文章编号：1004-4353(2017)03-0259-05

基于术语自动抽取的科技文献翻译辅助系统的设计

黄政豪， 崔荣一*

（延边大学工学院 计算机科学与技术学科 智能信息处理研究室，吉林 延吉 133002）

摘要：设计了一种中韩科技信息综合平台中的翻译辅助系统.首先，依据关键词确定的组词特征获取候选术语，并使用互信息评估候选术语以实现术语自动提取.其次，将已有术语、抽取到的新术语、术语译文和历史翻译记录等信息存储到系统数据库中建立术语库.最后，设计翻译工作者的用户接口，使其通过该接口获取已有术语的译文信息、新术语的相似译文信息和译文记忆库为基础的历史翻译数据.测试结果表明，本文设计的术语自动抽取功能和辅助译文生成功能达到了预定的设计目标，术语自动抽取算法召回率达到61.8%，结合优化方法进行优化后达到 66.9%；辅助译文生成平均延时为 0.031 s，MRR 为 0.951，测试结果满足用户需求.

关键词：术语自动识别；术语抽取；翻译辅助系统

中图分类号：TP391.41 **文献标识码**：A

Design of translation assistant system based on automatic extraction of terms

HUANG Zhenghao， CUI Rongyi*

（*Intelligent Information Processing Lab.*, *Dept. of Computer Science & Technology*, *College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*）

Abstract: This paper describes the design method of Chinese and Korean science and technology information aided translation system. Firstly, extracting candidate terms based on word formation characteristics of keywords, and using mutual information to evaluate candidate terms for automatic term extraction. Secondly, existing terminology, extraction of the new terminology, terminology translation and history of translation records and so on are stored in the system database and the established terminology database. Finally, design the user interface for translators, so the translators can obtain the translation of exiting terminology, the similar translation of additional terms, and the history translation data based on translation memory through this interface. The results of the system test show that automatic extraction of term and the auxiliary translation function reach the desired goals. The recall rate or term automatic extraction algorithm is 61.8%, and after optimization the rate is improved by optimization method to reach 66.9%. The generation of auxiliary translation averagely delays 0.031 seconds, and the MRR is 0.951, so the test results fulfil the users' needs.

Keywords: automatic term recognition; term extraction; computer aided translation

计算机翻译辅助(Computer Aided Translation，CAT)是在机器翻译(Machine Translation，MT)基础上演变而来的[1].它能够为翻译个人或翻译团队建立可以反复使用的翻译记忆库和搜索工具，使翻译工作者能够在短时间内快速地完成翻译工作.计算机翻译辅助系统主要涉及 3 个关键技

收稿日期：2017-04-19 * **通信作者**：崔荣一(1962—)，男，博士，教授，研究方向为模式识别、智能计算.

基金项目：吉林省自然科学基金资助项目(20140101186JC)；延边大学-延边州科技信息服务中心合作项目(延大科合字[2016]1 号)

术[2],即辅助译文生成、译后编辑和系统反馈学习.在辅助译文生成策略研究中,Bowker[3]提出翻译记忆是一种用于储存原文文本信息和对应的目标语言译文的语言数据库,其工作原理是:存储和记录翻译工作者已翻译完成的目标语言译文及其对应的原文文本,在后续翻译中如果出现与数据库中相同或相似的原文内容,系统将自动搜索与其相同或相似的目标语言译文信息并提供给翻译工作者.在译后编辑研究中,冯全功等[4]认为译后编辑是指根据一定的目的对机器翻译的原始产出进行加工与修改的过程.在译后编辑工具设计上罗季美等[5]对机器译文与人工译文进行了对比研究,描述了机器翻译在词汇、句法、符号等方面表现出的典型错误形式,并通过补充建立形式化规则为机器翻译系统提供反馈.在系统反馈学习研究中,黄河燕等[6]描述了一个智能译后编辑系统的设计原理和实现算法,该系统将一段作为编辑处理的基本单位,并且可以形成编辑反馈信息提供给知识处理模块.基于上述研究,本文设计了一种基于术语自动抽取的科技文献翻译辅助系统,即具有领域术语自动抽取、管理功能的计算机翻译辅助系统,并通过测试验证了本文方法的有效性.

1 结构特征与互信息相结合的术语自动抽取方法

首先根据已收集的术语语料进行分词并分析词性组合,提出符合该领域的词性规则模板;其次对测试语料进行分词,通过词项顺序与词性规则模板匹配获得候选术语;最后通过计算相邻词项之间的互信息评估稳固程度和设定的阈值分类出术语.其中,术语语料库和测试语料库的分词利用NLPIR 汉语分词系统实现.

1.1 术语语料库

冯志伟在《现代术语学引论》[7]中将术语定义为通过语音或文字来表达或限定专业概念的约定性符号.在科技领域等专业性较强的文章中术语的出现频率较高,尤其文章作者所设定的关键词都是最能体现文章主题和领域特征的标准术语[8].根据这一特点,本文对收集到的航空领域科技文献的关键词进行筛选后获得 17 330 条具有较高的航空领域特征的关键词.本文利用这些关键词作为术语语料,通过对其进行分析归纳出该领域术语的结构特征.

1.2 术语结构特征

文献[7]提出术语可以是词也可以是词组,例如,"固体燃料发动机"是由"固体""燃料""发动机"3 个名词词项组成,这一特征能很好地确定组成术语的词项长度范围,能够更好地确定领域内组成术语的词项长度规则.为了分析词项长度,首先对术语语料库中的 17 330 条关键词进行分词,然后统计组成每个术语的词项数量,最终确定其数量范围在 1～10 之内.统计结果如图 1 所示.

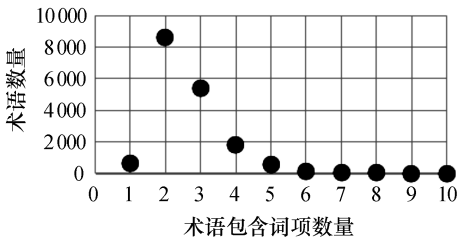

图 1 组成术语的词项数量

由图 1 可知,大多数术语包含的词项数分布在 1～6 范围内,占整个术语语料库的 99.6%.张榕[9]在研究中统计出组成术语的词项长度一般为 2、3、4 个,共占总数的 71.723%,大部分术语长度都在 1～6 之间,这与本文研究统计结果一致.但本研究中主要研究词组型术语,所以将使用术语长度确定为 2～6,占术语总数的 95.85%.

除词项数规则外,术语在构成术语的词性组合上也存在着一定规则,如"名词＋名词"短语在术语语料库中包含 2 332 个,占术语语料库的 13.5%.术语语料中的所有术语的词性组合前 3 种统计结果如表 1 所示.

表 1 词性组合统计

词性组合	数量	实例
名词＋名词	2 332	光学卫星、数字滤波器
动词＋名词	2 010	减速伞、克隆平台
名词＋动词	1 885	图像检索、路径优化

李丽双等[10]在抽取汽车术语时制定了一套词性规则,本文在这一词性规则基础上结合术语语料库中航空领域术语的特点对其进行了改进,改进后的规则如下:①所有位置不包含叹词、处所词、状态词、代词、除"—"以外的标点符号.②开头不包含助词、连词、后缀、量词、标点符号.③结

尾不包含助词、连词、前缀、方位词、标点符号.

1.3　互信息

互信息(Mutual Information)是信息论中信息度量方法，它可以看成是一个随机变量中包含的关于另一个随机变量的信息量，即是一个随机变量由于已知另一个随机变量而减少的不确定性[11]；因此，可以通过互信息评价相邻词项在语料中结合的稳固程度. 互信息的计算公式为

$$I(x;y)=\lg\frac{P(xy)}{P(x)P(y)},\tag{1}$$

其中：x 和 y 分别是相邻的两个词项；$P(xy)$ 表示 x,y 在语料中相邻出现的概率；$P(x)$ 和 $P(y)$ 分别是 x,y 在语料中出现的概率.

1.4　术语抽取算法

利用上述词项数量、词性组合和互信息等方法，本文提出术语自动抽取算法，其具体步骤如下：

Step 1　利用 NLPIR 中文分词系统对术语语料 T 中的每个术语 t_i 进行分词和词性标注，以此获得每个术语的词性组合集合 S，然后再根据词性规则 R 删除不符合的词性组合获得 S'；

Step 2　对测试语料 D 中的每个摘要 d_j 进行分词和词性标注，以此获得词项序列 $w_1\cdots w_k$ 和词性序列 $s_1\cdots s_k$；

Step 3　选第 k 个词性 s_k，开始在 S' 中查找是否存在所选词性开头的词性组合；

Step 4　如果不存在选下一个词项，执行 Step 3 操作；

Step 5　如果存在，则选 s_ks_{k+1} 作为新的词性组合并继续在 S' 中查找，直到获得 $s_ks_{k+1}\cdots s_{k+n}$，其中 $1\leqslant n\leqslant 5$；

Step 6　根据式(1)计算 s_ks_{k+1} 对应的词项组合 w_kw_{k+1} 的互信息；

Step 7　如果计算到 s_{k+n} 时互信息小于阈值，则选择 $w_kw_{k+1}\cdots w_{k+n-1}$ 为术语；

Step 8　如果 s_{k+n} 为止互信息全部大于阈值，则选择 $w_kw_{k+1}\cdots w_{k+n}$ 为术语；

Step 9　选下一个词项执行 Step 3 操作，直到完成语料库中的所有词项的计算.

1.5　优化提取结果

通过术语提取算法获取到的术语中包含一些结构不完整的词组. 例如，抽取到“动力学模型”，而语料中出现的是“无人机动力学模型”，这类问题主要是因为部分术语中包含 1 个或多个变化的词项. 为了解决这类问题，先假设术语语料库中存在类似结构术语“车体动力学模型”，然后分别对此 2 个词组进行分词得到“动力学 /n 模型 /n”“车体 /n 动力学 /n 模型 /n”，再以词项为比较对象计算编辑距离，最后得出“动力学模型”左侧添加词项“车体 /n”，词性为名词，而测试语料中“动力学模型”左侧也是名词“无人机”. 根据这个特点，将“无人机”合并到“动力学模型”形成一个新的术语. 此方法可以进一步完善术语抽取质量，但需要使用术语语料库作为对比库，所以术语语料库的大小直接影响着优化效果.

2　辅助译文生成与系统管理

翻译辅助系统是面向翻译工作者和术语管理者的主要操作接口，也是连接用户界面和术语自动抽取系统模块的功能核心. 翻译辅助系统主要包含辅助译文生成模块、用户反馈模块、术语管理模块和翻译记忆管理模块.

2.1　辅助译文生成

辅助译文生成模块作为翻译辅助系统的主要功能模块，实现面向翻译工作者的术语标记、术语翻译、相似翻译反馈和翻译记忆提示等功能，是翻译工作者在翻译时所能使用到的主要接口. 其具体工作流程(图 2)如下：

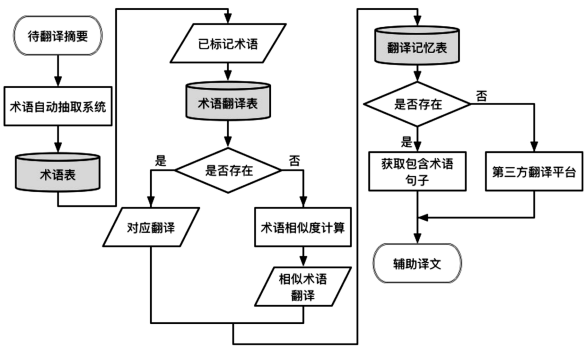

图 2　翻译辅助系统流程图

1)点击翻译时，待翻译摘要进入术语自动抽取系统将包含的术语用高亮方式进行标记；

2)对每一个标记术语进行术语翻译表搜索，如果存在已翻译结果则直接返回对应翻译内容(如果有多个，则全部返回)，如果不存在对应翻译则采用术语相似度计算找出最接近的术语并返回其翻译结果；

3)通过翻译记忆表获取历史翻译过的包含所查询术语的句子并返回,如果不存在历史翻译数据,则借助第三方翻译平台(本文中采用百度翻译接口)获取对应翻译结果并返回;

4)整理最终生成的辅助译文并显示到用户界面中.

经过上述操作,翻译工作者在翻译摘要时能够获得多数已翻译术语和相似历史翻译参考句子,无需频繁切换不同翻译工具,可有效地提高翻译效率.

2.2 相似术语查找方法

如果术语自动抽取系统抽取到的新术语不存在于术语语料库中,则无法获取到精确的翻译对.为了解决这类问题,本文从计算效率角度考虑,采用编辑距离作为两个术语相似度计算算法.编辑距离[12](levenshtein distance)的意义是一个字符串 S 修改为另一个字符串 S' 所需要的最小编辑操作次数,这些操作包含插入、删除、修改3种操作.编辑距离能够计算两个术语之间的差异,有助于快速获取待翻译术语的相似术语.根据本文中术语是由多个词项组成这一特征,计算编辑距离时也将采用通过分词得到的词项为单位进行编辑距离计算.

2.3 第三方翻译平台

对于新录入的术语本系统无法提供任何翻译内容,所以遇到这类术语时本系统将借助第三方翻译平台接口提供翻译内容.本系统中采用百度翻译模块进行辅助翻译.百度翻译支持多个语言对之间的文本和网页翻译,只需输入想要翻译的文本或者网页地址,即可轻松获得对应语言的翻译结果.

2.4 系统管理模块

术语管理模块主要对术语数据库和翻译辅助系统中的参数和数据进行管理,该模块主要针对系统管理者,翻译工作者无法使用.其主要功能是:①术语管理.可对已录入术语进行修改操作,录入新的术语,对在术语自动抽取系统中抽取到的数据进行审核并追加到库操作.②术语翻译管理.包含已翻译术语修改功能、新录入术语的翻译功能,还可以获取到用户通过反馈模块反馈的针对某个术语的翻译推荐或者错误指正等信息,以便于完善术语翻译质量.③翻译记忆管理.可完成包含术语的摘要编号和对应历史翻译句子进行数据修改操作,设置手动或自动方式的相似度计算操作.

2.5 用户反馈模块

翻译工作者在翻译工作中可以随时利用该模块进行问题反馈,让系统管理员及时了解系统存在的问题,尽可能避免后续大规模问题的发生.其主要功能是:①术语反馈.可快速反馈当前翻译术语的翻译错误,同时也可以反馈用户的翻译建议.②翻译记忆反馈.可反馈翻译实例句子中的问题和排序问题.③系统问题反馈.可反馈在操作中遇到的所有问题,以便进行系统更新,防止后续一系列未知错误发生.

3 测试及结果分析

3.1 评价指标

术语自动抽取系统采用标准评价指标:准确率、召回率、F 值.辅助译文生成模块采用准确率、搜索时间和MRR作为评价标准[13].

准确率 P 是抽取到的正确术语数 A 与抽取到的术语总数 B 的比值,衡量的是系统的查准率;召回率 R 是指抽取到的正确术语数 A 和测试语料中包含的所有术语数 C 的比值,衡量的是系统的查全率.F 值(F-Measure)是准确率和召回率的加权调和平均,是一种综合评价指标,具体公式如下:

$$F=\frac{2\times P\times R}{P+R}. \tag{2}$$

MRR是一个国际上通用的对搜索算法进行评价的指标,中文翻译为首现正确排序倒数,即首个正确结果出现的位置,即第 n 个记录是正确的,则MRR分数为 $1/n$;如果没有匹配的句子,则MRR分数为0.

3.2 术语自动抽取测试

根据本文实验流程,对测试数据分别通过词性组合、互信息以及两种方法相结合的方式进行比较,并采用正确率、召回率和 F 值对实验数据进行评测,测试结果如表2所示.由表2可以看出,词性组合和互信息相结合的方式比单独使用其中一种方法在性能上有一定提高,特别是抽取术语总数上有着较为显著的变化,但准确率和召回率并不是很高.优化提取结果,同时与文献[10]方法进行比较,结果如表3所示.由表3可以看

出，通过优化提取结果能够提高准确率和召回率，抽取效率比文献[10]方法有较大程度的提高.

表 2　术语抽取测试结果

测试方法	准确率/%	召回率/%	F 值/%
词性组合	27.4	48.6	35.1
互信息	23.9	49.9	32.4
词性组合＋互信息	57.2	61.9	59.4

表 3　优化提取结果后的数据

测试方法	准确率/%	召回率/%	F 值/%
文献[10]方法	47.0	34.7	37.5
词性组合＋互信息＋优化方法	61.2	66.9	63.8

3.3　辅助译文生成测试

本测试采用选取术语的方式进行. 在术语数据库中随机抽取 1 000 条出现频率大于 10 次的术语作为测试样本进行测试，并利用前 10 个抽取到的历史翻译句子进行 MRR 和准确率计算. 为了自动化完成测试，该测试过程未在客户端上进行，而是单独编写了测试程序. 测试结果如表 4 所示.

表 4　辅助译文检索测试结果

术语数量	平均用时/s	平均 MRR	平均准确率/%
1 000	0.031	0.951	92.17

根据表 4 可知，平均检索用时为 0.031 s，平均 MRR 为 0.951，平均正确率为 92.17%. 1 000 条随机术语中有些术语在单一个句子中出现多次，而包含的翻译句子少于 10 个，所以一定程度上影响到了平均准确率.

由图 3 可以看出，随着术语包含词项数量增加，查询所需要的耗时也会增加. 由于本系统中采用的术语长度只有 2～6 个词项，所以不会因术语长度而大幅增加查询时间.

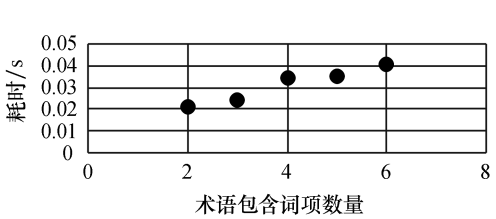

图 3　不同长度术语查询耗时

4　结论

本文从实际工程的角度设计并实现了基于术语自动抽取的科技文献翻译辅助系统，功能上完成了术语自动抽取和辅助译文生成模块. 测试结果表明，本文设计的术语自动抽取功能和辅助译文生成功能达到了预定的设计目标，术语自动抽取算法召回率达到 61.8%，结合优化方法进行优化后达到 66.9%，提高了 5.1%；辅助译文生成平均用时为 0.031 s，MRR 为 0.951，测试结果满足用户需求. 由于本系统的术语自动抽取算法受到测试语料和术语语料规模的限制，使得部分低频术语无法被正确地抽取出来. 另外，抽取规则是根据特定领域语料制定的，无法直接适用于其他领域. 针对以上问题，在今后将研究更有效的术语抽取算法，提高抽取效率，尽可能地降低系统中人工干预频率.

参考文献：

[1] 朱玉彬，陈晓倩. 国内外四种常见计算机辅助翻译软件比较研究[J]. 外语电化教学，2013，149(1)：69-75.

[2] 叶娜，张桂平，韩亚冬，等. 从计算机辅助翻译到协同翻译[J]. 中文信息学报，2012，26(6)：1-10.

[3] Bowker L. Computer-aided translation technology: a practical introduction[J]. Linguistics, 2007, 8(2)：229-231.

[4] 冯全功，崔启亮. 译后编辑研究：焦点透析与发展趋势[J]. 上海翻译，2016(6)：67-89.

[5] 罗季美，李梅. 机器翻译译文错误分析[J]. 中国翻译，2012(5)：84-89.

[6] 黄河燕，陈肇雄. 一种智能译后编辑器的设计及其实现算法[J]. 软件学报，1995(3)：129-134.

[7] 冯志伟. 现代术语学引论[M]. 北京：语文出版社，1997.

[8] 周浪，张亮，冯冲，等. 基于词频分布变化统计的术语抽取方法[J]. 计算机科学，2009，36(5)：177-180.

[9] 张榕. 术语定义抽取、聚类与术语识别研究[D]. 北京：北京语言大学，2006：76-77.

[10] 李丽双，王意文，黄德根，等. 基于信息熵和词频分布变化的术语抽取研究[J]. 中文信息学报，2015，29(1)：82-87.

[11] 傅祖芸. 信息论[M]. 2 版. 北京：北京电子工业出版社，2010.

[12] Levenshtein V I. Binary codes capable of correcting deletions, insertions and reversals[J]. Problems of Information Transmission, 1965, 1(1)：8-17.

[13] 费巍. 搜索引擎检索功能的性能评价研究[D]. 武汉：武汉大学，2010：11-12.