一种基于Web的术语翻译获取及验证方法

张 晶[1] 曹存根[2] 王 石[2]
(江苏科技大学计算机科学与工程学院 镇江 212003)[1] (中国科学院计算技术研究所 北京 100190)[2]

摘 要 中文术语及未登录词的翻译是机器翻译、跨语言检索中的一个重要问题,这些翻译很难从现有的词典中获取。提出了一种通过搜索引擎从网页中自动获取中文术语英文翻译的方法。通过术语的部分翻译信息,构造出3种查询项模式,提出了多特征的翻译抽取方法。针对传统方法结果准确率不高、候选翻译干扰项多的问题,提出端类比对齐验证、双语对齐度验证、构词法验证3种验证模型来对候选翻译进行有效验证。实验结果表明,获取的双语翻译对准确率高,TOP1的准确率达到97.4%,TOP3的准确率达到98.3%。

关键词 中文术语,网络挖掘,术语翻译,信息检索
中图法分类号 TP391 **文献标识码** A

Web-based Term Translation Extraction and Verification Method

ZHANG Jing[1] CAO Cun-gen[2] WANG Shi[2]
(School of Computer Science and Engineering, Jiangsu University, Zhenjiang 212003, China)[1]
(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)[2]

Abstract Extracting translation of Chinese term and oov words is an important problem in many applications such as machine translation and cross-language information retrieval. However, these translations are difficult to access from traditional dictionary. This paper proposed a method to automatically extract English translation of Chinese term from Web by search engine. It constructs three query modes using the partial translation of terms, proposes multi-featured extraction method. Aim at the problem of low accuracy and unrelated translations, it verifies the candidate translations by three verification modes: side analog alignment verification, Bilingual alignment degree verification and word-building verification. Experimental results show that TOP1 accuracy arrives 97.4% and TOP3 98.3%.

Keywords Chinese term, Web mining, Term translation, Information retrieval

1 引言

目前双语词典被广泛应用于语言翻译中。双语词典具有查询方便、准确性高的特点。然而,词典词汇量有限,许多词语无法在词典中查询。这样的问题被称为未登录词(OOV, Out of Vocabulary)问题。由于双语词典的特点,使得未登录词问题经常发生在翻译复合词(如专有名词、名词词组或者新术语)的时候。

网络上有很多包含双语信息的页面,即中文术语及其翻译同时出现在页面中。当人们在翻译文章时,遇到无法在词典中找到翻译的词,一些有经验的用户会尝试使用搜索引擎,但是大量无关重复的信息会妨碍他们获取有用的信息。于是,一种基于网络的自动翻译获取系统是十分必要的。

基于网络的翻译获取系统应用范围很广。用户在撰写或阅读文章时,经常会遇到词典中没有收录的词汇,该系统能帮助他们从网络中挖掘准确及地道的翻译,同时可以为编纂词典提供翻译候选,还能评价并排序词典中的候选列表[1-3]。

从网络中获取双语翻译的难点在于网络中信息量大,信息来源广,需要从海量网页资源中获取相关网页。由于网页是一种非结构化的文本,因此从网页中抽取出术语翻译难度较大,准确性不高。本文研究了一种有效地从网页中获取专业术语翻译的方法,可以有效弥补双语词典的不足。通过构造适当的查询项模式访问搜索引擎,利用多特征的抽取方法从网页中获取术语的翻译,并采用多种验证模型排除干扰项。目前的自动获取翻译的研究中,主要是面向领域专业术语的翻译获取,本文研究将实验对象从领域专业术语扩展到一般性未登录词。实验结果证明,使用本方法可以有效地获取高质量的双语翻译对。

2 相关工作

基于搜索引擎的自动获取术语翻译在过去的文献中已有一定研究。一般使用搜索引擎在Web中搜索原术语,利用目标翻译的特征信息(如与源术语位置关系、频率、上下文特征、词性等)在搜索结果网页集中抽取一组目标语言的候选翻译项,构成候选翻译项集合;然后根据候选翻译评价模型,从候选翻译项集中选择最佳翻译项[4]。

到稿日期:2011-08-29 返修日期:2011-11-22 本文受国家自然科学基金(61035004)资助。
张 晶(1987—),男,硕士,主要研究方向为知识获取、中文信息处理;曹存根(1964—),男,研究员,博士生导师,主要研究方向为大规模知识获取与管理;王 石(1981—),男,博士,助理研究员,主要研究方向为知识获取、中文信息处理。

Zhang,Huang 提出利用搜索引擎的返回结果来获取双语翻译知识。他们采用启发式的方法构造查询词,交给搜索引擎。在返回结果中,利用统计方法获得对应翻译[5,6]。该方法的特点是翻译的召回率高,但需要构造大量的查询项,耗时多、运行效率不高。

方高林的汉英翻译获取系统将中文术语每个汉字的英文作为预测信息,进而将其作为源术语的扩展搜索词,以此搜索目标网页,从中抽取候选翻译项。然后使用词汇分布特征、长度比率与中文术语的距离、关键符号与边界信息等多种特征,对候选翻译项集进行排序[7]。而在一个基于 Web 的英中术语翻译系统中,方高林又利用后缀数组构造候选翻译,使用子集冗余和词缀冗余两种方法来解决噪音干扰问题,最后基于互信息方法从候选翻译集中选择最佳翻译项[8]。方高林的方法取得了较好的实验结果,然而对候选翻译的验证方面仍有不足,干扰的翻译项较多、TOP1 准确率不高、实用性不强。

目前基于网络的双语翻译获取方面虽然取得了一定进展,但我们认为,仍有许多问题没有解决。尽管网络资源已非常丰富,但仍需考虑充分、高效利用网络资源,以挖掘更多、更准确的翻译。此外,目前的自动翻译系统对候选翻译的验证方面研究较少,方法仍比较单一,不能有效地解决候选翻译中包含错误翻译和干扰项的问题。本文针对已有研究的不足,首先通过术语的部分翻译信息构造出 3 种查询项模式,对获取到的搜索结果利用多特征的抽取方法获得候选翻译。为了排除翻译中的错误和干扰项,提出了端类比对齐验证、双语对齐度验证、构词法验证 3 种验证方法从候选翻译中过滤掉无关的干扰项。实验中 TOP1 准确率达到 97.4%,有效提高了双语翻译对获取的正确率。

3 基于 Web 的术语翻译抽取

3.1 基于多查询项的相关网页检索

搜索引擎会检索大量的网页。对于每个查询项,搜索引擎 Google 最多只返回 1000 条结果。为了获取更多的翻译项,需要构造合适的查询项,提高返回结果的相关性。

借鉴 Fang 等提出的基于语义预测构建查询项的方法,利用术语的部分翻译作为原术语的扩展搜索词,以此来进行检索。术语的部分翻译是指中文术语经过分词处理后分割开的部分英文翻译。该方法可以有效提高返回结果的相关性[7]。利用吕学强等提出的单词级最小求交对齐的方法,对一部手工收集的汉英术语词典进行单词级最小求交对齐,最后得出汉英子串翻译词典,用于获取术语的部分翻译[9]。

为了提高抓取网页的效率,本文直接获取搜索引擎返回的摘要,而不进入每个网站。通过考察网页中存在的术语翻译及其在搜索引擎返回的结果摘要中的形式,我们发现,大部分术语翻译在网页中与原术语直接相邻,或者相隔标点符,或相隔空白,或相隔一些标志词,比如"即","英文是","翻译是"等。在以上分析的基础上,本文提出了 3 种限制不同的查询项模式:

1)位置约束查询项模式:"C Ti" OR "Ti C",如""汽车设计师 designer" OR "designer 汽车设计师"",C 是待查的中文术语,Ti 是中文术语的第 i 个部分翻译。双引号的作用是固定位置,让搜索引擎严格遵循这种格式。

2)共现约束查询项模式:"C" "T1" OR "T2" OR…"Ti" … OR "Tn",例如""汽车设计师" "car" OR "automobile" OR "automotive""。

3)相关词约束查询项模式:"C" "翻译" OR "中译英" OR "汉英" OR "专业英语" OR "术语英文"。

Fang 等提出的基于语义预测构建查询项的方法仅考虑了以部分翻译扩展查询项,而没有考虑到搜索结果中术语与翻译的位置相关性。位置约束的查询项模式通过双引号进行位置相关约束,匹配术语与翻译相邻或仅相隔标点符号的情况,实验表明,这种查询模式的精度很高,返回结果的相关性也最高。由于部分术语无法通过语义预测获取正确的部分翻译,因此相关词约束模式通过组合特定词语获取相关网页。这种查询项模式是最宽松的限制,其可能出现更多不相关的网页,但对于专业术语的效果不错,并能解决低频词及术语不能分词或分词错误的问题。

3.2 基于多特征的翻译抽取

查询项返回的搜索结果中,首先需要对每条结果进行过滤,以过滤掉网页中的 HTML 标签。在 Google 中进行搜索时,Google 返回结果会对搜索词进行红色加亮处理,再找到每条返回结果中的加亮词,并进行标记。

在抽取候选翻译时,为了最大程度地抽取正确翻译,同时减少干扰项,本文针对搜索结果的特点,通过分析术语及翻译的位置、格式、出现形式等特征,总结出以下 3 种抽取方法:

1)基于模板的翻译抽取;

2)基于词典模式的翻译抽取;

3)基于位置的翻译抽取。

3.2.1 基于模板的翻译抽取

一些术语翻译的出现形式会有明显的特征或者标志词,如术语翻译可能出现在术语后的括号中,或者与术语以冒号或者"翻译为"、"英文是"等相隔。本文根据这些规律,研究确定多种模板,用于提取翻译。

一些模板翻译的例子:

1)C(,)(的)(翻译|英文)(名|名称|名字)(为|是|即)(:|,)E→(C,E)

例句:"电气工程的翻译是 electrical engineering。"

2)E(C)C′→(C,E)

例句:"电气工程(electrical engineering)是现代科技核心学科之一。"

3)E_1(,)C:E_2→(C,E_2)

例句:"English word 电气工程:electrical engineering。"

基于模板匹配的方法,由于其规律性较强,抽取时要求严格,抽取出的候选翻译准确性高,因此优先采用此抽取方法。但又由于网页是非结构化文本,仍会有大量搜索结果无法通过模板匹配方法抽取翻译,故转而采用其他抽取方法。

3.2.2 基于词典模式的翻译抽取

在网页中以词典型存在的英文翻译往往来自手工整理的学科或领域翻译列表或在线词典,具有较高的准确性。这类翻译在原网页以列表或表格形式存在,在搜索结果的摘要中会呈现英文词与中文词交错连续出现的情况(见图 1)。当以术语为中心的网页片段连续出现 3 次英文词组与中文词交替出现的情况时,可以认定其为词典型。词典型搜索结果由于

网页可靠性高、排列规整，因此采用基于词典模式的翻译抽取方法单独处理。但对于在线词典不包含的生僻词或一些未登录词，往往没有满足要求的词典型搜索结果。

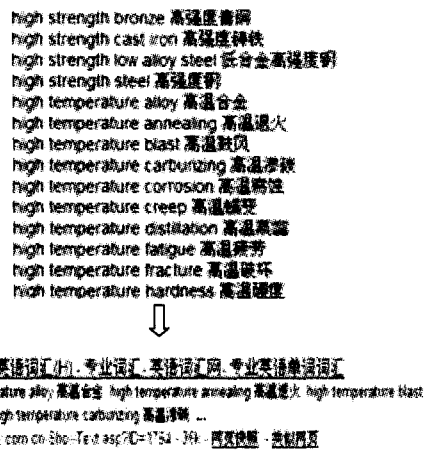

图 1　网页原文与该网页在搜索引擎中的片段

3.2.3　基于位置的翻译抽取

在多查询项获取的网页结果中，虽然其包含一定量的术语翻译，然而页面中有大量非术语翻译的英文词组和无关的干扰项。首先考察正确的术语翻译与原术语在网页中的分布。

Fang 查看了 5800×200 个网页，发现 5800 个中文术语与其英文翻译的距离分布符合高斯分布，70% 的英文翻译位于与中文术语直接相邻的位置[8]。

He 对来自 6 个不同领域的 25 个术语进行实验，结果表明，正确翻译为术语左、右前两个英文词组占所有正确翻译数的 98.7%[10]。

由以上分析可以得出，即绝大部分翻译候选项存在于术语左侧或右侧的前两个英文词组中。所以仅在术语两侧各两个英文词组中寻找翻译。

对于保留下的英文词组，继续考察其中是否有英文词组包含有红色加亮标记。如果英文词组的一个或几个单词位于红色加亮标记内，表明其包含有术语的部分翻译，则该英文词组为术语的翻译的可能性很大。

对于无法满足模板匹配和词典模式的搜索结果，采用基于位置的翻译抽取。相对前两种抽取方法，基于位置抽取准确率较低，抽取出的候选翻译包含无关的干扰项较多，但该方法适用于大多数网页形式，保证了候选翻译的召回率。

4　候选翻译验证和排序

由于网页内容的复杂性、不确定性以及抽取方法的局限性，使得网页中获取的候选翻译存在许多不相关的干扰项，影响获取结果的准确性。本文将错误的候选翻译类型定义为以下 3 类：

设标准翻译 T，候选翻译 C

1)C⊆T (C≠T)。如“咖啡厂”的一个候选翻译是“coffee”，标准翻译是“coffee plant”。

2)T⊆C (C≠T)。如“模范工人”的一个候选翻译是“The model worker is always tough at work”，标准翻译是“model worker”。

3)余下的情况。

在传统的翻译获取系统中，对获取的翻译只进行简单的验证处理，返回的结果仍包含有大量不正确的翻译，影响了结果的准确性、可读性和实用性。由于干扰翻译项的不同类型及中英文的语言特点，使得单一的验证方法无法取得良好的效果。本文利用手工整理的双语翻译对资源，提出了 3 种验证模型，对候选翻译包含多个单词的翻译项，需要满足端类比对齐验证或双语对齐度验证的条件；对候选翻译项为单个单词的情况，则采用构词法验证。

4.1　端类比对齐验证

端类比验证的方法通过已有的汉英术语翻译资源对候选翻译的开头和结尾词或词组进行类比验证。该方法可以有效地排除第一类和第二类错误候选翻译。

引入几个定义。

定义 1(头端对齐)　选翻译的开头单词或词组经过类比验证为正确。

定义 2(尾端对齐)　候选翻译的结尾单词或词组经过类比验证为正确。

接下来定义头端对齐度。当候选翻译头端对齐时，其头端对齐度大于 0；头端不对齐时，其头端对齐度小于 0。如果无法验证头端是否对齐，则头端对齐度等于 0。尾端对齐度的定义方法类似。最后定义端对齐度，候选翻译的端对齐度为其头端对齐度和尾端对齐度的较小值。

利用端对齐度，可以将候选翻译分为 3 类。

1)端对齐：头、尾两端都对齐，端对齐度大于 0。

2)部分端对齐：一端对齐，一端无法验证或两端均无法验证，端对齐度等于 0。

3)端不对齐：头、尾两端至少有一端不对齐，端对齐度小于 0。

本文使用图模型定义头、尾端对齐度的计算方法，对候选翻译进行端类比对齐验证。以验证中文术语“工友”的候选翻译“work mate”为例，计算尾端对齐度，如图 2 所示。

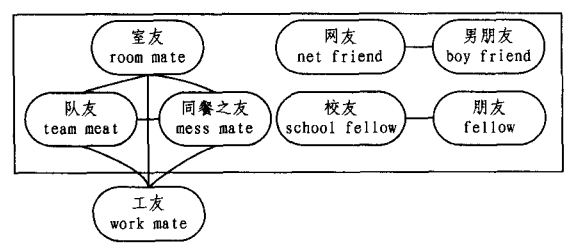

图 2　验证“工友”的候选翻译“work mate”尾端对齐度

图 2 中的节点表示中文短语与英文短语的翻译对。首先从已有中、英翻译对资源中选出与待验证术语“工友”有相同中文结尾词或词组的翻译对，组成图中的点。如果某两个点所表示的中、英文短语翻译对具有相同的英文尾端，则两点之间存在一条边。图中的一个连通子图中的点表示具有相同中文和英文尾端的翻译对。

由此定义尾端对齐度的计算方法为：如果表示候选翻译的节点可以加入到一个连通子图中，则候选翻译尾端对齐，尾端对齐度为候选翻译节点的边的个数，即已有翻译对中与待验证候选翻译有相同中文词结尾和相同英文词结尾的翻译对数量。如果候选翻译不能加入到一个连通子图中，则表示候

选翻译尾端不对齐，尾端对齐度为图中点数量的负数，即从已有翻译对中选出的与待验证中文术语有相同中文结尾词的翻译对数量的负数。

在计算出头、尾端对齐度后，可以求出候选翻译的端对齐度，保留端对齐的候选翻译，提高候选翻译的准确性。

4.2 双语对齐度验证

通过对中文术语及其候选翻译进行词汇级对齐，计算对齐度，对齐度越高的候选翻译的准确率越高。为了进行双语对齐，本文采用计算双语词语相似度的方法。中文词与英文词的意义越相似，其对齐的可能性就越大。

本文采用以下 3 种方法计算双语词语相似度。

1)基于词典的相似度计算

通过已有的双语翻译资源，构建双语词典，以索引表的形式存储英汉、汉英词语翻译。首先利用双语词典中的中英文词短语表，对中文术语 C_{term} 和英文翻译 E_{term} 采用逆向最大匹配算法进行分词，得到 $C_{term}=\{C_1\cdots C_i\cdots\}$，$E_{term}=\{E_1\cdots E_j\cdots\}$。若 C_i 与 E_j 在词典中互为翻译，则 C_i 与 E_j 相似。

2)基于统计的相似度计算

基于统计的方法已被前人证明是有效的。本文采用戴斯系数来计算双语词语的相似度，公式如下：

$$\mathrm{Dice}(S1,S2)=2|S1\cap S2|\ /\ (|S1|+|S2|)$$

给定中文词 C 和英文词 E，假设 C 和 E 在句子对齐的语料库中出现的句子集合分别是 Sc 和 Se，则可以用 Dice(Sc, Se)估计词语 C 和 E 的相似度。

3)基于语义的相似度计算

翻译中常会有利用同义词替代翻译词的情况。同义词词林是一部常用的同义词词典，整个词典形成一个树形结构，两个词 S1 和 S2 之间的语义距离 SenLength(S1,S2)表示两个词在树形结构上两个结点的最短路径，即两个词的语义近似程度。

定义词语 S1 和 S2 的相似度为：

$$\mathrm{SenSim}(S1,S2)=1/\mathrm{SenLength}(S1,S2)$$

定义中文词 C、英文词 E，C 在词典中的翻译为 Te，C 与 E 的语义相似度为：

$$\mathrm{SenSim}(C,E)=\mathrm{MAX}(\mathrm{SenSim}(C,D))\ (D\in T)$$

通过确定一个阈值，利用语义相似度就可以对候选翻译进行有效对齐，从而弥补了双语词典内容不足的问题。

采用基于词典的相似度计算最为简单也最为有效。但由于词典的规模有限，只要词典中没有该翻译对，就无法进行对齐，因此先采用基于词典的方法进行对齐，对无法对齐的部分，采用基于统计和语义的方法进行补充。

在定义了相似度计算的方法后，可以定义双语对齐度的计算公式。$D=\{C_i,E_j\}$ 表示中文词 C_i 与英文词 E_j 相似。

双语对齐中，中文短语对齐度

$$\mathrm{AliDegreeCn}=\sum(C_i,E_j)/C_{num}$$

双语对齐中，英文短语对齐度

$$\mathrm{AliDegreeEn}=\sum(E_j,C_i)/E_{num}$$

双语对齐度

$$\mathrm{AliDegree}=(\mathrm{AliDegreeCn}+\mathrm{AliDegreeEn})/2$$

利用双语对齐度将候选翻译分成 3 类。

1)对齐双语术语：每个 C_i 与某个 E_j 相似，每个 E_j 与某个 C_i 也相似，术语对齐度为 1。

2)部分对齐双语术语：某个 C_i 与每个 E_j 均不相似，术语对齐度在 0～1 之间。

3)非对齐双语术语：每个 C_i 与每个 E_j 均不相似，术语对齐度为 0。

设双语对齐度阈值为 α，仅保留 $\alpha>0.8$ 的术语，通过双语对齐度信息进行过滤，除掉可能的无关干扰项。

4.3 构词法验证

对于候选翻译为单个单词的情况，无法使用端类比对齐和双语对齐度的验证方法。针对这种候选翻译的特点，本文利用英文构词法的特点，采用构词法验证模型，对单个单词的翻译进行验证。这里列举几个构词法的例子。

构词法一：单词＋后缀

一些单词在添加某些后缀后会变成新单词，同时在含义上会有一定变化。如“ceram(陶瓷)”加上后缀“ist”后，变成“ceramist(陶瓷技师)”；“engrave(雕刻)”加上后缀“er”后，变成“engraver(雕刻师，雕刻家)”。

构词法二：前缀＋单词

一些单词在添加某些前缀后会变成新单词，同时在含义上会有一定变化。如“histamine(组胺)”加上前缀“anti”后，变成”antihistamine(抗组织胺)”

构词法形成的单词在含义上与原单词及前缀或后缀有着很大的相关性，通过已有的汉英翻译词典及构词法原则可以对候选翻译为单个单词的情况进行验证。

4.4 候选翻译的排序

对通过验证的候选翻译利用其置信度进行排序。置信度的计算依赖每个候选翻译通过每种翻译抽取模式从网页中获取到的频数 f 及每种抽取模式的权重 λ，权重 λ 取决于每种翻译抽取模式的准确性。

由此，本文定义候选翻译 t 的置信度 C 的计算方法为：

$$C(t)=\sum_{k=1}^{k}\lambda_k f_k(t)$$

式中，λ_k 是 $f_k(t)$ 对应的权值，k 是抽取模式的数量。计算出候选翻译的置信度之后，将其按照从大到小的顺序排序。

5 实验结果与分析

实验数据包括 1523 个随机抽取的各领域专业中文术语，每个术语包含 2～10 个汉字，部分中文术语包含英文单词或标点符号。通过手工或词典整理等方式获取了 120 余万条中英翻译对，用于获取中文术语的部分翻译，并在候选翻译验证时，将其作为双语翻译资源来验证候选翻译的准确性。

用 Google 搜索引擎获取网页摘要信息，对每个查询项，最多可以返回 1000 条网页摘要。

5.1 翻译正确率

在 1523 个测试中文术语的实验中，1369 个术语能够从网页中找到翻译项，召回率为 90%。实验的正确率统计如表 1 所列。

表 1 实验准确率

	TOP1	TOP3	TOP5
本文方法准确率	97.4%	98.3%	99.1%
Fang 的方法准确率	71.8%	94.5%	97.0%
Zhang 的方法准确率	77.13%	89.3%	95.1%

TOPn准确率是指召回翻译的术语中前n个翻译项中存在正确翻译的术语比率。实验数据表明，TOP1准确率达到了97.4%，翻译对的质量令人满意。选取Fang Gaolin在文献[7]与文献[8]的方法及Zhang在文献[5]的方法进行对比，可以看出，本文提出的3种翻译验证模式确保了翻译的准确性，准确率方面，特别是TOP1准确率较Fang Gaolin及Zhang的方法有较大幅度的提高，并增强了系统的实用性和可靠性。

传统的基于搜索引擎的翻译获取系统，需要构建查询项进行搜索引擎查询，耗时较多。本文基于相关性反馈动态控制网页下载量，定义一个有效翻译递减率来刻画一个查询项返回结果中每页获取到的候选翻译数相对前一页的递减率，以动态终止查询项。实验中每个术语的平均耗时如表2所列。

表2 术语平均耗时

	位置约束
本文方法平均耗时	8.3s
Fang的方法平均耗时	9.2s
Zhang的方法平均耗时	15s

5.2 翻译获取

考察3种查询项模式各自返回的候选翻译的比率和候选翻译的正确率，如表3所列。

表3 查询项模式返回候选翻译比率及正确率

	位置约束	共现约束	相关词约束
比率	21%	66%	13%
正确率	41%	24%	17%

表3的结果表明，位置约束的查询项模式虽然获取的翻译数量较少，但是获取到的翻译的准确性最高。

通过3种翻译抽取模式抽取出的候选翻译的比率及候选翻译的正确率如表4所列。

表4 翻译抽取模式抽取的候选翻译比率及正确率

	基于模板	基于词典	基于位置
比率	7%	5%	88%
正确率	72%	68%	20%

表4的结果表明，基于模板模式和词典模式抽取出的翻译准确性较高，基于位置的抽取方法则可以最大限度地保证翻译的召回率。

5.3 候选翻译验证

表5 不同验证模式下的实验结果

	TOP1正确率	召回率
All	97.4%	90.0%
端类比对齐+构词法验证	93.8%	94.5%
双语对齐度+构词法验证	90.4%	92.2%
无验证	71.4%	99.2%
Fang的验证方法	82.0%	95.3%

表5列出了实验中候选翻译经过不同验证模式后，实验结果的正确率和召回率。表中的4组实验表明，在候选翻译没有经过验证时，TOP1正确率仅有71.4%，采用单一验证方法效果不佳。针对候选翻译的特点，将3种验证方法进行结合，可以在保证召回率的情况下，有效提高实验结果的准确性，TOP1准确率上升到97.4%。作为对比，采用Fang在文献[7]的验证方法对实验数据进行验证。Fang的方法是基于互信息的方法，去除候选数据中的前缀或后缀冗余信息。实验表明，验证后的TOP1正确率提高到82.0%。但由于去除冗余信息需要通过比较多个候选数据，因此其不适用于网页搜索结果较稀疏的术语，且它易受到无关干扰项的影响。

结束语 本文提出了一种通过网页获取专业术语翻译的方法。通过术语的部分翻译构建查询项，极大地提高了搜索引擎返回结果的相关性。利用多特征的翻译抽取方法，在保证召回率的基础上抽取出准确的候选翻译。最后提出了3种验证模型，排除候选翻译中的干扰项，保证了结果的准确性。实验结果表明，抽取出的翻译对准确性高，错误的干扰项极少，具有较好的实用价值。

参考文献

[1] Huang F, Vogel S, Waibel A. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization[C]// Proceedings of ACL 2003 Workshop on Multilingual and Mixed Language Named Entity Recognition. 2003: 9-16

[2] Huang F, Vogel S. Improved Named Entity Translation and Bilingual Named Entity Extraction[C]// Proceedings of ICMI. 2002: 253-258

[3] Zhang Y, Vines P. Detection and Translation of OOV Terms Prior to Query Time[C]// Proceedings of SIGIR. 2004: 524-525

[4] Cao Gui-hong, Gao Jian-feng, Nie Jian-yun. A System to Mine Large-scale Bilingual Dictionaries from Monolingual Web Pages [C]// Proceedings of MT Summit XI. 2007

[5] Zhang Y, Vines P. Using the Web for Automated Translation Extraction in Cross-language Information Retrieval [C]// Proceedings of SIGIR. 2004: 162-169

[6] Huang F, Zhang Y, Vogel S. Mining Key Phrase Translations from Web Corpora [C]// Proceedings of HL T2EMNLP. 2005: 483-490

[7] Fang Gao-lin, Yu Hao, Nishino F. Chinese-English Term Translation Mining Based on Semantic Prediction [C]// Proceedings of the COL ING/ACL 2006 Main Conference Poster Sessions. 2006: 199-206

[8] Fang Gao-lin, Yu Hao. Web Translation Mining Based on Suffix Arrays[J]. Journal of Chinese Language and Computing, 2007, 17 (1): 1-141

[9] 吕学强，吴宏林，姚天顺. 无双语词典的英汉词对齐[J]. 计算机学报，2004，27(8)：1036-1045

[10] 何彦璋. 从Web中获取中文术语的英文翻译的方法研究与实现[D]. 北京：北京航空航天大学，2008

[11] 符建辉，曹存根，王石. 基于区分词的汉语隐喻短语识别[J]. 计算机科学，2010，37(10)：193-196

(上接第160页)

[29] Zhu J, Nie Z, Wen J R, et al. Simultaneous Record Detection and Attribute Labeling in Web Data Extraction[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(SIGKDD 2006). ACM Press, 2006: 494-503

[30] Zhu J, Nie Z, Zhang B, et al. Dynamic Hierarchical Markov Random Fields and Their Application to Web Data Extraction[C]// Proceedings of the 24th International Conference on Machine Learning(ICML 2007). ACM Press, 2007: 1175-1182