文章编号：1003-0077(2008)06-0103-07

一种有效的基于 Web 的双语翻译对获取方法

郭 稷[1]，吕雅娟[2]，刘 群[2]

(1. 北京大学 软件与微电子学院，北京 102600；
2. 中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190)

摘 要：命名实体和新词、术语的翻译对机器翻译、跨语言检索、自动问答等系统的性能有着重要的影响，但是这些翻译很难从现有的翻译词典中获得。该文提出了一种从中文网页中自动获取高质量双语翻译对的方法。该方法利用网页中双语翻译对的特点，使用统计判别模型，融合多种识别特征自动挖掘网站中存在的双语翻译对。实验结果表明，采用该模型构建的双语翻译词表，TOP1 的正确率达到 82.1%，TOP3 的正确率达到 94.5%。文中还提出了一种利用搜索引擎验证候选翻译的方法，经过验证，TOP1 的正确率可以提高到 84.3%。

关键词：计算机应用；中文信息处理；双语翻译对；统计判别模型；网络挖掘

中图分类号：TP391 **文献标识码**：A

An Effective Method to Extract Translation Pairs from Web Corpora

GUO Ji[1], LV Ya-juan[2], LIU Qun[2]

(1. School of Software and Microelectronics, Peking University, Beijing 102600, China;
2. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The translations of named entities, out of vocabulary words and terms play an important role in many application systems such as machine translation, cross-language information retrieval and question answer. However, these translations are hard to access from traditional bilingual dictionary. This paper proposes a method to automatically extract high quality translation pairs from Chinese web corpora. It analyzes the features of bilingual translation pairs in web pages, and then a statistical discriminative model combined with multiple features is used to extract translation pairs. Experimental results show that the quality of the extracted bilingual translations is improved greatly: Top1 accuracy 82.1%, and Top3 94.5%. The paper also proposes a verification method to further improve the accuracy of the initial extractions with the help of search engines. Top1 accuracy grows up to 84.3% after the verification.

Key words: computer application; Chinese information processing; bilingual translation pairs; statistical discriminative model; web mining

1 引言

随着互联网的普及和发展，互联网已经成为人们获取知识的主要来源。近几年，中文成为世界上网页数量增长速度最快的语种。据百度数据显示，到 2005 年底，中文网页总数达到约 24 亿。互联网上的中文资源越来越丰富。同时，由于国际化的需要，越来越多的中文网站成为双语网站。许多网站都加入了双语甚至多语信息。互联网已经成为获取双语或多语翻译资源的巨大来源。

双语翻译词典是重要的翻译资源。由于易实现和翻译词典的可读性，基于词典的方法被很多机器翻译应用，如跨语言信息检索中被广泛采用。但是

收稿日期：2008-06-05 **定稿日期**：2008-09-03

基金项目：国家自然科学基金资助项目(60603095)

作者简介：郭稷(1983—)，男，硕士生，主要研究方向为自然语言处理；吕雅娟(1972—)，女，博士，副研究员，主要研究方向为自然语言处理、机器翻译；刘群(1966—)，男，博士，研究员，主要研究方向为自然语言处理、机器翻译。

传统的双语词典通常不包含新词术语以及人名、地名等命名实体的翻译，而这些词的翻译对于机器翻译、跨语言信息检索、自动问答等系统的性能有着重要的影响。利用互联网丰富的资源，研究大规模、高质量的双语翻译对自动获取方法，已经成为目前的研究热点。前人在双语翻译资源获取方面做了很多尝试。搜索引擎、双语平行语料库和中文网页是获取双语翻译资源的主要来源。本文研究了一种有效地从中文网页中获取高质量双语翻译对的方法。该方法利用网页中双语翻译对的特点，使用统计判别模型，融合多种识别特征自动挖掘网站中存在的双语翻译对。实验证明，使用该方法可以有效地获得高质量的双语翻译对。

2 相关工作

在获取双语翻译知识方面已经存在一些研究工作。

Zhang[1]，Huang[2]提出利用搜索引擎的返回结果来获取双语翻译知识，他们使用不同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。他们的方法可以获得较好的翻译，但是由于搜索引擎的限制，这种方法不易用于获取大规模双语翻译资源。双语平行语料库已被用于构建大规模双语翻译词典。Huang[3,4]从句子对齐的双语语料库中训练双语命名实体之间的多特征的统计对齐模型，然后利用统计对齐模型进行双语翻译对的抽取。实验证明，他们的方法效果令人满意，然而高质量的双语平行语料库不太容易获取。张永臣[5]利用词间关系矩阵法从特定领域非平行语料中抽取双语词典。其中种子词的选择对抽取结果影响较大，抽取出来的双语词典的质量不高。

Zhang[6]在研究过程中发现，在中文网页中，如果英文出现在括号中，那么周围的中文很可能是其对应的翻译。她将出现在括号中的英文前面的中文分为两种情况：一种是前面的中文出现在书名号或者引号当中，例如，“东亚奇迹”(East Asia Miracle)，《银行保密法》(Bank Secrecy Act)，「独立公投」(independence referendum)；另一种是前面的中文不出现在书名号或者引号中，如据考克斯新闻社(Cox news service)。具有这样特征的中文网页是一个获取大量双语翻译对的潜在来源。Cao[7]在大规模的中文网页上做了相应研究。他们训练一个音译对齐判别模型用于音译对的抽取，然后训练一个翻译判别模型用于翻译对的抽取。然而网页内容的复杂性影响了音译判别模型的效果，例如博格斯(Tom Burgis)，前面的中文往往只是英文中一个单词的音译。确定用于音译对齐判别的中文和英文，不仅烦琐而且容易出错。此外，日文名也不能用于音译对齐判别。实验结果显示，他们抽取出来的音译对和翻译对正确率较低，质量不能令人满意。

本文与Cao[7]的研究相似，希望能够在中文网页中抽取双语翻译对。与Cao[7]的工作不同的是本文采用统计判别模型Perceptron对候选翻译进行训练和识别，其优点是可以有效地融合多种特征。实验证明本文方法有效提高了双语翻译对抽取的正确率。

3 双语翻译对获取

3.1 术语定义

本节开始将要使用的术语有：

候选行，是指中文网页中的一行中文，其中有英文出现在括号中。

固定格式翻译对，是指在候选行中，当英文出现在括号中时，其前面的中文出现在书名号或者引号中的中英文翻译对。

候选翻译单元，是指候选行中抽取出来的不包含非法翻译字符(如。，！等)的中文文字。

候选翻译对，是指从候选翻译单元生成的用于翻译判别的中英文。

下面是一个简单的例子。在候选行“那就是他要承担责任。这也正是丹尼尔·布东(Daniel Bouton)”中，候选翻译单元是“这也正是丹尼尔·布东(Daniel Bouton)”，可能产生的一个候选对是“布东(Daniel Bouton)”。

3.2 候选翻译对生成

从一个候选翻译单元中抽取正确的双语翻译对，可以归结为中文边界划分问题。因为英文已经出现在括号中，要找到正确的翻译对，只需在英文前面的中文中划分出正确的边界，边界之内的中文就认定为英文的翻译。为了找到正确的边界，我们使用中文分词工具对候选翻译单元的中文进行切分，然后组合切分得到的词构成候选翻译对。例如，经过切分后的候选翻译单元是“足球/教练/佐/夫(Zolf)”，那么可以构成下面四个候选翻译对：

夫　Zolf

佐夫　Zolf

教练佐夫　Zolf

足球教练佐夫　Zolf

可以看出，一个候选翻译单元可以生成多个候选翻译对。这样，我们就把中文边界划分问题转化为从多个候选翻译对中选择正确的翻译对问题。由于当前中文分词系统具有很高的精度，候选翻译对中基本上会包含正确的翻译。然而，还是会存在切分错误造成候选翻译对中不包含正确翻译的情况。不使用分词工具来确定划分边界的问题，将留给下一步研究工作。

3.3　翻译判别模型

翻译判别模型是一个基于多特征的判别式模型。设 T 是候选翻译对的集合。t_i 表示第 i 个候选翻译对，其特征表示为 $f_k(t_i)$。t_i 的得分如公式(1)所示：

$$Score(t_i) = \sum_{k=1}^{K} f_k(t_i) \times \lambda_k \qquad (1)$$

λ_k 是 $f_k(t_i)$ 对应的权值。K 是总的特征个数。翻译判别模型计算每个候选翻译对的得分，然后以得分最高的候选翻译对作为翻译对抽取的结果。

3.4　特征选择

我们先对候选翻译单元的中文部分进行分词、词性标注和命名实体识别，然后选取了以下特征：

1. 候选翻译共现频率

在生成候选翻译对时，我们将具有相同英文翻译的中文放在一起统计。某候选翻译和英文的共现频率越高，它越可能成为该英文的翻译。

2. 候选翻译的长度

候选翻译的长度是指候选翻译包含的汉字个数，长度过长或过短，其成为英文翻译的可能性就越小。

3. 是否是命名实体

如果某个英文是一个命名实体，那么候选翻译中的命名实体就很可能成为其翻译。

4. 是否包含“·”

在外国人名全称的翻译中，·号是姓和名的分隔标志。如果候选翻译包含这个符号，该候选翻译可能包含了外国人名的全称翻译。这个特征可保证外国人名全称的翻译不会丢失。

5. 候选翻译首词的词性

以名词、形容词等开头的候选翻译成为对应英文翻译的可能性比以介词、连词等开头的候选翻译大。

6. 候选翻译前一个词的词性

在中文里面，尤其是中文网页中，对于特定的词性而言，如介词、连词、助词等，其后面的中文成为相应英文的翻译的概率较大。

7. 候选翻译前一个词

在中文网页中，有些词语带有明显的暗示信息，其后面的中文是对应的英文的翻译。比如“基地组织高级指挥官利比(Abu Laith al-Libi)”，“财政部长梅隆(Andrew Mellon)”，“已经与英国航空(British Airlines)”，其中，“指挥官”、“部长”、“与”就带有很好的暗示信息。

3.5　模型训练

我们利用感知机来训练翻译判别模型。

对于频率特征，我们将其离散化，转换成二值特征。将翻译实例出现的频率分为 7 个等级：等级 1 是出现 1 至 2 次，等级 2 是出现 3 至 5 次，等级 3 是出现 6 至 8 次，以此类推，最后等级 7 是出现 18 次以上。这样，频率的特征函数如公式(2)所示。

$$f_k(x) = \begin{cases} 1 & \text{if } x \text{ 的频率等级是 } 1 \\ 0 & \text{else} \end{cases} \qquad (2)$$

对于长度特征，采用类似的离散化方法，将其转换成二值特征。这样，整个模型使用的全是二值特征。

由 3.2 节的叙述可以知道，一个候选翻译单元会产生多个候选翻译对。由于我们将具有相同英文的中文候选对放在一起统计，其产生的候选翻译对的数目可能会很大。而在众多的候选翻译对中，只有一个正确翻译对，因此训练数据是倾斜的，不利于感知机的训练。为了克服训练数据倾斜问题，受 Reranking 思想[8,9]的启发，我们将训练过程看成一个类似于重排序的过程。首先将具有相同英文的训练实例划为一组。在每轮训练中，对于每组训练实例，计算其中每个训练实例的得分，然后选出得分最高的实例。如果得分最高的训练实例是正例(即正确翻译对)，则继续进行下一组实例训练，不调整参数；如果是负例，则调整权值参数。整个训练过程迭代进行，直到满足收敛条件。训练过程如图 1 所示。

输入：训练实例集 $S_i(i=1\cdots N)$

输出：权值向量 $\vec{\lambda}$

1. 初始化权值 $\lambda_k^1=1(k=1,2\cdots K)$
2. for $i=1$ to N
3. 计算第 i 组每个训练实例得分 $\text{Score}(s_{i,j})$
4. $q=\arg\max Score(s_{i,j})$
5. if $s_{i,q}$是负例
   设 $s_{i,p}$是正例
6. $\lambda_k^{i+1}=\lambda_k^i+\eta(f_k(s_{i,p})-f_k(s_{i,q}))k=1,2\cdots K\ \eta=0.001$
7. 重复步骤 2 直至收敛

图 1 翻译判别模型训练

4 翻译对抽取实验

4.1 数据准备

实验使用的网页数据分为训练网页数据和测试网页数据。训练网页数据来自《联合早报》、《欧洲时报》、《华盛顿观察报》三个网站。测试网页数据分为两部分。一部分和训练网页数据一样，来自于《联合早报》网站。为了测试系统的性能，另一部分来自于与训练网页数据来源不同的“星岛环球网”。表 1 显示了实验的网页数据情况。

表 1 实验网页数据

	来 源	大小
训练集	《联合早报》《欧洲时报》《华盛顿观察报》	1.6 G
测试集	《联合早报》	835 M
	“星岛环球网”	858 M

4.2 预处理

对于训练网页数据，首先抽取候选行，然后从中获得固定格式翻译对，接着进行候选翻译单元抽取。在生成候选翻译对前，我们使用 ICTCLAS 对候选翻译单元的中文进行分词、词性标注和命名实体识别。最后生成候选翻译对。表 2 列举了一个实例，说明处理后的一个候选翻译对的各个特征。

经过预处理，训练网页数据一共产生了 99 082 个候选翻译对。对于测试网页数据，采用相同的预处理方式。

表 2 举例：当前翻译对是“受到卡特里娜飓风(Katrina)”，对应的各个特征值。

特征模板	特征值	注 释
FQ_i (i=1…7)	FQ_1 = 1	频率是 1，频率等级是 1
LEN_i (i=1…7)	LEN_5=1	长度是 16，长度等级是 5
NE	NE = 0	不是 NE
DOT	DOT = 0	不包含 ·
FT_pos	FT_v = 1	首词词性是 v
PT_pos	PT_p = 1	前一词词性是 p
PW_word	PW_在 = 1	前一个词是“在”

FQ_i：频率等级是 i 的特征值；LEN_i：长度等级是 i 的特征值；NE：是否是命名实体的特征值；DOT：是否包含 · 的特征值；FT_pos：首词词性是 pos 的特征值；PT_pos：前一词词性是 pos 的特征值；PW_word：前一个词是 word 的特征值

4.3 翻译判别模型训练和测试

将训练网页数据生成的候选翻译单元，按照 4∶1 的比例，构造翻译判别模型的训练实例集和开发实例集。在训练实例集中，人工标注了 29 953 个训练实例。按照图 1 所示的训练方法进行模型训练。在每轮迭代训练之前，我们随机产生 N 组训练实例的训练顺序，这样可以避免训练数据过拟合问题。实验中，迭代过程的收敛条件是前后两次训练的正确率小于阈值 0.000 1。

开发实例集一共包含 19 661 个实例。由于在训练过程中，每轮迭代时训练实例的顺序都不同，所以每次训练产生的模型也不同。开发集就是用于从训练得到的多个模型中选取最佳的训练模型。另外，实验的目的是从中文网页中抽取双语翻译对。对抽取出来的双语翻译对质量的评价客观上表现了翻译判别模型的性能，所以实验中不使用测试实例集。

表 3 说明了训练集和开发集的实例数据。表 4 显示了翻译判别模型在开发集上使用不同特征的实验结果。我们发现，在候选翻译识别中，首词词性特征是一个重要的特征，去掉首词词性，正确率下降了 6.1 个点。长度特征也起到了重要的作用。去掉长度特征，正确率下降了 4.1 个点。接下来，去掉前一个词的特征，正确率下降了 3.2 个点。去掉是否是命名实体的特征带来了 3 个点的正确率下降。频率特征起到了一定的作用，去掉该特征，带来了 2.7 个点的正确率下降。前一个词词性的特征所起的作用较小，去掉该特征使得正确率下降了 1.5 个点。最

后，是否包含·的特征起到的作用最小，去掉该特征正确率只下降了 0.3 个点。我们认为，该特征的主要作用在于获取外国名字的全称翻译，在候选翻译对方面的判别能力不如其他特征。

表 3　翻译判别模型数据集

	实例数目(个)
训练集	29 953
开发集	19 661

表 4　不同特征下翻译判别模型实验结果

特征	开发集正确率(%)	特征	开发集正确率(%)
All	76.0	All-DOT	75.7
All-FQ	73.3	All-FT	69.9
All-LEN	71.9	All-PT	74.5
All-NE	73.0	All-PW	72.8

All 表示所有特征，FQ 表示频率特征，LEN 表示长度特征，NE 表示是否是命名实体，DOT 表示是否包含·，FT 表示首词词性，PT 表示前一个词词性，PW 表示前一个词。

4.4　翻译对抽取

利用上面训练得到的翻译判别模型，我们在测试网页数据上进行了两组实验，测试抽取的翻译对的正确率。

第一组实验的网页数据的来源和训练数据相同，来自于《联合早报》网站。经过预处理，得到了 1 181 个固定格式翻译对。经验证，准确率为 98.4%。使用翻译判别模型，我们从剩下候选翻译对中抽取了 5 118 个翻译对。随机选择了 600 个翻译对进行人工验证，实验结果如表 5 所示。

表 5　《联合早报》网站翻译对质量

	翻译对正确率(%)
TOP 1	78.5
TOP 2	89.7
TOP 3	93.7

为了验证翻译判别模型的健壮性，在第二组实验中，我们使用了与训练网页数据来源完全不同的网页数据。该组实验的网页来自于“星岛环球网”。预处理后，得到 666 个固定格式翻译对，经验证，准确率为 96.9%。使用翻译判别模型，我们从剩下的候选翻译对中共抽取了 4 267 个翻译对。随机选择 600 个翻译对进行人工验证，实验结果如表 6 所示。

表 6　“星岛环球网”网站翻译对质量

	翻译对正确率(%)
TOP 1	79.8
TOP 2	86.7
TOP 3	89.7

表 5 和表 6 中，翻译对正确率是指从候选翻译对中抽取出来的翻译对的正确率。TOP1 是指每个翻译对取得分最高的中文翻译作为翻译抽取结果。TOP2 和 TOP3 是指以得分最高的前两个和前三个中文翻译作为翻译抽取结果。在表 5 中，TOP1 的翻译对正确率达到了 78.5%。TOP3 的翻译对正确率达到了 93.7%。在表 6 中，TOP1 的翻译对正确率达到 79.8%，TOP3 的翻译对正确率达到了 89.7%。翻译对的质量令人满意。两组不同的实验表明，在不同来源的中文网页中，我们的方法可以抽取高质量的双语翻译对。附录 A 是我们获取的双语翻译对的一些实例。

另外，抽取的双语翻译词表的正确率是指加上固定格式翻译对后总的翻译对的正确率。第一组实验中，TOP1 的总正确率为 82.1%，TOP3 的总正确率为 94.5%。第二组实验中，TOP1 的总正确率是 82.1%，TOP3 的总正确率是 90.6%。

5　翻译对后验证

从表 5 和表 6 可以看到，TOP3 的翻译对正确率比 TOP1 的正确率高出很多。这意味着，使用第 3 节介绍的翻译判别模型从中文网页中抽取翻译对，绝大部分的正确结果都在得分最高的前 3 个候选翻译中。相对而言，TOP1 的正确率比较低。那么，对翻译判别模型抽取出来的候选翻译结果进行验证，选出正确的结果，提高 TOP1 的正确率，会是非常有意义的事情。本节介绍一种借助搜索引擎进行后验证的方法，其思路来源于跨语言检索领域中解决 OOV 翻译问题的方法[2]。

我们使用第一组实验中随机选择的 600 个翻译对作为初始抽取结果，其中每个英文包含得分最高的 3 个中文翻译。后验证的过程就是利用搜索引擎的返回结果，从这 3 个候选中文翻译中选择正确的翻译。每个英文及其对应的 3 个中文翻译构成一组，一共 600 组实验数据。验证过程如图 2 所示。

输入：600组实验数据，每组包含英文及其3个中文候选翻译
输出：600组实验数据，每组包含英文及其中文翻译

1. 对于每组实验数据 T_i，$c_{i,j}$表示第j个候选翻译（$i=1\cdots600$，$j=1,2,3$）
2. for $j=1$ to 3
3. 使用 $c_{i,j}$和英文e构成查询词，交给搜索引擎，取返回的前100个结果
4. 统计 $c_{i,j}$在返回结果中的出现频率 $f_{c_{i,j}}$
5. $S(c_{i,j})=score(c_{i,j})+\beta\cdot\mu\cdot f_{c_{i,j}}$
6. $l=\arg\max_j S(c_{i,j})$，取 $c_{j,l}$为英文e的中文翻译

图2 后验证过程图

图2中，$score(c_{i,j})$表示候选翻译$c_{i,j}$在翻译判别模型下的得分。μ是$f_{c_{i,j}}$的调整因子，目的是将$f_{c_{i,j}}$的值调整到合适的数量级。实验中取$\mu=0.001$。β是$f_{c_{i,j}}$的加权系数，它的值使用类似于最小错误率训练的方法进行调整。实验中$\beta=0.4$。经过后验证，对这600个翻译对重新进行人工验证，TOP1的翻译对正确率为81.7%。可见，后验证提高了TOP1的正确率。

6 未来工作展望

未来的工作可从以下三个方面开展：

1. 在3.2节中，我们提到，从候选翻译单元中抽取正确的翻译对，就是中文边界划分问题。本文采用的方法是将英文前面的中文进行分词，然后逐个组合，形成多个候选翻译对，再从候选翻译对中选出正确的翻译对。然而，中文分词的错误，会造成候选翻译对中不包含正确的候选翻译对。在实验使用的开发集的候选翻译对有2 811组，其中有12组由于分词错误造成没有包含正确的翻译对。研究中不使用的中文分词边界划分的方法，将是下一步的工作。

2. 到目前为止，研究人员从中文网页中抽取双语翻译对，获取中文网页资源的方法有两种：一是在大规模的中文网页中去抽取双语翻译对，如Cao的实验，他们使用的中文网页达到300G；另一种是人工识别特定网站，然后下载相应网页。本次实验处理的是英文出现在括号当中的中文网页，因此我们人工地寻找具有这种特点的网站，然后下载网页。研究自动识别具有这种特点的网站的有效方法，将是很有意思的事情。

3. 本文对利用搜索引擎进行后验证的方法进行了初步尝试，实验证明，这种方法能够提高双语翻译对抽取的正确率。但是后验证方法的效果还有待改进。搜索引擎返回结果的处理，包括词频统计，去重问题等，将对后验证方法效果产生重要的影响。

7 结论

本文提出了一种从中文网页中抽取双语翻译对的有效方法。我们目前只处理有英文出现在括号中的中文网页。首先，对中文网页进行预处理，然后对候选翻译单元的中文进行分词、词性标注和命名实体识别，最后得到候选翻译对。我们训练一个翻译判别模型，并使用这个模型从候选翻译对中抽取翻译对。实验结果表明，翻译判别模型对于不同来源的中文网页是健壮的。抽取的翻译对正确率高，质量令人满意。此外，我们还提出了利用搜索引擎进行翻译对后验证的方法。经过后验证，抽取的翻译对的正确率得到了进一步的提高。

参考文献：

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

[2] F. Huang, Y. Zhang and S. Vogel. Mining Key Phrase Translations from Web Corpora[C]//the Proceedings of HLT-EMNLP 2005: 483-490.

[3] F. Huang, S. Vogel and A. Waibel. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization[C]//the Proceedings of ACL 2003 workshop on Multilingual and mixed-language named entity recognition, 9-16.

[4] F. Huang and S. Vogel. Improved Named Entity

Translation and Bilingual Named Entity Extraction [C]//the Proceedings of ICMI 2002, 253-258.

[5] 张永臣，孙乐，等. 基于 Web 数据的特定领域双语词典抽取[J]. 中文信息学报，2006,20(2)：16-23.

[6] Y. Zhang and P. Vines. Detection and Translation of OOV Terms Prior to Query Time[C]//the Proceedings of SIGIR2004,524-525.

[7] G. H. Cao, J. F. Gao and J. Y. Nie. A System to Mine Large-Scale Bilingual Dictionaries from Monolingual Web Pages[C]//MT Summit XI, 57-64.

[8] M. Collins and N. Duffy. New Ranking Algorithms for Parsing and Tagging: Kernel over Discrete Structures, and the Voted Perceptron[C]//the Proceedings of ACL2002, 263-270.

[9] M. Collins and T. Koo. Discriminative Reranking for Natural Language Parsing[J]. Computational Linguistics,2005,31 25-70.

附录 A　双语翻译对示例

类　型	中　文	英　文
人名	阿当·阿济兹	Aadam Aziz
	加海姆	Peter Garnham
	卢武铉	Roh Moo Hyun
	中川昭一	Shoichi Nakagawa
地名	罗尼湾	Rodney Bay
	卓源路	Toh Guan Road
	王府井书店	Wangfujing Bookstore
	阿巴拉契亚中心山脉	Central Appalachian
机构名	国家环境保护总局	State Environmental Protection Administration
	统一俄罗斯党	United Russia Party
	埃塞克斯大学	University of Essex
电影、报刊、书籍	《得克萨斯城太阳报》	Texas City Sun
	《爱上大姐大》	Marrying the Mafia
	《广角镜》	Panorama
新词	新加坡式英语	Singlish
	就业人息补助	Workfare Income Supplement
	血管成形术	angioplasty
	“菠菜人”	Spinach Man

（上接第 102 页）

[12] Jinxi Xu and W. Bruce Croft, Query Expansion Using Local and Global Document Analysis[C]//SIGIR '96, Zurich, Switzerland, Aug. 18-22, ACM Press, New York, NY: 4-11.

[13] Iadh Ounis, Maarten de Rijke, Craig Macdonald, Gilad Mishne, Ian Soboroff. Overview of the TREC-2006 Blog Track[C]//TREC 2006.

[14] Craig Macdonald, Iadh Ounis, Ian Soborof, Overview of the TREC2007 Blog Track[C]//TREC 2007.

[15] Ricardo Baeza-Yates, Berthier Ribeiro-Neto Modern Information Retrieval[M]. Boston, MA, USA: Addison-Wesley Longman Publishing Co. 1999, 117-118.

[16] Christopher J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery,1998. 2: 121-167.

[17] 代六玲，黄河燕，陈肇雄. 一种文本分类的在线 SVM 学习算法[J]，中文信息学报，2005,19(05)：11-15.

[18] Craig Macdonald and Iadh Ounis. The TREC Blog06 Collection: Creating and Analysing a Blog Test Collection [R]. DCS Technical Report TR-2006-224. Department of Computing Science, University of Glasgow. 2006.

一种有效的基于Web的双语翻译对获取方法

作者： 郭稷， 吕雅娟， 刘群， GUO ji， LV Ya-juan， LIU Qun
作者单位： 郭稷,GUO ji(北京大学,软件与微电子学院,北京,102600)， 吕雅娟,刘群,LV Ya-juan,LIU Qun(中国科学院,计算技术研究所,智能信息处理重点实验室,北京,100190)
刊名： 中文信息学报 ISTIC PKU
英文刊名： JOURNAL OF CHINESE INFORMATION PROCESSING
年，卷(期)： 2008，22(6)
引用次数： 0次

参考文献(9条)

1. Y.Zhang. P.Vines Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval

2. F.Huang. Y.Zhang. S.Vogel Mining Key Phrase Translations from Web Corpora

3. F.Huang. S.Vogel. A.Waibel Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization

4. F.Huang. S.Vogel Improved Named Entity Translation and Bilingual Named Entity Extraction

5. 张永臣. 孙乐. 李飞. 李文波. 西野文人. 于浩. 方高林 基于Web数据的特定领域双语词典抽取[期刊论文]-中文信息学报 2006(2)

6. Y.Zhang. P.Vines Detection and Translation of OOV Terms Prior to Query Time

7. G.H.Cao. J.F.Gao. J.Y.Nie A System to Mine Large-Scale Bilingual Dictionaries from Monolingual Web Pages

8. M.Collins. N.Duffy New Ranking Algorithms for Parsing and Tagging:Kernel over Discrete Structures,and the Voted Perceptron

9. M.Collins. T.Koo Discriminative Reranking for Natural Language Parsing 2005

相似文献(10条)

1. 期刊论文 李宁 XML--中文信息处理的变革之路 -中文信息学报2003,17(2)
本文从中文信息面临的问题出发,阐述了中文信息处理走Internet开放变革之路的必要性.文中还介绍了Intemet上已经开展的与中文信息处理相关的部分工作,重点论述了XML在中文信息处理方面的优势,指出以XML为基础的Web服务是分布式环境中文信息处理技术的发展方向.作者为此提出了一个中文信息处理服务体系框架的构想.

2. 期刊论文 张玮. 孙乐. 冯元勇. 李文波. 黄瑞红. ZHANG Wei. SUN Le. FENG Yuan-yong. LI Wen-bo. HUANG Rui-hong 词汇搭配和用户模型在拼音输入法中的应用 -中文信息学报2007,21(4)
中文输入法是中文信息处理的难题之一.随着互联网上中文用户的不断增加,中文输入法的重要性也变得日益突出.本文在对句子中长距离词汇依赖现象观察的基础上,抽取出语料库中的词汇搭配来获取长距离特征,并以此构建基于词汇搭配关系的拼音输入法系统;同时将词汇搭配的思想应用到拼音输入法的用户模型中,从而使我们的输入法系统能够辅助用户更加有效的输入.实验表明基于词汇搭配关系的改进方法对提高输入法的准确率有积极的作用.

3. 期刊论文 唐慧丰. 谭松波. 程学旗. TANG Hui-feng. TAN Song-bo. CHENG Xue-qi 基于监督学习的中文情感分类技术比较研究 -中文信息学报2007,21(6)
情感分类是一项具有较大实用价值的分类技术,它可以在一定程度上解决网络评论信息杂乱的现象,方便用户准确定位所需信息.目前针对中文情感分类的研究相对较少,其中各种有监督学习方法的分类效果以及文本特征表示方法和特征选择机制等因素对分类性能的影响更是亟待研究的问题.本文以n-gram以及名词、动词、形容词、副词作为不同的文本表示特征,以互信息、信息增益、CHI统计量和文档频率作为不同的特征选择方法,以中心向量法、KNN、Winnow、Naive Bayes和SVM作为不同的文本分类方法,在不同的特征数量和不同规模的训练集情况下,分别进行了中文情感分类实验,并对实验结果进行了比较,对比结果表明:采用BiGrams特征表示方法、信息增益特征选择方法和SVM分类方法,在足够大训练集和选择适当数量特征的情况下,情感分类能取得较好的效果.

4. 会议论文 吕肖庆. 北京大学文字信息处理技术国家重点实验室. 尹江红. 北京大学文字信息处理技术国家重点实验室. 汪宾成. 北京大学文字信息处理技术国家重点实验室. 张建国. 北京大学文字信息处理技术国家重点实验室. 赵学亮. 北京大学文字信息处理技术国家重点实验室. 任力. 北京大学文字信息处理技术国家重点实验室 超大字库及其相关技术的研制 2002
自计算机发明以来,汉字集合的选择、组织形式、特别是汉字编码问题,曾长期困扰着中文信息科技的发展.直到八十年代初,正式确立的中文简体字国家标准(GB2312)后,中文信息才有了统一的交换平台,应用软件也才得以蓬勃发展.但是,由于硬、软件环境的限制,以及编码工作本身的复杂性,虽然继

GB2312标准之后推出了扩展标准(GBK)和GB18030,但常规的应用软件只能处理2万多个汉字,即使普通人在使用电脑时都可能遇到一些生僻字无法处理,比如人名、地名以及一些专用名词,而面对浩如烟海的中华古籍,2万余字的处理能力远远不够,长久以来一直让编纂辞书、整理古籍的专业人士扼腕痛惜.为了让源远流长的华夏文明能够凭借先进的计算机技术发扬光大,我们研制了超大字库及其相关的应用技术,不仅彻底解决了大量汉字的编码、显示问题,还经过长期积累,探索出超大字库录入的全新方法,并配备了排版、检索等工具.在中文信息处理方面为专业出版单位开拓了更为广阔的应用领域.近年来,该成果已不仅成功应用于古籍、辞书的编排与印刷,而且在医疗保险、户籍管理和历史档案的检索等方面,建立起了高水信的应用系统.

5.期刊论文 李茹.王文晶.梁吉业.宋小香.刘海静.由丽萍.LI Ru.WANG Wen-jing.LIANG Ji-ye.SONG Xiao-xiang.LIU Hai-jing.YOU Li-ping 基于汉语框架网的旅游信息问答系统设计 -中文信息学报2009,23(2)

该文借助汉语框架网(Chinese FrameNet,简称CFN)在语义表达方面的独特优势,探讨用本体描述语言建立面向特定领域的汉语框架语义知识库,并且以旅游交通领域中问答系统设计为例分析方法的有效性.方法中首先利用TREC分类与本体分类相结合的方式为查询问句分类,然后提出基于CFN的问句分析策略,通过CFN语义分析得到问句中三元组:语义谓词、语义主体和语义客体,在问句分析的基础上从旅游本体知识库中对答案进行抽取并对答案处理,同时用本体编辑工具Protégé编码,实验证实方法是有效的.

6.期刊论文 江敏.肖诗斌.王弘蔚.施水才.JIANG min.XIAO Shi-bin.WANG Hong-wei.SHI Shui-cai 一种改进的基于《知网》的词语语义相似度计算 -中文信息学报2008,22(5)

中科院刘群的基于<知网>的词语相似度计算是当前比较有代表性的计算词语相似度的方法之一.在测试中我们发现对一些存在对义或反义的词语与同义、近义词语一样具有较高的相似度,一些明显相似的词反而相似度较低,如"美丽"与"贼眉鼠眼"的相似度为0.814 815,与"优雅"的相似度为0.788 360,"深红"与"粉红"的相似度仅为0.074 074,这将不利于进行词语的极性识别.基于文本情感色彩分析的需要,把词语相似度的取值范围规定为[-1,+1],在刘群论文的基础上,进一步考虑了义原的深度信息,并利用<知网>义原间的反义、对义关系和义原的定义信息来计算词语的相似度.在词语极性识别实验中,得到了较好的实验结果:P值为99.07%,R值为99.11%.

7.期刊论文 乐明.YUE Ming 汉语篇章修辞结构的标注研究 -中文信息学报2008,22(4)

汉语篇章修辞结构标注项目CJPL采用大陆主要媒体的财经评论文章为语料,依据修辞结构理论(Rhetorical Structure Theory,RST),定义了以标点符号为边界的篇章修辞分析基本单元和47种区分核心性单元的汉语修辞关系集,并草拟了近60页的篇章结构标注工作守则.这一工作目前完成了对97篇财经评论文章的修辞结构标注,在较大规模数据的基础上检验了修辞结构理论及其形式化方法在汉语篇章分析中的可移用性.树库所带有的修辞关系信息以及三类篇章提示标记的篇章用法特征,可以为篇章层级的中文信息处理提供一些浅层语言形式标记的数据.

8.期刊论文 冯元勇.孙乐.李文波.张大鲲.FENG Yuan-yong.SUN Le.LI Wen-bo.ZHANG Da-kun 基于单字提示特征的中文命名实体识别快速算法 -中文信息学报2008,22(1)

近年来条件随机场(CRF)模型在自然语言处理中的应用越来越广泛.标准的线性链(Linear-chain)模型一般采用L-BFGS参数估计方法,收敛速度慢.本文在分析模型复杂度的基础上提出了一种改进的快速CRF算法.该算法通过引入小规模单字特征降低特征的规模,并通过在推理过程中引入任务相关的人工知识压缩Viterbi和Baum-Welch格搜索空间,提高了训练的速度.在中文863命名实体识别评测语料和SIGHAN06语料集上进行的实验表明,该算法在不影响中文命名实体识别精度的同时,有效地降低了模型的训练代价.

9.期刊论文 冯元勇.孙乐.董静.李文波.FENG Yuan-yong.SUN Le.DONG Jing.LI Wen-bo 基于分类信心重排序的中文共指消解研究 -中文信息学报2007,21(6)

共指消解是自然语言处理的核心问题之一.本文针对分步消解中分类器全局信息的不足,依据分类信心对全体提及配对进行排序,优先根据可靠的分类结果对提及进行聚集或分离.实验表明,该算法在多个学习框架下显著地改善了系统的整体性能.

10.期刊论文 董静.孙乐.冯元勇.黄瑞红.DONG Jing.SUN Le.FENG Yuan-yong.HUANG Rui-hong 中文实体关系抽取中的特征选择研究 -中文信息学报2007,21(4)

命名实体关系抽取是信息抽取研究领域中的重要研究课题之一.通过分析,本文提出将中文实体关系划分为:包含实体关系与非包含实体关系.针对同一种句法特征在识别它们时性能的明显差异,本文对这两种关系采用了不同的句法特征集,并提出了一些适合各自特点的新的句法特征.在CRF模型框架下,以ACE2007的语料作为实验数据,结果表明本文的划分方法和新特征有效的提高了汉语实体关系抽取任务的性能.