跨语言文档对齐

王洪俊[12]，施水才[1]，俞士汶[2]

（北京拓尔思信息技术有限公司,北京 100101；北京大学计算语言学研究所,北京 100080）

摘　要： 本文提出了一种新的双语文档对齐算法，该算法用 TfIDf 方法进行文本特征提取和权重计算，使用统计翻译模型进行双语词汇对齐，用 Dice 方法的改进算法计算双语文档的相似度。实验表明，该算法可以准确地发现一种语言书写的文档在另一种语言中的译稿，可应用于双语重稿检测、跨语言相似文本检索等领域。

关键词： 跨语言文档对齐，文档相似度。

Cross-Language Document Alignment

Wang Hongjun[12]， Shi Shuicai[1], Yu Shiwen[2]

(Beijing TRS Information Technology, Beijing 100101, China; Institute Of Computing Linguistics Peking University, Beijing 100080, China)

Abstract: We proposed a new algorithm for Cross-Language Document Alignment，which uses statistical translation model to align bilingual words-pairs, uses TfIDf method to refine text features and uses a new Dice-Method-based method to compute Cross-Language document similarity. This algorithm can be applied to find translation equivalent of a document in other languages.

key words: Cross-Language Document Alignment, Document similarity

1　概述

近年来，大型双语平行语料库（Parallel Documents）作为机器翻译和多语言信息处理的重要资源正在发挥越来越大的作用，被广泛应用于词汇知识获取、统计翻译模型、跨语言信息检索等领域。然而，对于大多数双语对（Language Pairs）来说，可以利用的双语资源范围非常狭窄，主要集中在政府公文、宗教文本等有限的领域。同时，与其他语言资源一样，双语语料库往往需要付费或授权

基金资助：北京市重大科技计划项目(H030130050610)；国家自然科学基金资助项目(60272084)；北京市教委重大项目(KZ200310772013)；

作者简介：王洪俊（1975-），男，山东人，在职博士生 email: wang.hongjun@trs.com.cn

后才能使用。

互联网上包含着大量的用各种语言书写的双语平行网页，这些网页由无数的作者创作，覆盖了各种不同的领域，并且这些网页会随着语言的变化和社会的发展不断进行更新。因此，互联网是一个免费双语语料的巨大宝藏。许多研究者尝试开发一种算法，从互联网上自动寻找双语平行网页。

其基本方法是，首先从互联网下载网页，对网页进行预处理，提取正文信息；然后计算双语文档之间的相似度，确定一篇文档是否是另一篇文档的翻译稿。（由于网页下载和预处理等都是比较成熟的技术，本文只讨论双语文档之间的相似度计算方法，该方法又称为双语文档对齐技术。）

Resnik[1][2]提出了一种根据网页的名字、结构和长度等信息来计算双语文档相似度的方法，在英法文档对齐任务上取得了比较好的结果；Nie[2]把这种方法应用于英汉网页对齐，结果并不理想，Nie 认为其中的一个原因是汉语网页的命名不规范。

Noah[3][4]提出了一种通用算法，在不同语言的文档库中检测双语对齐文档。该算法基于网页的重稿检测（Duplication Detection）技术，并使用了从双语语料中学习的知识。

Ralf Steinberger[5]等提出了一种语言无关的方法计算文档之间的语义相似度，这些文档可以是用同一种语言写成的，也可以用几种不同的语言写成。文档相似度的计算的前提是用一种语言无关的方式进行文档表示，使用多语言词典EUROVOC中的词表示文档，然后计算这些词语之间的距离。

Hasan[6][7]等提出了一种方法，使用语言学知识和统计方法对一个未对齐的汉日双语语料库进行文档对齐。双语文档的相似度计算使用互信息（Mutual Information,MI）和 RIDF 两种方法。

本文对 Noah 的方法进行了改进，提出了一种更加有效的双语文档对齐算法。该算法采用统计翻译模型（Statistical Translation Model）进行汉英词汇对齐，使用 TfIDf 方法进行文档特征提取和权重计算，在此基础上，使用 Dice 方法的改进算法进行双语文档相似度计算。汉英双语文档对齐的实验结果表明，该算法的精度是已报道的结果中最好的。

2 双语文档对齐

Noah 方法的基本原理是，两个双语文档之间的相似度是由二者共同拥有的翻译词对来决定的，拥有的翻译词对越多，两篇文档越相似，二者互为翻译对的可能性越大。

所谓翻译词对，就是这样两个词，二者属于不同的语言，之间有互译的可能，或者存在一个连接关系，在一定条件下， 其中一个词可以译成另一个词。

下图是一个汉英句对的互译关系图，直线连接的两个词就是翻译词对，包括：（暴风雨，storm），（后，after），（天气，sky），（晴朗，cleared）。（过，NULL）、（NULL，The）等。指向 NULL 的词是空的翻译词对，表示该词没有直接的翻译词，计算的时候我们不考虑这样的翻译对。

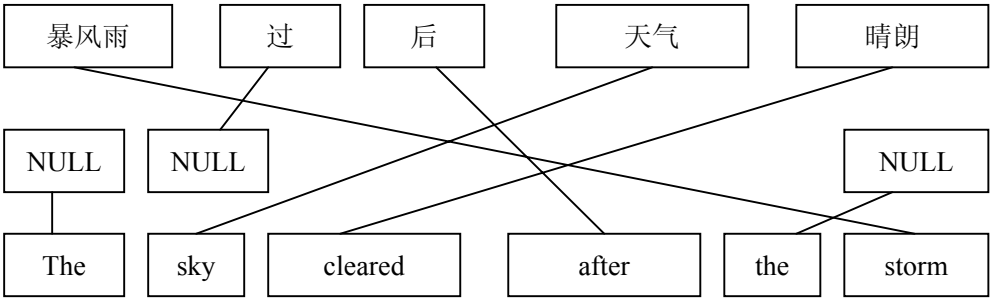

图 1

双语文档对齐需要解决两个主要问题：

（1） 如何查找翻译词对，即进行双语词汇对齐。

双语词汇的对齐一般采用基于双语词典的方法，该双语词典可以是人工编辑的翻译词典，也可

是统计翻译模型。

查找翻译词对的方法是：从源文中取一个词，在双语词典中找到一组可能的翻译词，然后把这组翻译词到目标文本中进行匹配，如果目标文档中有某个翻译词存在，则认为找到了一个翻译词对。由于词语的多义性，一个词往往有多个可能的翻译词，如何选择合适的翻译词，是这里需要考虑的问题。

（2） 双语文档相似度的计算。

找到两篇文档的所有翻译词对后，根据翻译词对的数量、文档的长度等信息计算文档的跨语言相似度，以找到最相似的文档，即该文档用另一种语言书写的译稿。Noah 在其文档对齐实验中只考虑了翻译词对的数量，我们认为，翻译词对在文档中的权重对于文档相似度计算也有很大的影响，并采用了 Tf 和 TfIdf 两种权重进行了实验，获得了良好的结果。

下面我们对这两个问题进行详细的阐述。

3 双语词汇的对齐方法

我们使用的双语知识库是统计翻译模型，格式为（汉语词，英语词，两词之间的翻译概率）。该统计翻译模型有如下特点：

（1）使用共享软件 Egypt[8]和 GIZA++[9]进行统计翻译模型训练，得到上述的翻译模型。

（2）训练前，对语料中的英文进行了词根处理和大小写转换，以缩小文本特征空间；对中文进行了分词，但没有进行未登录词处理。

（3）训练后，对得到的统计翻译模型进行过滤：把翻译概率值过低的翻译对去掉，把标题符号和部分虚词如连词、介词等的翻译对去掉。这样可以缩小参数空间，同时减少噪声信息的干扰。

（4）与一般双语词典相比，使用翻译模型只能进行词与词的对齐，不能进行词与短语和短语与短语之间的对齐。另外，生成一个翻译模型比构造一个双语词典更容易，耗费人工更少。

双语词汇对齐的具体算法是：

（1）对源文和目标文本进行中文分词和英文词根处理、大小写转换。

（2）从源文切分结果中取一个词，在翻译模型中找到该词的一组可能的翻译词；

（3）按翻译概率从高到低的顺序，到目标文本的切分结果中进行翻译词对匹配，如果目标文档中有某个翻译词存在，则认为找到了一个翻译词对。采用按翻译概率从高到低的顺序进行匹配，可以保证优先对齐翻译概率大的词对。

为了提高对齐的效果，我们做了以下限制：

（1）每个目标词只能对齐一次，对齐过的词不再与其他源词进行对齐。

（2）一个源词只能对应一个目标词。但一个源词如果在原文出现了 n 次，则可以与多个目标词进行对齐，这些目标词的词频总和不能超过 n。

（3）有些词如数字，在中文和英文中都是一样的形式，这些词可以直接作为翻译对。如源文中的 2000 和目标文本中的 2000 也算做一个翻译对。

4 双语文档相似度计算

在计算文档相似度前，先对文档应用向量空间模型进行文本表示，然后计算两个文档向量之间的相似度。常用的相似度计算方法有 Cosine 距离法、欧氏距离法(Euclidean)等。我们这里使用的是 Dice 方法及其改进算法，其优点是计算简单、结果归一化。

整个算法流程如下：

（1） 文档特征提取

对文档进行分词，包括中文分词、英文的词根处理和大小写转换。

在对文档进行分词后，需要对分词结果进行特征提取。这样做，一方面可以降低文档向量的维数，提高计算效率；另一方面，也可以减少噪音的干扰，因为文本中的一些与主题无关的特征被过滤掉了，留下的词语更能够反映文档的主题。

常用的特征提取方法有互信息（MI），交叉熵（CHI）、Tf、IDf 等。我们采用的特征提取算法是信息检索中常用的 TfIDf 权重计算方法，同时与 Tf 权重计算方法进行了比较。

方法是，计算文档中每个词语的 TfIDf 值，根据 TfIDf 值的大小衡量词语在文本中的重要程度，把前 N 个 TfIDf 值最大的词语作为文本的特征，把 TfIDf 值作为该词语的特征值。这里 N 取值是 100。

（2） 相似度计算方法

采用类似于 Dice 距离的方法进行计算。

设两篇文档 A 和 B 的特征词语的 Tf 权重个数分别为 a 和 b，两篇文本之间拥有的相同的文本特征权重总数为 c；公式为:

$$Sim(d1,d2)=\frac{2*c}{(a+b)} \tag{1}$$

为了提高该算法的性能，我们对该公式进行了如下改进：

设两篇文档 A 和 B 的特征词语的 TfIDf 权重之和分别为 a’和 b’，两篇文本之间拥有的相同的文本特征权重之和为 c’；公式为:

$$Sim'(d1,d2)=\frac{2*c'}{(a'+b')} \tag{2}$$

经过实验，发现公式（2）对于单语文档相似度计算效果很好，对于双语相似度计算则结果并不理想，我们又采用了如下改进公式：

$$Sim''(d1,d2)=\frac{2*c'}{(a+b)} \tag{3}$$

但是这个公式存在问题，没有进行归一化，由于 a’>>a，b’>>b，使得可能 2*c’>(a+b)，不能保证 Sim”(d1,d2)<=1。所以我们对公式（3）进行了归一化：

$$Sim'''(d1,d2)=\frac{2*c'}{(a+b)}*\frac{\min(a',b')}{a'} \tag{4}$$

后面的实验结果证明，引入了 TfIdf 权重之后，公式（4）还是相当有效的，其性能超过了公式（1）。

5 算法描述和结果分析

5.1 算法流程

我们提出的算法流程如下（见图 2）：

（1）首先对文档进行分词，包括中文分词、英文的词根处理等。

在对文本进行分词后，采用 TfIdf 方法对分词结果进行特征提取，提取前 N（100）个权值最大的词。

（2）将源语文档的特征词与目标文档的特征词进行对齐。

这里我们使用了一个统计翻译模型。对齐的时候按照特征词的 TfIDf 权值从高到低的顺序进行对齐。

（3）双语文档相似度计算

词汇对齐后，我们采用公式（4）进行双语文档相似度计算。

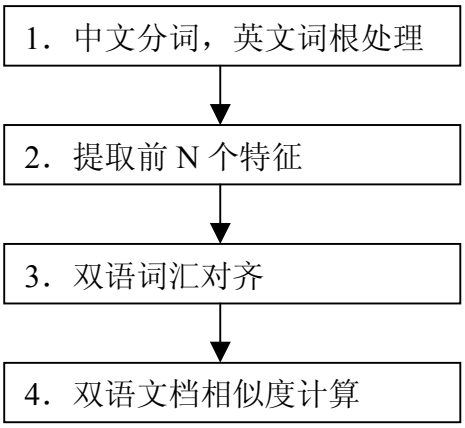

图 2

5．2 测试语料

我们采用以下语料作为测试语料库：

奥运基础数据库中的双语对齐文档共 500 篇，中英文各 250 篇，总共 14M 字节；新浪网英语频道（http://edu.sina.com.cn/en/sy.html）下载的双语对照文档共 120 篇，中英文各 60 篇，总共 252K。

20 Newsgroup 英文语料 19,997 篇，共计 43.9M 字节。这部分语料与上述语料不存在对齐关系，只是作为噪声语料参与测试，测试算法的抗干扰能力。

这些文档的长度从几百字到数十万字不等，共计 58.2M 字节。

5．3 测试方法和结果

将测试语料中的中文文档共 310 篇，做成批处理文件，到测试语料库中逐篇查找中文文档的对应英文文档，看相似度排名最高的文档是否就是其英文对照文档。

在实验中我们使用了两个不同的翻译模型，测试翻译模型的大小与对齐效果的关系：

（a）大的统计翻译模型：奥运基础数据库中的 20 万句汉英双语句对，中英文各 10 万句，共 25.9M。

（b）小的统计翻译模型：自己收集的大约 4.5 万句汉英双语句对，中英文各两万多句，共 1.62M。

另外，我们还分别测试了公式（1）和公式（4）的对齐性能；

没有排在第一位的对应英文文档的个数/正确率	公式（1）	公式（4）
小翻译模型	5 篇/98.4%	3 篇/99.0%
大翻译模型	18 篇/94.2%	1 篇/99.7%

测试结果表明：

（a）尽管使用了干扰语料库（近两万篇英文文档），四组测试结果的精度都达到了 90%以上，其中采用公式（4）的两个测试结果的精度都达到了 99%以上，表明该算法的性能完全满足双语文档对齐这个任务。这个测试结果是国内外已发表的结果中最好的。

（b）采用了 TfIDf 权重的公式（4）比采用了 Tf 权重的公式（1），性能更稳定，效果更好。

（c）该算法受统计翻译模型的影响不大，大的翻译模型未必就比小的翻译模型好。一个 4 万多句的双语句对集就足以获取很好的对齐效果。这个结论可以指导我们搜集双语句对，进行双语文档对齐实验。

（d）20 万句的句对集训练的翻译模型在公式（1）下表现很差，甚至不如 4 万句的句集。我们认为，大的统计翻译模型中包含了更多的噪声信息，在双语词汇对齐阶段，这些噪声信息影响了对齐效果。公式（4）中由于引入了 TfIDf 权重信息，抑制了噪声信息的不利影响。

6 存在的问题和未来的研究方向

本文提出了一种双语文本对齐算法，可以准确地发现一种文档用另一种语言写的译稿。该算法可以应用于双语文档对齐、跨语言相似文档检索等领域。

该算法有如下优点：核心算法与语种无关，对于某个语言对，只要提供一定数量的双语句对训练一个统计翻译模型，就可以方便地进行双语相似度计算。由于训练语料的限制，我们现在只对汉英文档对齐做了实验。前人的实验结果表明，汉英文档对齐任务相比欧洲语言的双语对齐任务更难一些。下一阶段，我们将应用该算法进行欧洲语言的双语对齐，进一步验证该算法的性能。

另外，虽然本算法在测试中已经取得了相当好的结果，但仍然存在一个缺点，无法找到一个重稿的相似度阈值。这个缺陷使得暂时无法直接根据阈值进行双语重稿判别。

下一阶段，我们打算采用下列方法解决这个问题：采用段落对齐算法、句对齐算法对排名靠前的相似文档进一步对齐，获取对齐的更有力证据。

致谢：

感谢中科院计算所提供的奥运基础数据库。

参考文献：

[1] Philip Resnik, "Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text." Third Conference of the Association for Machine Translation in the Americas (AMTA-98), Langhorne, PA, Lecture Notes in Artificial Intelligence 1529, Springer, October, 1998.

[2] Philip Resnik, "Mining the Web for Bilingual Text." 37th Annual Meeting of the Association for Computational Linguistics (ACL'99). College Park, Maryland, June 1999.

[3] Md. Maruf Hasan and Yuji Matsumoto. “Multilingual Document Alignment - A Study with Chinese and Japanese.” Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001), Tokyo, November 2001, pp.617-623.

[4] Md. Maruf Hasan. “Cross-language Information Retrieval, Document Alignment and Visualization -A Study with Japanese and Chinese.” PHD thesis(2001),Nara Institute of Science and Technology

[5] Ralf Steinberger, Bruno Pouliquen, Johan Hagman. “Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC.” CICLing 2002: 415-424

[6] Noah A. Smith. “Detection of Translational Equivalence.” Bachelor Thesis(2001),University of Maryland

[7] Noah A. Smith. "From Words to Corpora: Recognizing Translation." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Philadelphia, Pennsylvania.

[8] http://www.clsp.jhu.edu/ws99/projects/mt/

[9] http://www.isi.edu/~och/GIZA++.html

[10] Wessel Kraaij Jian-Yun Nie. "Embedding Web-based Statistical Translation Models in Cross-Language Information Retrieval." Computational Linguistics 29(3): 381-419 (2003)