	22.50	
Olympic five rings	20	0	+	27.72	奥运五环
the Olympic rings	0	24	-	33.27	
middle-sized	35	1	+	40.77	中等大小
medium-sized	4	30	-	22.50	
important significance	35	3	+	31.69	重要意义
great significance	6	42	-	30.37	

对数似然值最大的关键词簇排在列表的顶端,表明该词簇在本族英语和中式英语之间频次分布差异比较大。比如,某些关键词簇在中式英语中被过度使用或者欠使用。依据对数似然值的变化差异,可以发现中式英语与本族英语的区别特征,为本项目自动提取基于本族英语的翻译模板和语义多词表达研究提供重要参考。

4　实验结果

本研究利用关键词簇自动剖析技术(最大为5元词串)从词语表达层面分析了本族英语和中式英语的区别特征,计算出三元组可比语料库中本族英语和中式英语在词语表达层面的差异显著性。根据给定的 p 值和 LL 值,生成关键词簇过使用(overused)和欠使用(underused)对照表。表2仅列出对数似然值 LL 大于20的典型关键词和关键词簇。因为在计算期望值时,已经考虑到两个语料库的词次规模(即 c 和 d),所以在运用公式前,不需要做归一化处理[16]。事实上,表中给出的频率可以认为是以百万词次做归一化处理的,因此对表中所给数字可作直接比较。

从表2的中式英语语料库与本族英语语料库(参考语料库)的词语表达层差异显著性剖析结果可以看出,e-government construction(电子政务建设)、second-hand data(二手资料)和important significance(重要意义)等均为词汇负迁移现象引起的过使用词语,而e-government development,indirect data和great significance为欠使用词语。

通过上述分析,可以在三元组可比语料库中发现中式英语与本族英语的区别特征,实现自动提取基于本族英语模型的多词表达(MWEs)和翻译模板,改进和发展机器翻译等自然语言的处理应用。

5　结论

目前,构建和剖析三元组可比语料库的研究已在词汇表记层面取得有效成果,对克服中英平行语料库存在固有的扭斜的语言模型,建造和挖掘基于本族语言模型的双语词库,改进机器翻译等自然语言处理应用具有很大的实用价值。嵌入本项研究成果的机译系统已在国内外得到广泛使用。今后,按照本文的研究方法,也可以进行词性层面和语义层面的差异显著性剖析研究。本项研究今后的目标,是把基于关键词和关键词簇方法的可比文本微观研究扩展到基于关键语义场(key semantic fields)的可比文本宏观研究,使其支持内容分析。这样,就可以把当前对特定的三元组可比语料库的定量分析扩大到泛化的基于内容的可比文本的定性分析,有效地扩展了可比语料库的研究与应用。2013年8月召开的第6次"可比语料库构建和应用(BUCC)"研讨会的中心议题,是"改进和发展可比语料库经典的术语挖掘技术,提高数据挖掘准确度,扩展应用覆盖面",值此之际发表本项研究成果更具有现实意义。最后,感谢英国Lancaster大学Paul Rayson博士在本项研究中给予的理论和方法上的指导。

参考文献:

[1] Daille B.Building bilingual terminologies from comparable corpora:the TTC TermSuite[C]//Proceedings of the 5th Workshop on Building and Using Comparable Corpora,2012:29-32.

[2] TTC Annual Public Report 2012[R].2012.

[3] Wu Zhibiao,Palmer M.Verbs semantics and lexical selection[C]//Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics(ACL'94),Association for Computational Linguistics,1994:133-138.

(下转186页)