利用网络挖掘技术建立英语学习平台

周明　刘晓华　蒋龙　Matt Scott
微软亚洲研究院 北京 100190
E-mail: mingzhou@microsoft.com

摘　要：本特邀报告将介绍我们最新的一个英语学习的垂直搜索平台英库（英文名 Engkoo）(http://www.engkoo.com)。我们开发了一套完整的网络挖掘技术，从数以百亿计的网页中获取大规模、多样化、新鲜的语言知识和翻译知识。然后在此基础上，构建了一个专门用来进行英语学习的网络服务，在一个垂直搜索平台下，提供词典释义、例句、词汇对照表示、拼写纠正、语音输出等各种功能。我们的目标是不论用户在翻译和写作中遇到任何问题，比如词汇和术语的定义和翻译，词汇的选择，句型的选择，词语的搭配，都可以在本系统中通过搜索我们在网络上获取的超大规模语言数据库，找到答案。我们的研究表明，互联网挖掘和发现对自然语言的研究和应用具有广泛的前景。我们同时展现利用 Engkoo 所建立的平台，使得机器翻译、跨语言检索、转述、深层语义搜索的研究巧妙地融合在一起，从而促进各项研究有机地融合并且做到有的放矢地发展。
关键词：英语学习，网络发掘，垂直搜索，翻译知识获取，机器翻译

Building an English Learning Service with Web Mining

Ming Zhou, Xiaohua Liu, Long Jiang, Matt Scott
Microsoft Research Asia, Beijing 100190
E-mail: mingzhou@microsoft.com

Abstract: Engkoo is a vertical search engine for using and exploring language – currently applied to English for Chinese users – however the technology itself is language independent. At a system level, Engkoo is an application platform that supports a multitude of NLP technologies such as cross language retrieval, alignment, paraphrasing, and statistical machine translation. The data set that supports this system is primarily built from mining a massive set of bilingual terms and sentences from across the web. Specifically, web pages that contain both Chinese and English are discovered and analyzed for parallelism, extracted and formulated into clear term definitions and sample sentences. This approach allows us to build the world's largest lexicon linking both Chinese and English together - at the same time covering the most up-to-date terms as captured by the net. In addition, our data set is intelligently merged with licensed data from sources including Microsoft Office and Encarta. Finally, the resulting vast, ranked, high quality composite data set is analyzed by a machine learning based classifier, allowing users to filter down sample sentences by combinable categories. A working system of Engkoo, currently exposing a subset of these technologies, is accessible via www.engkoo.com..

1　引言

语言学习尤其是英语学习具有巨大的市场价值。单单在中国，就有七千五百万人每天学习和使用英语。在日本、韩国等亚洲国家还有众多的欧洲国家把英语作为第二语言从而需要学习英语的用户更是数不胜数。已经有很多英语学习产品比如电子词典，各类网站等等。但是存在的主要问题是，由于词汇量比较小，而且不能反映新译，因此词典不能与时

俱进。比如，当美国选出新总统, Barack Obama, 如何翻译成中文？还有“一次性处理”，“小心碰头”，“水煮鱼”等如何翻译成英文？这些词汇的翻译很难在现有的词典得到答案。同时每一个词条针对每一个意义的例句比较有限，语言的细微用法难以展示。我们近十年来一直从事英语的写作和英语学习的研究，发展了一整套的从数十亿网页中挖掘翻译数据，包括词汇和例句的方法。试图覆盖所有的词汇和表达，并且定期挖掘，及时地发现最新的词汇和最新的翻译。在所挖掘的翻译数据的基础上，自动抽取词汇的搭配知识，词汇、搭配和句子级的转述，关键词、搭配和句子的翻译引擎。并且研究了多语言检索和多种语言搜索结果的合并和再排列。利用以上的数据和工具，我们建立了一个专门用于英语学习的垂直搜索服务（http://www.engkoo.com）。

2 系统的概况

Engkoo 的设计理念简单来讲就是从 web 中来，到 web 中去；数据驱动和自动学习；通过实施，用户参与系统的完善。具体来讲，除了人工编纂的词典和例句之外，其他所有的数据都是从 web 中自动获取而来。因此强大的数据挖掘系统就显得非常重要。然后利用统计机器学习，自动建立机器翻译引擎，转述引擎，搜索引擎。最后，利用基于实施的策略(deployment-based research)，把系统放在 web 上供用户使用。在使用的过程中，根据用户的反馈进一步修正系统的内核和用户界面，使得系统日趋完善。

3 双语数据（词典和例句）的获取

双语数据是建立机器翻译系统，跨语言检索和语言学习系统的重要资源。然而， 以有的双语数据要么已经过时，要么仅限于狭窄的领域，要么数据量不够。由于利用人工方式建立大规模的双语数据通常要耗费巨大的人工成本，近年来许多研究人员开始尝试从互联网上自动获取双语的资源。

目前获取双语例句的主要方法，比如(Nie et al., 1999; Shi et al., 2006), 都是试图首先获取双语的网站，然后获取双语的文档，再利用句子对齐技术获取双语的句子。获取双语的网站的方法大致可以分为基于互联网网页目录分类，利用特殊设计的 query 在搜索引擎上搜索得到候选的网站，利用模式匹配技术在数以十亿计的网页列表中找到可能的双语对照的网站。 我们在 2006 年的工作（Shi et al., 2006)提出利用 DOM Tree 对齐技术获取更多的网页。至于获取双语的词典，主要的方法是，比如（Cao et al.，2007；Lin et al. ，2008)的工作，利用了这样的观察：就是很多人在写文章的时候，会把英文的原词放在中文词的后面，加以括号。比如下面的网页。

泰坦尼克号邮船简介 [编辑本段]

英国皇家邮船泰坦尼克号(RMS Titanic)是奥林匹克级邮轮的第二艘邮船，20世纪初，由英国白星航运公司（White Star Line）制造的一艘巨大豪华客轮。由位于爱尔兰贝尔法斯特（Belfast）的哈兰德与沃尔夫（Harland and Wolff）造船厂兴建。泰坦尼克号是当时世界上最大的豪华客轮，被称为是“永不沉没的船”或是“梦幻之船”。泰坦尼克号共耗资7500万英镑，吨位46328吨，长882.9英尺，宽92.5英尺，从龙骨到四个大烟囱的顶端有175英尺，高度相当于11层楼,是当时一流的超级豪华巨轮。计划与姐妹船 奥林匹克号（RMS Olympic）和 不列颠尼克号（RMS Britannic）一道为英国白星航运公司的乘客们提供快速且舒适的跨大西洋旅行。

1912年4月10日，泰坦尼克号从英国南安普敦（Southampton）出发，途经法国 瑟堡-奥克特维尔（Cherbourg-Octeville）以及爱尔兰 昆士敦（Queenstown），计划中的目的地为美国的（New York），开始了这艘“梦幻客轮”的处女航。4月14日晚11点40分，泰坦尼克号在北大西洋撞上冰山（大约在41°43'55.66"N 49°56'45.02"W附近），两小时四十分钟后，4月15日凌晨2点20分沉没，由于缺少足够的救生艇，1500人葬生海底，造成了当时在和平时期最严重的一次航海事故，也是迄今为止最为人所知的一次海难。电影《泰坦尼克号》就是根据这一真实海难而改编。

图1　括号模式的双语词汇

利用以上的括号模式，可以获取大量的双语词汇。还有，在很多双语的网页里，有很多双语的数据被集中收集在一起。图二表示了一个网页的片段，全部是各种狗的名字。双语对照。这些词汇表并没有括号模式，但是他们服从统一的格式。比如“{Number}。{English name}{Chinese name}{EndOfLine}”. 同样，在句子的级别，也有这样的网页。

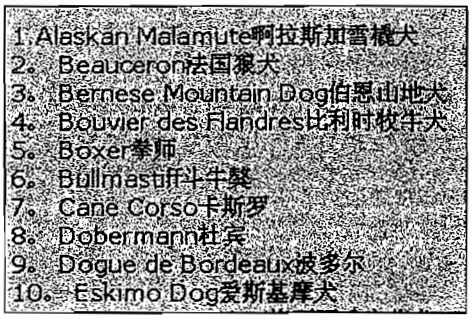


图 2　双语术语对照的网页

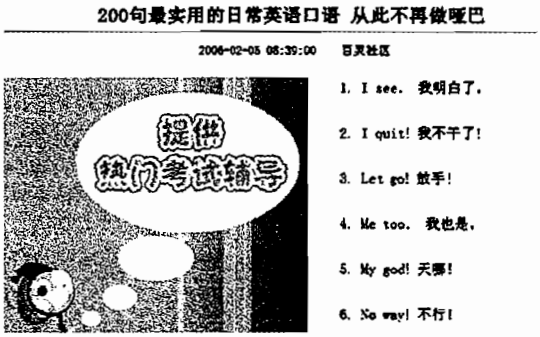


图3　双语句子对照的网页

根据我们的初步调查，这样几种的中英双语对照网页有千万计。而且每一个网页里都有大量的例句和词汇。但是获取起来并不是容易的事情。主要有如下的原因。

1)每一页的模式各不相同。因此不容易用一组实现定义好的模版抽取。

2)有一些网页虽然按照某些模版排列，但是却不是翻译对照。

3)在一个网页中，也许会存在多种对照模式。

由于这些问题，简单地用一个分类器从相邻的文本中选择翻译效果并不理想。我们提出了基于模版的方式：从集中式的双鱼网页中自动学习自适应的排列模版。具体来讲，包括如下步骤：

1)预处理：首先把网页分析称 DOM tree，然后把树中每一个节点的文本切分为若干片断;

2)种子模板的获取：使用一个兼顾翻译和音译的对齐模型来判断互为翻译的词汇对。

3)模式学习：学习通用的模版

4)利用所学习的模板抽取该网页的双语数据

详细介绍，参看（Long, et al, 2009）的文章。

4 Engkoo 的主要功能

目前的 Engkoo1.5 版包括了如下功能：

(1) 输入中文词或者短语，得到英文的翻译。同时得到一组例句。

比如输入“微软亚洲研究院”得到 Microsoft Research Asia, 以及 MSRA

(2) 输入英文词或者短语，得到中文的翻译. 同时得到一组例句。

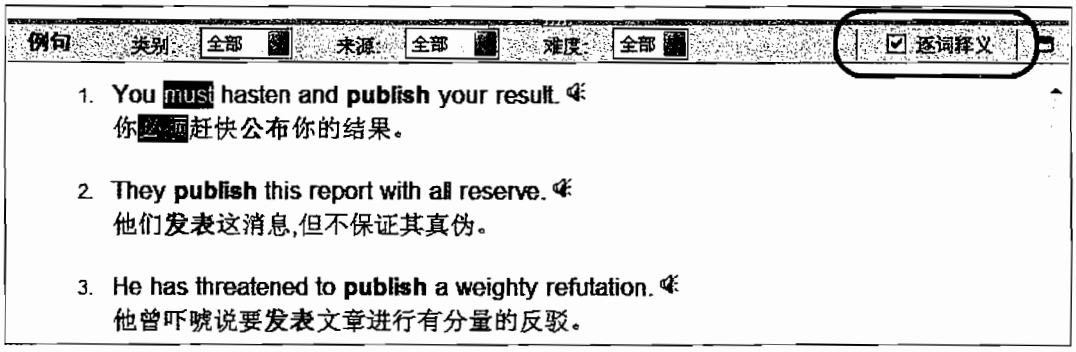
比如输入 swine flu 得到 “猪流感”

(3) 允许输入词性 wildcard 进行搜索。

比如 “adj. girl” 搜索满足形容词修饰名词 “girl”的例句。

1. She had been a **stunning girl**. 她从前是一个非常漂亮的姑娘。	2. He is squiring a **pretty girl**. 他正随侍在一位漂亮的姑娘身边。
3. The **little girl** soaked her clothes. 小女孩已把她的衣服泡上了。	4. She was a **dull-looking** country **girl**. 她是一个外表阴郁的乡村女孩。

(4) 例句可以显示词词的对应。方便英语的学习。



(5) TTS 朗读功能。可以用微软亚洲研究院开发的语音合成系统把英文句子朗读出来。

其他功能还包括：query 自动补全功能。同义词和近义词功能。今后还将继续增加更多的功能。

5 结论

本文简要地介绍了微软亚洲研究院自然语言组所从事的 engkoo 英语学习的垂直搜索项目（http://www.engkoo.com）。重点介绍了在互联网上获取双语数据和知识。今后将在本版本基础上陆续增加更加高级的功能，比如跨语言检索，机器翻译等等。

目前，该组主要从事多语言的分本分析和多语言的搜索引擎，机器翻译，问答系统和聊天机器人，以及中文对联诗词的自动生成（http://www.duilian.live.com）。有关详细介绍可参见主页：

http://www.msra.cn/Research/Group.aspx
http://research.microsoft.com/en-us/groups/nlc/

感谢

本工作是微软亚洲研究院自然语言组、成果转化组、语音组和人机界面组的很多同事共同努力的结果。在此表示感谢。

参 考 文 献

[1] Cheng, P., Teng, J., chen, R., Wang, J,. Lu, W., and Cheng, L. Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval. In the Proc. of SIGIR204, pp 162-169.

[2] Fung, P. and Yee, L. Y. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. COLLING-ACL, 1998 p414-420

[3] Huang, F., Zhang, Y., and Vogel, S.(2005) Mining Key phrase Translations from Web Corpora. In the Proceedings of HLT-EMNLP2005

[4] L Jiang, M Zhou, L Chien, C Niu. Named Entity Translation with Web Mining and Transliteration, Proceedings of the 20th IJCAI-07, pp 1629-1634

[5] D.Lin, S.Zhao, B.Durme and M.P. Mining Parenthetical Translations from the Web by Word Alignment. Proceedings of ACL-08, pp 994-1002

[6] Lu, W. and Lee, H. (2004). Anchor text mining for translation of Web queries: A transitive translation approach. ACM transactions on Information Systems, Vol.22, April 2004, pages 242-269.

[7] D.Marcu, W.Wong, a phrase-based, joint probability model for statistical machine translation, Proceedings of EMNLP, 2002, pp 133-139

[8] Nie, J-Y., Simard, M., Isabelle, P., and Durand, R. Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of parallel Text from the Web. SIGIR 1999,pp.74-81

[9] Shao and Ng, 2004 Li Shao and Hwee Tou Ng. 2004.Mining new-word translations from comparable corpora. In Proc. of Coling 2004, pp. 618–624

[10] Lei Shi, Cheng Niu, Ming Zhou, Jianfeng Gao: A DOM Tree Alignment Model for Mining Parallel Data from the Web. ACL 2006

[11] Jung H. Shin , Young S. Han , Key-Sun Choi, Bilingual knowledge acquisition from Korean-English parallel corpus using alignment method: Korean-English alignment at word and phrase level, Proceedings of the 16th conference on Computational linguistics, August 05-09, 1996, Copenhagen, Denmark

[12] J.C.Wu, T.Lin, J.S.Chang, Learning Source-Target Sur-face Patterns for Web-based Terminology Translation, ACL Interactive Poster and Demonstration Ses-sions,. Pp 37-40, Ann Arbor, 2005

[13] Zhang, Y. and Vines, P. (2004). Using the Web for Au-tomated Translation Extraction in Cross-Language Information Retrieval. In the Proceedings of SIGIR 2004, pp. 162-169.

[14] C.-H. Li, M. Li, D. Zhang, M. Li, M. Zhou and Y. Guan. A Probabilistic Approach to Syntax-based Reordering for Statistical Machine Translation. *In Proceedings of ACL 2007.*

[15] L. Jiang, M. Zhou, L. Chien, and C. Niu. Named Entity Translation with Web Mining and Transliteration. *In Proceedings of IJCAI 2007.*

[16] D. Zhang, M. Li, C.-H. Li and M. Zhou. Phrase Reordering Model Integrating Syntactic Knowledge for SMT. *In Proceedings of EMNLP 2007.*

[17] Long, Shiquan Yang, Xiaohua Liu, Ming Zhou, Mining Bilingual Data from the Web with Adaptively Learnt Patterns, ACL 2009

[18] S.Zhao, C. Niu, M. Zhou Combining Multiple Resources to Improve SMT-based Paraphrasing Model *In Proceedings of ACL 2008*