基于搜索引擎的专有名称译文挖掘研究

葛运东　孙常龙　房璐　姚建民
苏州大学江苏省计算机信息处理重点实验室苏州 215006
E-mail: geyundong@gmail.com,jyao@suda.edu.cn

摘要： 查询翻译是影响跨语言信息检索的关键因素之一，而查询中有很大比重是专有名称，因此专有名称译文的挖掘对改进查询系统性能具有重要意义。本文首先利用主题词译文查询扩展方法从搜索引擎获取有效双语摘要资源；其次，利用频度变化信息和邻接信息，从含有噪声、规模相对较小的摘要资源中抽取复合词、短语等多词候选翻译单元；最后综合音译特征、统计特征和模板特征进行专有名称译文选择。

关键词： 跨语言信息检索，专有名称，查询翻译，译文挖掘，查询扩展，多词单元抽取

Search Engine Based Proper Names Translation Mining

Ge Yun-Dong, Sun Chang-Long, Fang Lu, Yao Jian-Min
Provincial Key Laboratory of Computer Information Processing Technology
Soochow University, Suzhou, China, 215006
E-mail: geyundong@gmail.com,jyao@suda.edu.cn

Abstract: Query translation is one of the major challenges to cross-language information retrieval (CLIR).The vast of queries are proper names, so proper names translation mining is essential to improving the performance of the CLIR system. This paper introduced translations of the topic words based query expansion method for gathering effective bilingual snippets. Frequency change information and adjacent information was utilized to extract MLUs (multi-lexical unit, words, compound words, phrases etc.) from noisy, small scale snippets. We combined transliteration model, statistic model and surface pattern model to evaluate the appropriate translation.

Key Words: Cross-Language Information Retrieval, Proper Names, Query Translation, Translation Mining, Query Expansion, Term Extraction

1 引言

跨语言信息检索（CLIR）通过不同语言的查询，使人们能够检索多种语言的文档。虽然近来 CLIR 获得了快速的发展，但查询翻译仍是制约其性能的关键因素之一。文献[1]中指出查询中大约 50%为专有名称（Proper Names），包括人名、地名等，如果不能对这些专有名称进行正确的翻译，会使不相关文档排序较高。因此专有名称译文的挖掘对 CLIR 具有重要意义。

早期的 CLIR 多采用词典进行查询翻译。源查询在词典中对应译文通常有多个，所以要面临翻译歧义问题；即使使用质量很高的词典，还是有一些查询不能从词典中查找到对应译文，即所谓的 OOV（out of vocabulary）问题。文献[2]分析了此方法存在的问题，并给出相应的解决方法。

随着网络资源的日渐丰富，许多学者尝试从语料库中挖掘专有名称译文，文献[1][3]利用平行语料库进行查询翻译；文献[4]利用从网络挖掘的可比语料库进行查询翻译；文献[5]利用统计

词特征、词关联矩阵从非平行语料库中抽取翻译词对。基于语料库的翻译方法的关键问题是如何自动构建大规模的、多领域的语料库。

Nagata 等[6]首次尝试利用搜索引擎下载前 100 个网页进行日语查询的翻译；Cheng 等[7]利用上下文向量和卡方进行译文选择，只需利用搜索引擎返回的前 100 个摘要信息（snippets）作为双语语料库；Chengye Lu 等[8]也利用前 100 个摘要和共现信息进行源查询译文的抽取。以上方法由于没有进行跨语言扩展造成返回的前 100 个摘要通常是单语的，没有包含有效的目标语言译文，不能用来进行抽取目标语言译文。Fei Huang 等[9]利用搜索引擎挖掘关键短语的译文；Gaolin Fang 等[10]首先将源查询拆分，跨语言扩展后收集双语网页进行抽取译文；Sun Jun 等[11]采用正向-逆向最大匹配拆分方法进行跨语言扩展后抽取最后的译文。由于都要对源查询进行拆分，拆分后每个单元的组合意义和源查询的意义往往不同，这将会引入额外的噪声甚至错误。

针对以上问题，本文首先采用基于共现信息的主题词扩展方法进行跨语言扩展，从搜索引擎收集有效的双语摘要信息；其次采用改进的基于频度变化信息和邻接信息的候选翻译单元抽取方法抽取高质量的候选单元；最后结合音译特征、模板匹配、频度和距离信息进行专有名称译文的选择。实验结果表明，我们的专有名称译文挖掘方法取得了很好的效果。

本文剩余部分安排如下，第 2 节将详细介绍本文提出的专有名称译文挖掘方法；接着在第 3 节给出实验结果和结果分析；最后进行简要的总结和介绍未来工作。

2 专有名称译文挖掘

2.1 双语摘要资源自动获取

采用基于共现信息的主题词译文对源查询进行跨语言扩展。将待翻译的专有名称提交搜索引擎获取源语言摘要信息。对源语言摘要信息进行预处理；从中获取名词词汇列表，采用 TF*IDF 进行加权排序；选取前 5 个作为源专有名称的主题词。接着在双语词典中查找主题词对应的目标语言译文；把源查询与获得的主题词的译文一起提交搜索引擎获取双语摘要资源。比如，我们将“Macedonia”（马其顿）提交搜索引擎后，获得的主题词分别为：“republic”，“country”，“politics”，“culture”，“population”，通过查找英汉双语词典，主题词对应的译文分别为：“共和国”，“国家”，“政治”，“文化”，“人口”；双语扩展后得到的新查询为：“Macedonia”＋“共和国”，“Macedonia”＋“国家”，“Macedonia”＋“政治”，“Macedonia”＋“文化”，“Macedonia”＋“人口”。

2.2 候选翻译单元抽取

由于所获取的双语摘要资源规模较小，仅包含几个句子或部分句子片段，本文采用频度变化信息[8]与邻接信息抽取候选翻译单元。此方法对于频度较低的候选单元，其抽取效果也较好。方法基于以下两个观察。一是同一个合法的候选翻译单元中每个字符的频率是相似的；二是若对一个合法的候选翻译单元用额外的一个字符进行扩展，扩展后的单元频度会显著降低。采用如下公式判断一个字符串是否为候选单元。

$$R(S)=\frac{f(S)}{1+\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad (1)$$

其中，S 是一中文字符串，f（S）是字符串 S 的频度，x_i 是 S 中每一个字符的频度，$\overline{x}$ 是 S 中所有字符的平均频度。

此方法抽取的候选翻译单元集合中仍包含一些合法翻译单元的子串，有时一个合法的翻译单元加一个额外的字符后频度并没有降低，因此，正确的候选翻译单元可能不能被抽取出来。合法的候选翻译单元一般具有多样的邻接字符而其子串却具有相对稳定的较少的邻接字符，因此引入邻接信息来改进上面的问题，公式修改如下：

$$R(S)=\frac{LN(S)\times f(S)\times RN(S)}{1+\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}} \qquad (2)$$

其中 LN（S）是与 S 左邻接的不同的字符总数，RN（S）是右邻接的不同字符的总数。我们没有直接对源字符串 S 去除停用词，而是将其中的停用词采用符号“+”替换，根据“+”与其他标点符号将 S 分成一系列子句，然后分别从子句中抽取候选单元，这避免了停用词前面的字符和停用词后的字符被抽取为候选单元。获得的候选集合中包含一些可以在词典中查找到的词汇，因为源查询是 OOV，所以其目标译文通常也是 OOV，因此能从词典中查找到的候选单元被删除。

2.3 译文选择

本文采用音译模型、统计模型和模板匹配模型来综合选取译文。

2.3.1 音译模型

Li Haizhou 等[12]利用音译技术进行译文的抽取，我们的音译模型与他们的不同之处在于：首先我们已经具有了对应的候选译文集合，不需要由源查询直接产生对应的目标语言音译，只需要计算源查询与每个候选译文的音译匹配相似度，选取相似度最高的候选作为最终的译文；其次，为了避免从英语到音素，从音素到拼音，从拼音到汉字转换的双重错误，我们采用[13]类似的思想，直接将源专有名称转换成音节，然后计算每个音节与每个汉字字符的相似度。根据启发式规则拆分成音节后，采用如下公式计算专有名称和某一候选单元的分值：

$$Trl(s,t)=P(s,t)/D(s,t) \qquad (3)$$

其中 s 为源专有名称，t 为任一候选单元，分子是 s，t 共现的概率，分母为 s 和 t 中不同音节的数目。P(s,t)的定义如下：

$$P(s,t)\approx\prod_{i=1}^{\min(m,n)}(1-\gamma)prob(e_i,c_i) \qquad (4)$$

其中，γ 为平滑系数，$prob(e_i,c_i)$ 为英语音节 e_i 与汉字字符 c_i 匹配的概率，这个概率从包含 37665 个英汉专有名称对的音译训练语料中通过动态规划算法训练获得。$D(s,t)$ 的定义为：

$$D(s,t)=\varepsilon+|m-n| \qquad (5)$$

其中，ε 为衰减参数，m 为英语音节总数，n 为一个候选单元中汉字字符的总数。为了避免音节拆分的错误，对源专有名称进行正向、逆向两种音节拆分，并分别计算两种拆分方法下与候选单元的得分，最后音译模型的得分修改为：

$$Trl(s,t)=(Trl_F(s,t)+Trl_B(s,t))/2 \qquad (6)$$

其中，$Trl_F(s,t)$ 为正向拆分下的音译得分，$Trl_B(s,t)$ 为逆向拆分下的音译得分。专有名称为多词的情况下，由于翻译习惯的不同，可能存在翻译顺序的调整。比如“Andersen,Hans,Christian”被翻译为“汉斯·克里斯蒂安·安徒生”。因此我们对词序进行调整，计算在每种可能的顺序下与候选单元的音译得分，把得分最高值作为最后的音译得分。

2.3.2 统计模型

在给定双语摘要集合中，真正译文通常与源专有名称一同出现，两者具有相似的频度；其次若一个候选单元与源专有名称距离越近，其为真正译文的可能性也就越大。源专有名称与某一候选单元的相似度采用如下公式计算：

$$F\_Q(s,t) = \sum_J \sum_k \frac{1}{d_k(s,t)} \Big/ \max_{fre-dis} \quad (7)$$

其中，s 为源专有名称，t 为其中某一个候选单元，J 为所有摘要的总数，K 为在一个摘要中 s，t 共现的次数，因为 s，t 可能在一个摘要里共现多次；$d_k(s,t)$ 为 s，t 在一个摘要中的第 k 次共现的距离。有两种方法计算 s，t 之间的距离，一种是利用字节距离，另一种是采用字符个数，我们采用 s，t 之间的字符个数作为其距离，若 s，t 之间存在一个字符，则距离为一，以此类推；$\max_{fre-dis}$ 为所有的候选单元中距离的倒数的最大值。

2.3.3 模板匹配模型

亚洲语言（如汉语，日语，韩语等）用户习惯在第一次使用一个术语的时候将其对应的译文标注在括号中。Jian-Cheng Wu 等[14]利用此启发信息从网络挖掘译文。专有名称和其对应译文之间的标点符号信息对译文抽取来说是相当重要的。首先将一些英语-汉语词对提交搜索引擎自动学习表层模板。如果一个候选翻译单元和源专有名称匹配了多数模板，那么其作为正确译文的概率则较大。模板匹配的贡献值采用如下公式计算：

$$SP(s,t) = N_{matching} / \max_{num} \quad (8)$$

其中，s 是源专有名称，t 为一候选单元，分子为 s，t 匹配的模板的总次数，分母为所有候选中匹配次数的最大值。

3 实验结果及分析

我们从维基百科的人名表中[15]的外文人名中获得了 300 个英语-汉语人名对；从世界地理索引中共获得 224 个英语-汉语国家（地区）词对，使用其中的英语部分作为专有名称查询，对应的汉语部分作为正确译文的参考。

我们从 524 个查询中随机抽取了 50 个查询，对这 50 个查询抽取候选单元，采用正确率、包含率作为抽取的候选单元的评价指标。如果抽取出的一个候选单元具有合法的词汇边界，即是一个具有实际意义的单元，则被认为是正确的。正确率、包含率的定义如下：

$$\text{正确率}=\frac{\text{正确的候选单元个数}}{\text{候选单元总数}}*100\% \quad (9)$$

$$\text{包含率}=\frac{\text{候选单元中含有正确译文的查询数}}{\text{查询总数}}*100\% \quad (10)$$

正确率反映了对合法多词单元的识别能力，包含率反映了识别源查询对应译文的能力。对译文挖掘系统来说，包含率具有更重要的意义，如果正确译文没有被正确的识别出来，即候选单元集合中没有其正确译文，译文选择模型是不可能挖掘到正确译文的。实验结果如表 1 所示：

表 1 候选单元抽取结果

	正确率	包含率
候选单元抽取	88.44%	88%

从上表看出，本文的候选单元抽取方法取得了很好的效果。高质量的候选单元集合为后面的译文选择排序奠定了坚实的后盾。

对于译文抽取，我们采用 TOP N 包含率（Inclusion Rate）[7]作为评价尺度。单独采用音译模型，以及综合了音译模型、统计模型和模板匹配模型在人名，国家地区名两个测试集合上进行了实验，实验结果如表 2 所示。

表 2 不同方法的译文抽取结果

		TOP 1	TOP 3	TOP 5	TOP10
外国人名	音译	54.3%	71%	76.7%	81%
	综合	55%	79.3%	84.3%	88%
国家地区名	音译	54%	66.5%	73.2%	81.3%
	综合	84.8%	97.8%	98.2%	98.7%

从上表可以看出，单独使用音译模型就取得了不错的效果，在两个测试集上 TOP 1 的包含率分别为 54.3%和 54%，TOP 10 的包含率分别为 81%，81.3%。在国家地区名上，音译模型的 TOP1、TOP3、TOP5 的包含率均低于在外国人名集合上的包含率，外国人名中音译词的比重高于国家地区名的比重，国家地区名中包含更多的非音译词，如“Iceland（冰岛）”，“Isle of Man（马恩岛）”等等。综合代表音译模型、统计模型、模板匹配模型的简单的线性综合，在两个测试集上性能均得到了改善 TOP 1 分别从 54.3%增加到 55%，54%增加到 84.8%，增幅分别为：0.7%，30.8%。综合方法在 TOP 10 取得最高性能 88%和 98.7%。

综合方法在两个测试集上对音译模型的改进程度不同，对外国人名改进程度没有在国家地区名上的改进程度明显。外国人名中某些名字在网上的资源非常少，比如“Vogt Alfred Eltonvan（范·沃格特）”等，搜索引擎返回的摘要中基本没有包含其对应中文信息，其次，人名翻译相对随意，不正规，所以不能抽取出对应的中文译文。国家地区名虽然音译的比重较低，但是其在网上的资源相对比较丰富，同时其翻译相对比较稳定和正规，因此综合方法取得了显著的性能提升。

在抽取的译文集合中发现，本文的方法能够抽取出查询的多个正确译文，比如“Addison Joseph”，抽取的译文为：“约瑟夫艾迪生”和“约瑟夫阿狄生”，这两个译文都是正确的，只是不同人在音译时选字不同造成；“Cape Verde”我们抽取的译文为“佛得角”，“佛得角共和国”，一个简称，一个全称，都是其正确译文。抽取出查询的所有形式的正确译文对 CLIR 是有积极意义，这无疑是简单自然的查询扩展，对改进 CLIR 系统的性能具有重要作用。

4 总结

本文利用共现信息，从搜索引擎获取与源专有名称相关的同一主题的主题词，对获取的主题词进行翻译后，利用主题词译文对源专有名称进行跨语言扩展，利用扩展后的查询再次从搜索引

擎获取高质量的双语摘要资源；其次，针对所获取的摘要资源含有更多噪声、规模较小等特点，利用基于频度变化信息和邻接信息抽取复合词、短语等候选翻译单元，所抽取候选单元集合质量高；综合利用音译特征、统计特征和模板特征进行最终的专有名称译文选择。实验结果表明，本文提出的专有名称译文挖掘取得了很好的效果。

在下一步的工作中，我们将探索利用其他的特征，如语义特征等以进一步提高所挖掘的译文的质量；另一方面我们将采用更大规模、更标准的测试集进行实验；最后将在 CLIR 系统中采用本文的译文挖掘方法，测试其对 CLIR 的性能改进程度。

参考文献

[1] Davis, M. W. and W. C. Ogden, Free Resources and Advanced Alignment for Cross-Language Text Retrieval, In Proc. of the Sixth Text Retrieval Conference (TREC6), Gaithersburg,Maryland, 1998,385-394.

[2] Pirkola, A., Hedlund, T., Keskustalo, H. & Järvelin, K. Dictionary-based cross-language information retrieval: problems, methods, and research findings. Information Retrieval, 2001,4(3/4): 209-230.

[3] J.Y. Nie, M. Simard, P. Isabelle, R. Durand. Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web. SIGIR-1999, 74-81.

[4] Tuomas Talvensaari , Ari Pirkola, Kalervo Ja¨rvelin, Martti Juhola,Jorma Laurikkala. Inf Retrieval , 2008,11:427–445.

[5] Fung, P. and Yee, L.Y. An IR Approach for Translating New Words from Nonparallel , Comparable Texts. In Proc. Of COLING-ACL, 1998,414-420.

[6] M. Nagata, T. Saito, and K. Suzuki. Using the web as a bilingual dictionary, Proc. ACL 2001Workshop Data-Driven Methods in Machine Translation, 2001, 95-102.

[7] P.–J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, and L.-F. Chien. Translating unknown queries with web corpora for cross-language information retrieval. In the Proceedings of 27th ACM SIGIR, 2004,146-153.

[8] Chengye Lu, Yue Xu, and Shlomo Geva. Web-Based Query Translation fro English-Chinese CLIR. Computational Linguistics and Chinese Language Processing, 2008, 13(1), 61-90.

[9] Fei Huang, Ying Zhang and Stephan Vogel. Mining Key Phrase Translation from Web Corpora. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP),2005,483-490.

[10] Gaolin Fang, Hao Yu, and Fumihito Nishino. Chinese-English Term Translation Mining Based on Semantic Prediction. In Proceedings of the COLING/ACL 2006 Main Conference Poster Session,2006,199-206.

[11] Sun Jun, Yao Jian-Min, Zhang Jing, Zhu Qiao-Ming. Web Mining of OOV translations. Journal of Information & Computational Science,2008,5: 1, 1-6.

[12] Li Haizhou, Zhang Min, Su jian. A Joint Source-Channel Model for Machine Transliteration. In Proc. Of 42th Annual Meeting of the Association for Computational Linguistics, Forum Convention Centre, Barcelona, 2004,160-167.

[13] Wen-Hsiang Lu, Jiun-Hung Lin and Yao-Sheng Chang. Improving Translation of Queries with Infrequent Unknown Abbreviations and Proper Names. Computational Linguistics and Chinese Language Processing, 2008,13(1), 91-120.

[14] Jian-Cheng Wu, Tracy Lin and Jason S.Chang . Learning Source-Target Surface Patterns for Web-based Terminology Translation. In Proceedings of the ACL Interactive Poster and Demonstration Sessions, June, 2005,37-40.

[15] http://zh.wikipedia.org/