一种面向 WEB 的生物医学领域英汉术语翻译对抽取方法*

何莉　林鸿飞
大连理工大学计算机科学与工程系 大连 116024
E-mail: lihedlut@yahoo.cn

摘　要：双语词典是信息检索及相关应用的基础资源。但是领域专业双语词典不易获得且规模有限，因此本文提出一种面向 WEB 的生物医学领域自动获取双语术语翻译对的方法，以补充、完善双语词典。该方法主要包括候选中文对译词识别和对译词选择两个部分。前者使用了统计规则和长度-标准差模型，后者采用感知器算法及共现模型实现。通过对比实验结果显示，本文的方法是有效的，提高了术语翻译对抽取的准确率。

关键词：双语术语翻译对，长度-标准差模型，共现模型，感知器模型

A Web-Based Bilingual Terminologies Extraction Approach in the Biomedical Literature

Li He, Hongfei Lin
Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024
E-mail: lihedlut@yahoo.cn

Abstract: Bilingual thesauruses are fundamental resources for many NLP applications, such as information retrieval, question-answer, etc. Unfortunately, it is usually difficult to obtain such professional dictionaries, especially in the field of biomedical, and the coverage is often limited. In this paper, we propose an automatic method for extracting plenty of bilingual terminologies from biomedical literature so as to complement the bilingual dictionaries. Our approach consists of two major steps. The first step is candidate phrase identification, which is based on statistical rules and a Length Standard Deviation measurement. The second step is translation selection, in which we use a Perceptron algorithm aided by a co-occurrence model. Comparative experiments show that the proposed approach is effective and efficiency, and can consistently improve the precision of bilingual terminology extraction.

Keywords: Translation pairs in bilingual terminology, Length Standard Deviation measurement, co-occurrence model, Perceptron model.

1　前言

双语字典是跨语言信息检索、机器翻译和问答系统等自然语言处理应用的基础资源。随着互联网络的普及、信息爆炸性增长和专业领域新词不断涌现，人工编纂双语词典因其更新困难、容量有限等特点已经无法及时满足用户的需求。通过大规模语料自动获取术语翻译对，从而补充、完善双语词典，特别是特定领域双语词典成为一种研究趋势。在这一背景下，互联网以其网页更新迅速、内容涵盖广泛等特点成为抽取双语术语翻译对的首选语料。因此，近年来许多研究工作都围绕着从 Web 中自动获取双语术语翻译对，以补充和完善双语词典这一主题展开。

文[1-4]试图利用平行语料，以锚文本信息从大量的网页中获取通用领域术语翻译对。但是平行语料通常都难以获得，规模较小且锚文本信息有限，因此大大的限制了研究的深入展开。文[5-6]在上述基础上，基于非对齐的网页，使用互信息为指标衡量英文关键词和其相应的中文对译词(即该英文关键词的中文翻译词)的关联度，并提出对于任意给定的两个英文词 e_i 和 e_j(i 和 j 不等)，

*本文承国家“八六三”计划项目（2006AA01Z151）和国家自然科学基金资助项目(60373095，60673039)的资助。

如果它们在语料中的共现度高，那么它们的对译词 c_i 和 c_j 之间的共现度也必然很高这一思想。

但是，面向生物医学领域的搜索，检索结果中存在大量和英文关键词共现度低并且出现频度低的词汇，从而导致以上方法很难直接应用。因此，本文中仅利用互信息抽取很难得到准确的结果，必须在上述方法的基础上考虑领域特性从而适应生物医学领域术语抽取的特点。

经统计发现，根据生物医学英文关键词检索得到的单语中文网页片段，包含大量的英汉双语术语翻译对信息，如："Aflatoxins"和其中文对译词"黄曲霉素"经常共现。因此，该网页片段可以作为抽取英汉双语术语翻译对的一种有效资源，从中获取翻译对信息。

本文余下章节的安排如下：第二部分描述了英汉双语术语翻译对抽取系统的结构。第三部分描述了抽取双语术语翻译对的过程，包括候选中文对译词的识别和对译词选择。第四部分介绍了详细的实验过程，包括第三部分中引入的特征和模型对实验结果的影响。第五部分对工作进行总结并提出了进一步研究的方向。

2 英汉双语术语翻译对抽取系统结构

英汉双语术语翻译对抽取系统主要是完成在搜索引擎中，通过给定英文关键词首先检索得到包含该关键词的双语网页片段，然后识别出这些网页中所有与之对应的候选中文对译词，最后从中选择最可能成为该英文关键词译项的部分，抽取出英汉双语术语翻译对。抽取系统的构架如下图 1 所示，其中主要包含两个步骤：候选对译词识别和对译词选择。

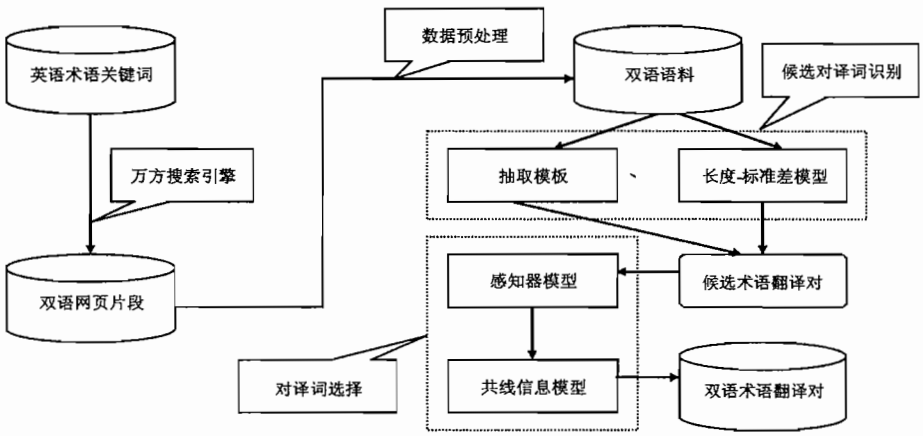

图 1 系统构架

候选对译词识别过程主要是从网页片段中，得到与英文关键词相对应的候选中文对译词。由于生物医学领域中专业术语的长度和出现频率差异均较大，如果事先固定对译词边界将导致识别不完整。因此本文不采用文[1]中方法而是基于模板提取英文关键词周围出现中文词串，进行分词后将所有中文词串的组合都作为候选中文对译词。例如：从英文关键词"Infectious Bronchitis Virus"周围提取出"传染性支气管炎病毒"这一短语，通过分词并逐个组合后，将"病毒"、"炎病毒"、"支气管炎病毒"和"传染性支气管炎病毒"都作为其候选对译词，最后对于模板未能提取出的部分采用长度-标准差模型进行处理。

对译词选择过程主要实现从所有候选对译词中选择出最可能成为英文关键词译项的部分。这一过程中，由于生物医学领域的特殊性，本文提出了基于单语医学词典，采用语言模型计算候选对译词的成词概率，并将其作为特征加入感知器模型进行训练，最后结合共线模型来提高选择准

确性这一方法。对译词识别和选择过程将在第 3 节详细说明。

3 英汉双语术语翻译对抽取过程

根据第 2 部分所述，本文将双语术语翻译对的抽取过程分为候选对译词识别和对译词选择两个主要部分。候选中文对译词的识别过程主要采用模板过滤和长度-标准差模型两种方法。而文[7]提出对译词的选择可以被看成一个 ranking 任务，即从所有候选的中文对译词中，挑选得分最高的部分作为正确的选项。基于该思想，本文训练一个二值分类器，以此判断候选对译词成为英文关键词译项的可能性，并在实验中设置一个阈值，保留超过该阈值的部分作为正确的译项，最后对于分类器判断错误的实例采用共现模型进行二次判别。

3.1 候选中文对译词识别

候选中文对译词识别解决的是根据英文关键词，从生物医学网页片段中识别出所有与之对应的候选中文对译词的问题。文[2]将所有中文关键词周围的英文词串都提取出来，作为该关键词的候选对译词。这种方法虽然充分利用了局部信息，但同时会加大干扰因素。因此，本文对文[2]方法进行改进，通过统计生物医学领域的网页中双语术语翻译对的分布规律，定义以下模板用以识别候选中文对译词。

(1) $c_1c_2\ldots c_i$(e)；

(2) $c_1c_2\ldots c_i$ 是/即(is/are)e；

(3) $c_1c_2\ldots c_i$e。

其中，e 代表的是英文关键词，$c_1c_2\ldots c_i$ 代表的是候选中文对译词。

同时，为了避免使用模板而导致数据缺失，文中对于使用模板后没有提取出候选对译词的英文关键词，采用长度-标准差模型进行二次提取处理，从而保证数据的完整性。

文[8]认为一个句子中，如果一个中文词串内各个词之间的结合度很高，那么该中文词串的标准差值就会比较低。因此，标准差值可以作为衡量一个中文词串在句子级别结合紧密程度(即它成为一个词的概率)的标准。但是在某些情况下，标准差并不能完全反映一个词的成词概率。如："perennial allergic rhinitis" 对应的候选对译词 "反应性鼻炎" 的标准差值和 "变态反应性鼻炎" 几乎相等，因而无法正确识别。但是该情况下，较短的中文词串往往是长的中文词串的一部分。因此，长度较长的词串应该得到更高的成词概率。所以，本文在文[8]的基础上加入字符串长度信息，从而得到长度-标准差模型，如公式(1)所示。

$$\Gamma = \frac{len}{1+\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}} \tag{1}$$

其中，x_i 代表中文串中每个词出现的次数，$\bar{x}$ 表示该中文串中各个词的平均出现次数，*len* 代表中文词串的长度。同时为了避免出现中文词标准差值过小而导致 Γ 值无法计算的情况，文中加入数据平滑处理。这样，引入长度-标准差模型就可以正确识别出英文关键词对应的候选中文对译词，供对译词选择部分使用。

3.2 对译词选择

对译词选择过程是从已识别出的候选中文对译词中，选择最可能成为英文关键词译项的部分。整个选择过程主要分为两个步骤：训练基于感知器算法的二值分类器，以区分正确和错误的候选中文对译词实例；采用共现模型对感知器模型判断错误的实例做二次判断。

3.1 节候选中文对译词识别过程认为，基于模板提取的英文关键词周围的中文词串，都是该英文关键词的候选对译词。但是这样处理容易导致在训练分类器的过程中，错误的训练实例要远

多于正确的实例这一数据失衡现象，从而使得模型训练不准确。

因此，本文使用 Averaged Perceptron model 以弥补上述不足，同时采用模型参数平均化(Averaging parameters)[9]的特征调整算法，避免在感知器的学习过程中，由于学习速率过大所引起的数据震荡现象，从而保证实验结果的准确性。这样，通过对感知器模型的训练可以赋予正确的例子更高的得分，同时降低错误的例子的得分，从而正确选择出中文对译词。

上述过程中，尽管感知器模型可以很好的防止数据不均衡带来的错误，但是对于其没有学习过的样本特征，感知器无法做出正确的判断。而共现模型提供的局部信息可以弥补感知器模型这一不足。因此对于感知器模型判别错误的中文对译词，文中采用以下共现模型进行二次判别，选择 Λ 值最大的候选中文词串作为英文关键词的正确中文对译词，进一步保证了抽取的准确性。

$$\Lambda = \frac{\lambda * len}{1+\sigma} + \gamma * frequence(s) + (1-\lambda-\gamma) * len \tag{2}$$

其中，$frequence(s)$ 代表中文词串出现的次数，len 代表中文词串的长度。λ, γ 是经验数据。

4 实验

4.1 实验数据

本文实验所用数据是从专业英文医学词典中，随即抽取 2,643 个生物医学领域英文关键词，在万方搜索引擎及其科技文章数据库[10]中，检索得到所有包含这些英文关键词的 13，2150 个英汉双语网页片段，利用这些双语网页进行英汉双语术语翻译对抽取实验。

4.2 实验过程

实验中首先将所有双语网页片段使用 Htmlparser[11]工具包进行信息过滤，其次利用模板对双语网页片段进行候选中文对译词识别。由表 1 可以看出，3.1 节所述三个模板已经涵盖了网页中双语术语的大多数出现形式。因此，本文采用以下三个模板进行初步处理，加快识别速度。

表 1 三个模板的出现概率和正确率

模板	出现概率	包含正确候选对译词比率
(1)	62.5%	92%
(2)	5.1%	2.55%
(3)	25.7%	44.45%

对于模板未能识别的候选中文对译词，我们采用长度-标准差模型补充识别，然后将全部识别出的候选对译词交由感知器模型进行判别。这一过程中，为了选择最贴近实验数据的特征训练感知器模型，本文对处理后的全部网页进行分析和统计，并参考文[12]初始选定了以下 3 个特征训练感知器模型。

(1) 候选中文对译词的周围是否有指定边界的词，如：“的”，“和”，“与”，如果有则提高该候选中文对译词的得分；
(2) 候选中文对译词和英文关键词一起出现的次数。
(3) 英文关键词与候选中文对译词的长度比值；

但是上述 3 个特征在感知器模型的训练过程中必定存在一定的偏差，因此需要分析初始特征训练下的感知器模型所得结果，从而进一步调整训练特征，并且对于感知器模型判断错误的部分

采用共现模型加以纠正，这一过程将在第 4 节详细说明。

4.3 实验结果与分析

4.3.1 感知器特征分析和调整

通过上述 3 个特征训练感知器模型，术语翻译对抽取准确率仅为 48.91%。通过分析我们可以明显观察到一些和英文关键词共现度高，并且包含正确中文对译词，但并不能构成名词或名词短语的词串频繁出现，严重干扰了抽取过程。造成上述现象的原因就是没有确定成词边界，因此在判别过程中一旦出现该类词串时就容易产生误判。例如："postmenopausal osteoporosis" 对应的候选对译词中，"建立绝经后骨质疏松症" 显然不是一个有意义的名词或名词短语。在 "建立绝经后骨质疏松症" 中，"绝经后骨质疏松症" 的成词概率就直观上来说，也要比 "建立绝经后骨质疏松症" 大的多。

为了解决上述问题，本文加入成词概率这一特征，以过滤无意义的词串。候选中文对译词的成词概率是通过引入中文医学词典，利用语言模型计算得到的。成词概率这一特征加入训练感知器模型后，很好消除上述抽取错误的现象，使得术语抽取准确率达到了 86.86%，比第一次实验结果有了明显的提高，进一步证明这一特征对于抽取过程是有效的。

为了验证 4.2 节初始所选特征的有效性，在引入上述 "成词概率" 特征训练感知器模型的基础上，本文依次加入初始 4 个特征，并分析引入这些特征前后抽取结果的变化。实验结果表明，特征 1、2 和 3 均能提高抽取准确度，但是特征 3 的加入容易导致长度较长但是很少出现的中文对译词不能被正确识别出来。因此，本文引入新的特征 "候选中文对译词长度"，以此保证长度长的中文词串有更高的优先级。通过加入这个特征，双语术语翻译对的抽取准确率达到了 90.51%，比之前的实验结果又有了进一步的提高。

综上所述，本文最终选择了以下 5 个特征训练感知器模型。

(1) 候选中文对译词的成词概率；
(2) 候选中文对译词的周围是否有指定边界的词，如："的"，"和"，"与"，如果有则提高该候选中文对译词的得分；
(3) 候选中文对译词和英文关键词一起出现的次数；
(4) 英文关键词与候选中文对译词的长度比值；
(5) 候选中文对译词的长度。

4.3.2 实验结果

通过上述 5 个特征训练感知器模型，翻译对的抽取准确率达到了 90.51%。但是，由于文中仅使用语言模型计算成词概率，以判断一个中文词串的组合是否合理。因此，对于语言模型没有学习过的陌生词串，这一判断过程显然无法进行。例如：英文关键词 "spironolactone" 相应的候选对译词中，语言模型没有学习过 "螺内酯" 一词，因此无法识别出抽取得到的候选中文对译词串 "受体拮抗剂螺内酯" 不是一个合理的词串组合，而造成错误的将其判别成 "spironolactone" 对应的正确中文对译词。基于这种情况本文进一步引入共现模型加以判断处理。

从类似 "螺内酯辅助治疗重度低钾型周期性瘫痪" 的句子中，共线模型可以判断出得分最高的词串 "螺内酯"，从而避免上述的错误发生。通过引入共现模型纠正感知器判别错误的实例，抽取准确率进一步提高到 93.43%

综上所述，整个实验过程中引入各个特征和模型对于实验结果的影响如表 2 所示。

表 2 实验结果表明，使用包括成词概率的五个特征训练感知器模型，并加入共现模型补充判断可以获得比较好的实验结果。我们从获取的 132,150 个网页片段中挖掘出 1672 个正确的英

汉双语术语翻译对，抽取结果正确率达到93.43%。抽取出的这些术语翻译对有效地挖掘了Web资源，可以很好的补充双语词典，达到了实验目的。

表2 各个特征和模型对实验结果的影响

实验所用特征和模型	实验结果准确率
初始选择的3个特征	48.91%
加入成词概率特征	86.86%
加入长度特征	90.51%
加入共现模型	93.43%

5 结语

本文主要研究一种面向WEB的生物医学领域英汉术语翻译对抽取方法。其中包括两个主要部分：候选对译词识别和对译词选择。文中分别对这两个部分及所涉及的模型和方法做出了详尽的说明，并以实验结果阐述了系统的有效性。

下一步，我们将在此基础上完成英文词的同义词扩展，解决以大写字母缩写形式出现的英文关键词，如何根据领域特性正确选择其中文对译词的问题，进一步提高术语翻译对抽取准确率。

参 考 文 献

[1] Lu W, Lee H. Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach. ACM Transactions on Information Systems, 2004, 22: 242-269.

[2] Zhang Y, Vines P. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In the Proceedings of SIGIR, 2004: 162-169.

[3] Fei H, Ying Z, Stephan V. Mining key phrase translation from web corpora. In the Proceedings of HLT-EMNLP2005, 2005: 483-490.

[4] Wai L, ShingKit C, Ruizhang H. Named Entity Translation Matching and Learning: With Application for Mining Unseen Translations. ACM Transactions on Information Systems, 2007, 25(1): 38-69.

[5] 蒋龙,周明,简立峰.利用音译和网络挖掘翻译命名实体.中文信息学报, 2007, 21(1): 23-29.

[6] HsinHsi C, WenCheng L, Changhua Y, Weihao L. Translation-Transliterating Named Entities for Multilingual Information Access. Journal of the American Society for Information Science and Technology, 2006, 57(5): 645-659.

[7] Joachims T. Optimizing Search Engines Using Click through Data. In the Proceedings of the 8th ACM Conference on Knowledge Discovery and Data Mining, 2002.

[8] Chengye L, Yue X, shlomo G. Web-Based Query Translation for English-Chinese CLIR.中文计算语言学期刊, 2008, 13(1): 61-90.

[9] 于浩,步丰林,高剑锋.感知器在语言模型训练中的应用.计算机研究与发展, 2006, 43(2): 260-267.

[10] http://www.ilib.cn/

[11] http://htmlparser.sourceforge.net/

[12] Guihong C, Jiangfeng G, JianYun N. A System to Mine Large-Scale Bilingual Dictionaries from Monolingual Web. In the Proceedings of MT Summit XI2007, 2007: 57-64.