基于专业领域平行语料的双语核心术语抽取研究*

章成志[1,2]　王惠临[1]

1. 中国科学技术信息研究所，北京 100038；2. 南京理工大学信息管理系，南京 210094

E-mail: {zhangchz, wanghl}istic.ac.cn

摘　要：双语术语抽取在双语术语词典编撰、双语本体构建、机器翻译以及跨语言信息检索中具有重要的作用。其中，双语核心术语是双语术语识别和抽取的关键资源之一。本文将专业领域文档的关键词作为候选核心术语，利用中文和英文的专业领域分类语料，通过关键词抽取、术语度计算等关键技术，分别进行中文和英文的核心术语的识别。接着，以中英文专业领域平行语料为基础，通过双语对齐技术，自动生成中英文对照的双语核心术语列表。实验结果表明，每个专业领域中，前200对中英文对照核心术语的平均正确率在50%以上，个别领域正确率达80%左右。

关键词：双语核心术语，术语抽取，术语对齐，平行语料

Extract Bilingual Core Terminology from Parallel Corpora in Special Domain

Zhang Chengzhi[1,2], Wang Huilin[3]

1. Institute of Scientific & Technical Information of China, Beijing 1000383;

2. Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094

E-mail: {zhangchz, wanghl}istic.ac.cn

Abstract: Bilingual terminology extraction plays an important role in the bilingual dictionary compilation, bilingual Ontology construction, machine translation and cross-language information retrieval etc. The bilingual core terminology is the key resource for bilingual terminology extraction. The authors use the keywords of the document in special domain as the candidate core terminology. After the keywords extraction and termhood computation, the core terminology in Chinese and English are extracted from the classified corpora in special domain respectively. Then, the bilingual terminology alignment method is used to extract the bilingual core terminology from the parallel corpora in special domain. The experiment result shows that the average precision of the Top-200 bilingual core terminology is about 50% and the precision in some domain is 80%.

Keywords: Bilingual Core Terminology, Terminology Extraction, Terminology Alignment, Parallel Corpora.

1　前言

术语是在特定专业领域中一般概念的词语指称，一个术语表示一个概念。在自然语言的计算机处理的诸多领域中，如信息检索、信息抽取、文本分类等任务中，基本单位常为是单词型术语或词组型术语，离不开术语的自动处理[1]。同时，双语术语抽取在双语术语词典编撰、双语本体构建、机器翻译以及跨语言信息检索中具有重要的作用。双语核心术语是双语术语识别和抽取的关键资源之一。以双语核心术语对为双语术语种子对，在大规模语料基础上，通过机器学习方法可以发现更大规模的双语术语对。另一方面，专业领域文档包含大量专业术语。文档关键词通常是有效的候选核心术语。因此，如何充分利用现

*本研究受“十一五”国家科技支撑计划重点项目(2006BAH03B02)、中国博士后科学基金特别资助项目(200801105)、教育部人文社会科学研究一般项目(08JC870007)资助。

有的大规模专业领域分类资源，进行中英文对照的核心术语抽取，是一项很有意义的工作。

本文将专业领域文档关键词作为候选核心术语，利用中英文专业领域分类语料，通过关键词抽取、术语度计算等关键技术，进行中英文核心术语识别。接着以中英文专业领域平行语料为基础，利用双语对齐技术自动生成中英文对照的核心术语。实验结果表明，每个专业领域的前200对中英文对照核心术语的平均正确率在50%以上，个别领域80%左右。

2 相关工作概述

本文以专业领域文档集为基础，将文档关键词作为候选术语，根据术语度抽取核心术语，再依据中英平行文档进行核心术语的对齐。与该研究相关工作主要包括如下几个方面。

（1）双语术语抽取相关研究。相关工作包括：2000 年，孙乐和金友兵等人从英汉平行语料库中自动抽取双语术语词典[2]。2005 年，Bèatrice Daille 和 Emmanuel Morin 在可比语料上进行法-英双语术语抽取[3]。2006 年，张永臣和孙乐等人提出了一种从非平行语料中抽取特定领域双语词典的算法，他们的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[4]。2008 年，Maike Erdmann 根据维基百科的多语言特性，进行了双语术语抽取研究[5]。同年，Le An Ha 将互学习机制，提高了双语术语抽取质量[6]。

（2）术语度计算相关研究。关于术语度（termhood）最早的定义，由 Kageura Kyo 和 Umino Bin 于 1996 年给出，他们将候选术语的术语度定义为：“候选术语与一个特定领域概念的相关程度[7]。术语度计算方法包括 TF*IDF[8] [9]、C-value/NC value[10]、类间分布熵 (inter-domain entropy，IDE)[11]、领域部件特征集（Domain Component Feature Set，DCFS）[12]等方法。Luning Ji 和 Qin Lu 等人曾利用简单核心术语抽取、切分后核心术语抽取以及类间核心术语抽取等方法，进行核心术语抽取的研究[13]。Takehito Utsuro、Mitsuhiro Kida、Masatsugu Tonoike 等人利用 Web 信息进行了术语的领域计算和分类研究[14] [15] [16]。

（3）其他相关研究。本文利用关键词作为候选术语，与此较相关的工作是 Masao Utiyama 和 Masaki Murata 等人所做的工作，他们将文档中原作者标注的关键词作为术语抽取训练集，在 NTCIR 测试集上取得 0.431 的精确率和 0.800 的召回率[17]。

与以往研究不同，本文结合关键词抽取技术和术语度计算方法，对文档进行关键词抽取，丰富关键词数量，在较大规模专业领域分类资源上，进行核心术语的识别和对齐。

3 基于专业领域平行语料的双语核心术语抽取

3.1 总体流程

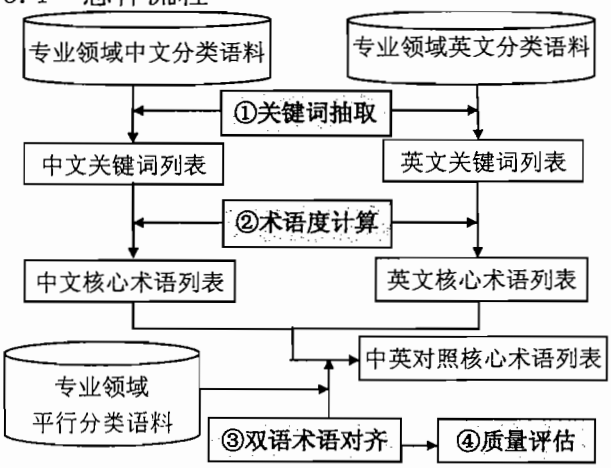

图 1　中英双语核心术语抽取流程图

如图 1 所示，基于专业领域平行语料的双语核心术语抽取过程包括五个方面：①获取中英文关键词。以真实的、较大规模专业领域中英文分类语料为基础，调用关键词抽取模块，抽取每篇文档中具有代表性的词语，作为每篇文档的关键词，将关键词作为该学科领域的候选术语。②获取中英文核心术语。根据专业领域语料计算候选术语的术语度，在一定术语度阈值的控制下，得到每个领域的核心术语列表。核心术语是利用词语在各

个类别中分布情况为基础，以类间分布熵作为计算依据，进行自动抽取。③获取中英文对照核心术语。在专业领域平行分类语料上，利用双语术语对齐技术自动生成中英文对照的核心术语。④结果评估。对每个领域的 Top-10、Top-50、Top-100、Top-200、Top-500 情况下的中英文核心术语进行人工检查，计算核心术语对齐的正确率。

3.2 关键技术

（1）关键词抽取。关键词抽取方法可以分为四类，即：基于统计的方法，该方法不需要复杂的训练过程，简单易行；基于语言学的方法，主要从词法分析、句法分析、语义分析及篇章分析等角度提高关键词提取质量；基于机器学习的方法，通过对训练数据进行训练获得统计参数，进行样本的关键词抽取；混合方法，即上述方法的综合运用或集成一些启发式知识。本文采用机器学习方法进行关键词的自动抽取。基于机器学习的关键词抽取，就是将关键词抽取看成一种分类问题。文[18]利用一定规模的文档集进行关键词抽取，结果表明条件随机场模型在改善关键词提取的性能方面，要优于支持向量机、多元线性回归模型等其他机器学习方法。本文采用文[18]中的条件随机场模型抽取文档的关键词。

（2）术语度的计算。术语度计算的基本出发点为：术语在各个专业领域语料中的分布不均匀，而非术语则一般均匀分布于不同的专业领域语料中。本文根据文[11]中的方法计算术语的类间分布熵，术语 w_i 的类间分布熵通过式（1）计算得到。

$$IDE(w_i)=\sum_j P_{ij}\log P_{ij} \tag{1}$$

其中，IDE(w_i)为术语 w_i 的类间分布熵。P_{ij} 为术语 w_i 在类别 j 上的出现概率，

$$P_{ij}=\frac{f_{ij}}{\sum_j f_{ij}} \tag{2}$$

其中，f_{ij} 为术语 w_i 的归一化词频，$f_{ij}=\frac{n_{ij}}{N_j}$，$N_j=\sum_i n_{ij}$。再根据文[11]中的方法，计算术语 w_i 在领域类别 j 上的权重，计算方式如式（3），

$$W_{ij}=n_{ij}\times\log_2\left[\frac{N}{Nd_i}\right] \tag{3}$$

其中，$Nd_i=2^{IDE(wi)}$，N 为领域类别总数。在每个专业领域中，依据术语的权重进行排序，取前 K 个（Top-K）权重最大的术语作为该领域的核心术语。

（3）双语核心术语对齐。以双语核心术语在双语平行语料（标题对齐、文摘对齐）上的共现信息为基础，进行双语核心术语的相关度计算，获取双语核心术语对齐信息。

表 1 双语术语出现频次联立表

	英文词 E 出现	英文词 E 不出现	
中文词 C 出现	*a*	*b*	*a+b*
中文词 C 不出现	*c*	*d*	*c+d*
	a+c	*b+d*	*N=a+b+c+d*

给定中文术语 C 和英文术语 E，双语核心术语 C 与 E 的相关度计算主要方法有 MI、Dice、χ^2 统计值以及 LogL 值等。按照表 1 所示，以上方法的计算公式分别如下。

$$MI(C,E)= \log2\ (N*a/\ ((a+b)*(a+c))) \quad (4)$$

$$Dice(C,E)=2*a/((a+b)*(a+c)) \quad (5)$$

$$\chi^2(C,E)= N*(a*d-b*c)/((a+b)*(a+c)*(b+d)*(c+d)) \quad (6)$$

$$LogL(C,E)=2*(a*\log2(a*N/((a+b)*(a+c))+ b*\log2(b*N/((a+b)*(b+d))+ c*\log2(c*N/((c+d)*(a+c))+ d*\log2(d*N/((c+d)*(b+d))) \quad (7)$$

已有的研究表明， LogL 方法方法可以处理其他相关度计算方法无法计算的低频二元组的关联强度[19]。因此本文采用 LogL 方法来进行中英文双语核心术语的对齐。

4 实验结果与分析

4.1 训练数据说明

表 2 学科分类领域语料分布情况

类别标记	类别	数量	比率	类别标记	类别	数量	比率
9	法律类	23813	5.16%	N	自然科学	1107	0.24%
A	马列主义、毛泽东思想	2548	0.55%	O	数学、物理、化学	826	0.18%
B	哲学	16643	3.61%	P	天文、地球科学	512	0.11%
C	社会科学	11189	2.43%	Q	生物科学	254	0.06%
D	政治	24508	5.32%	R	医药卫生	2070	0.45%
E	军事	519	0.11%	S	农业科学	756	0.16%
F	经济	128320	27.83%	T	工业科学	4929	1.07%
G	文化、科学、教育、体育	106442	23.09%	U	交通	76	0.02%
H	语言文字	10919	2.37%	V	航空航天	15	0.00%
I	文学	29973	6.50%	X	环境科学	1445	0.31%
J	艺术	11816	2.56%	Z	综合类	19	0.00%
K	历史地理	26972	5.85%		平均数量		17638

表 3 学科分类领域平行语料分布情况

类别标记	数量	比率	类别标记	数量	比率
9	7061	6.49%	*N*	732	0.67%
A	387	0.36%	*O*	373	0.34%
B	6516	5.99%	*P*	375	0.34%
C	3518	3.23%	*Q*	162	0.15%
D	4831	4.44%	*R*	1122	1.03%
E	105	0.10%	*S*	459	0.42%
F	31910	29.32%	*T*	1783	1.64%
G	28176	25.89%	*U*	46	0.04%
H	5864	5.39%	*V*	10	0.01%
I	6654	6.11%	*X*	924	0.85%
J	1277	1.17%	*Z*	6	0.01%
K	6557	6.02%	平均数量		4733

注： 本文将 D9 类别单独提取出来作为法律类，类别标识为 9。

本文采集了专业领域文献语料作为训练集，将学术论文作为专业领域文献语料，进行术语度计算和双语核心术语抽取和对齐。将中文数据部分（即包括中文标题、中文文摘、中文关键词三个部分）构成中文专业领域语料，包含46万余篇中文题录信息的记录。同理，得到英文中文专业领域语料，包含13万余篇英文题录信息的记录。中英文专业领域语料包含23个类别，每个类别包含的平均文档数为17638，具体分布情况如表2所示。由表2可以看出，分类语料是非均衡语料，例如经济类（标识为F）和文化、科学、教育、体育类（标识为G）的比率较高，两者之和达整个语料的50%左右，而个别类别多包含的文档极少，如综合类（表示为Z）。采集的专业领域平行分类语料的类别分布情况如表3所示，平均每个类别包含平行记录4733条，对齐方式为：中英文标题为句子级对齐、中英文文摘为段落级对齐，中英文关键词串可以进一步加工为词语级对齐形式。本文利用关键词抽取技术，给每个文档进行了关键词的补充或重新选择，将这些关键词作为候选术语。

4.2 抽取结果与分析

利用术语度计算方法分别对中英文专业领域分类语料进行了中英文候选术语的术语度计算，经过一定的阈值控制和人工检查，最终得到每个领域的核心词汇。

依据专业领域平行分类语料，通过双语核心术语对齐技术，最终得到每个领域的中英

文对照的核心术语。对 26 个专业领域的 Top-10、Top-50、Top-100、Top-200、Top-500 情况下的中英文核心术语进行人工检查，计算核心术语对齐的正确率。表 4 给出了双语核心术语对齐正确率分布情况。从表 4 可看出，每个领域的前 10 对核心术语平均正确率达 70%，前 200 对中英文对照核心术语的平均正确率在 50%以上，前 500 对的平均正确率为 45%左右。在 26 个类别中，类别 F 术语对齐正确率较高，前 200 对核心术语平均正确率在 81%，正确率较高的类还包括 G、B、I 等类别，而类别 V、X、Z 等的正确率则比较低。将表 4 与表 2、3 进行对照可以看出，领域的训练样本数对该领域的双语核心术语对齐有较大影响，样本数高，则双语核心术语对齐的质量较高，反之亦然。

表 4　双语核心术语对齐正确率

	9	A	B	C	D	E	F	G	H	I	J	K
Top-10	9	8	2	7	9	2	9	9	8	6	4	7
	90.00%	80.00%	20.00%	70.00%	**90.00%**	20.00%	**90.00%**	**90.00%**	80.00%	60.00%	40.00%	70.00%
Top-50	33	26	36	37	35	13	47	40	37	38	21	28
	66.00%	52.00%	67.00%	70.00%	65.00%	26.00%	**94.00%**	**80.00%**	74.00%	**76.00%**	42.00%	56.00%
Top-100	69	52	67	70	65	25	91	77	74	70	43	57
	69.00%	52.00%	67.00%	70.00%	65.00%	25.00%	**91.00%**	**77.00%**	**74.00%**	70.00%	43.00%	57.00%
Top-200	126	82	126	126	122	-	162	136	127	123	77	113
	63.00%	41.00%	63.00%	63.00%	61.00%	-	**81.00%**	**68.00%**	**63.50%**	61.50%	38.50%	56.50%
Top-500	214	-	238	197	223	-	291	234	233	216	-	194
	42.80%	-	**47.60%**	39.40%	44.60%	-	**58.20%**	**46.80%**	46.60%	43.20%	-	38.80%

	N	O	P	Q	R	S	T	U	V	X	Z	Avg
Top-10	6	6	6	4	4	4	6	6	1	3	2	5.5652
	60.00%	60.00%	60.00%	40.00%	40.00%	40.00%	60.00%	60.00%	10.00%	30.%	20.00%	55.65%
Top-50	21	18	22	13	22	23	28	15	-	27	-	26.3636
	42.00%	36.00%	44.00%	26.00%	44.00%	426%	56.00%	30.00%	-	54.00%	-	52.73%
Top-100	43	31	39	24	40	34	46	23	-	54	-	49.7273
	43.00%	31.00%	39.00%	24.00%	40.00%	34.00%	46.00%	23.00%	-	54.00%	-	49.73%
Top-200	83	-	62	-	-	57	-	-	-	89	-	107.4
	41.50%	-	31.00%	-	-	28.50%	-	-	-	44.50%	-	53.70%
Top-500	-	-	-	-	-	-	-	-	-	-	-	226.6667
	-	-	-	-	-	-	-	-	-	-	-	45.33%

我们对结果检查发现，有些具有包含关系的核心术语需要过滤，如 J 类中出现具有包含关系的核心术语对齐结果：（*音乐教育　music education*）与（*音乐　music*）、文化、G 类包含对齐结果如：（*网络　network*）与（*网络环境　network environment*）。这说明核心术语提取阶段，仅仅通过分布熵进行术语度计算还存在一些不足的地方。而 C-value/NC value 方法在解决术语间包含关系问题上具有很好的效果。因此，本文今后拟综合分布熵与 C-value/NC value 方法，进行领域核心术语的抽取。

5　结论与未来工作

本文将专业领域文档的关键词作为候选核心术语，利用中文和英文的专业领域分类语料，通过关键词抽取、术语度计算等关键技术，分别进行中文和英文的核心术语的识别。接着，以中英文专业领域平行语料为基础，通过双语对齐技术，自动生成中英文对照的双语核心术语列表。实验结果表明，每个专业领域中，前 200 对中英文对照核心术语的平均正确率在 50%以上，个别领域正确率达 80%左右，通过该方法可以快速获取每个专业领域的中英文对照核心术语，从而提高专业领域双语识别性能。本文下一步的工作包括：采集相对均衡的大规模专业领域语料作为训练集，提高每个领域的核心术语抽取的正确率与双语核心术语对齐的正确率；将分布熵方法与 C-value/NC value 方法结合起来，进行领域核心术语的抽取；借助双语短语分析方法提高中英文关键词抽取的质量与核心短语识别的正确率；利用对齐质量更好的双语术语对齐方法，提高双语核心术语对齐的正确率。

参 考 文 献

[1] 冯志伟. 一个新兴的术语学科——计算术语学. 术语标准化与信息技术, 2008,4: 4-9.

[2] 孙乐, 金友兵, 杜林. 平行语料库中双语术语词典的自动抽取. 中文信息学报， 2000，14 (06): 33-39.

[3] Bèatrice Daille, Emmanuel Morin. French-English Terminology Extraction from Comparable Corpora. In: Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-05), 2005: 707-718.

[4] 张永臣, 孙乐, 李飞, 李文波, 西野文人, 于浩, 方高林. 基于Web数据的特定领域双语词典抽取. 中文信息学报, 2006, 20 (02): 16-23.

[5] Maike Erdmann, Kotaro Nakayama, Takahiro Hara and Shojiro Nishio. Extraction of Bilingual Terminology from a Multilingual Web-based Encyclopedia. Journal of Information Processing, 2008, 16: 68-79.

[6] Le An Ha, Gabriela Fernandez, Ruslan Mitkov and Gloria Corpas. Mutual Bilingual Terminology Extraction. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 2008: 28-30.

[7] Kageura Kyo,Umino Bin. Methods of automatic term recognition: a review. Terminology, 1996, 3 (2): 259-289.

[8] Kiyotaka Uchimoto, Satoshi Sekine, Masaki Murata, Hiromi Ozaku and Hitoshi Isahara. Term Recognition by Using Different Field Corpora. In: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, 1999: 443-450.

[9] Yirong Chen, Qin Lu, Wenjie Li, Zhifang Sui and Luning Ji. A Study on Terminology Extraction Based on Classified Corpora. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), 2006: 2383-2386.

[10] Katerina T. Frantzi, Sophia Ananiadou and Junichi Tsujii. Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries, 2000, 3 (2): 115-130.

[11] Jing-Shin Chang. Domain Specific Word Extraction from Hierarchical Web Documents: A First Step toward Building Lexicon Trees from Web Corpora. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, 2005: 64-71.

[12] Qinlong Zhang, Qin Lu, Zhifang Sui. Measuring Termhood in Automatic Terminology Extraction. In: Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2007), 2007: 328-335.

[13] Luning Ji, Qin Lu, Wenjie Li, Yirong Chen. Automatic Construction of a Core Lexicon for Specific Domain. In: Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007), 2007: 183-188.

[14] Mitsuhiro Kida, Masatsugu Tonoike, Takehito Utsuro, Satoshi Sato. Domain Classification of Technical Terms Using the Web. Systems and Computers in Japan, 2007, 38(14): 11-19.

[15] Takehito Utsuro, Mitsuhiro Kida, Masatsugu Tonoike, Satoshi Sato. Towards Automatic Domain Classification of Technical Terms: Estimating Domain Specificity of a Term Using the Web. In: Proceedings of Asia Information Retrieval Symposium (AIRS 2006), 2006: 633-641.

[16] Takehito Utsuro, Mitsuhiro Kida, Masatsugu Tonoike, Satoshi Sato. Collecting Novel Technical Terms from the Web by Estimating Domain Specificity of a Term. In: Proceedings of the 21st International Conference on Computer Processing of Oriental Languages (ICCPOL 2006), 2006: 173-180.

[17] Masao Utiyama, Masaki Murata and Hitoshi Isahara. Using Author Keywords for Automatic Term Recognition. Terminology, 2000, 6 (2): 313-326.

[18] Zhang Chengzhi, Wang Huilin, Yao Liu, Wu Dan, et.al. Automatic Keyword Extraction from Documents Using Conditional Random Fields. Journal of Computational Information Systems, 2008, 4(3): 1169-1180.

[19] Dunning T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 1993, 19 (1): 61-74.