文章编号：1003-0077（2011）00-0000-00

基于词典的句对齐方法在专利文本上的应用研究*

姜涛，蒋宏飞，任智军，张凯，熊文

（中国专利信息中心，北京 100088）

摘要：专利文本句对齐是专利研究与应用的基础任务，被广泛应用于专利自动翻译、跨语言检索等领域。本文以开源的句对齐工具 Champollion[5,6]为基础，重点研究基于词典的句对齐方法在专利文本上的应用。本文通过实验证明了使用专利领域词典可以有效的提高句对齐的正确率。本文还改进了 Champollion 的对齐算法，使其能够使用多词表达式（multi-word expression）来扩展其词典。在此基础上，本文还将上述方法应用于专利文本的统计机器翻译，并通过实验证明了以上改进的有效性。

关键词：句对齐；Champollion Toolkit；专利文本

中图分类号：TP391　　**文献标识码：**A

Research of Applying Dictionary-based Sentence Alignment Method on Patent Text

Tao Jiang, Hongfei Jiang, Zhijun Ren, Kai Zhang, Wen Xiong

(China Patent Information Center, Beijing 100088, China)

Abstract: Sentence alignment is a fundamental task in research and application of patent text. It is widely used in patent machine translation, cross-language information retrieval, etc. In the paper, an open source sentence alignment toolkit, named Champollion[5,6], was used for the research of applying dictionary-based sentence alignment method on patent text. The experiment shows that the alignment precision is significantly improved by adding patent domain dictionary to the system. An improved alignment algorithm was also proposed to enable Champollion to use multi-word expression in its dictionary. Based upon work, a series of machine translation experiments are performed and the results readily justify the validity of the work in the paper.

Key words: sentence alignment; Champollion Toolkit; patent text

1 引言

文本句对齐是自然语言处理领域的基础任务，被广泛的应用于编纂词典，自动翻译、跨语言检索等领域。目前主要的句对齐方法包括基于长度的方法、基于词典的方法和基于统计的方法。基于长度的方法[1,4]根据句子中词的个数来对双语句对进行对齐。由于方法简单，并且不需要词典或训练语料，因此通常作为多次对齐中的第一次对齐方法。基于统计的方法[3]一般使用某种统计模型自动在平行语料间建立起词一级的对齐关系，然后通过迭代算法来优化句对齐结果。和基于长度的对齐方法类似，基于统计的对齐方法也不需要词典或训练语料。虽然存在上述优点，上面两种方法对噪声数据的处理能力较弱，在处理含有较多噪声数据的文本（例如专利）时，性能会有明显下降。基于词典的句对齐方法需要双语词典作为对齐依据，该方法将两个句子间的对齐关系分解为词之间的对齐关系。由于使用了词典作为

* **收稿日期：**　　**定稿日期：**

基金项目：国家高技术研究发展计划(863) (2012AA011104)；国家教育部博士点专项基金(授权号: 2013M540125, 2013M530026)。

作者简介：姜涛（1980——），男，工程师，机器翻译；蒋宏飞（1982——），男，助理研究员，自然语言处理，机器翻译，专利信息处理；任智军（1977——），男，高级工程师，机器学习，机器翻译；张凯（1988——），女，工程师，语料库建设，计算机辅助翻译；熊文（1968——），男，助理研究员/博士后，机器翻译，数据挖掘，文本挖掘。

对齐依据，因此和前两种方法相比，基于词典的对齐方法对噪声数据具有更好的健壮性，因此更适合噪声数据较多的专利文本的对齐。除此之外，还存在多策略句对齐方法。[7,9]使用的多策略句对齐方法中，在第一遍句对齐中使用了基于长度的对齐，在第二遍对齐中使用了基于 IBM Model1[2]的统计方法。 [13]使用了组合方法（ensemble）和过滤(filtering)来提高句对齐质量。

Ma 实现了基于词典的开源句对齐工具 Champollion，由于其改进了互译词间的算分方法，因此 Champollion 对噪声数据的文本对齐具有更好的健壮性。 Champollion 在汉英语料（Sinorama Magazine, Hong Kong Handsard and UN official documents）上得到了 96%以上的正确率。但在应用于专利文本的句对齐问题时，实验证明，Champollion 的性能会有较大下降。Lu 在[13]论证了专利中的松弛翻译（loose translation）是句对齐算法在专利语料上性能下降的主要原因。除此之外，专利文本中术语繁多也是导致 Champollion 性能下降的原因。

本文以 Champollion 作为基础。首先研究了领域词典对句对齐质量的影响。然后扩展了 Champollion 的句对齐算法，使其能够在词典中使用多词表达式。最后，我们还研究了上述方法对统计机器翻译译文质量的影响。

2 Champollion 的特点

Champollion 被设计成用来对齐噪声数据较多的平行语料。Champollion 通过赋予不常见的词更多的权重来增加其算法的健壮性。与其他句对齐工具相比，Champollion 具有两个特点。首先，Champollion 假设噪声数据作为输入，即双语文本中存在大量非 1 对 1 的情况。其次，与其它的基于词典的方法不同，Champollion 对互译词赋予权重。Champollion 采用了信息检索领域用来计算文档相似度的 tf-idf 权重来计算句子片段间的相似度。然后使用动态规划算法搜索文档内部最优的句对齐路径。Champollion 允许 1-2,0-1,1-1,,2-1,1-2,1-3,3-1,1-4 和 4-1 对齐关系。

3 Champollion 的改进

基于词典的对齐方法的核心思想是将句子间的对齐关系分解为两个句子词之间的对齐关系，而词之间的对齐关系可以通过双语词典获得。因此双语词典对输入句中词汇的覆盖度直接决定了对齐算法的有效性。

Champollion 的对齐算法要求双语词典中的词条必须是单词形式（词条内部不可以存在空格），不可以是多词表达式。表 1 是 Champollion 使用词典的部分示例。

表 1　Champollion 词典示例

abandon <> 背弃	abduct <> 诱拐
abandon <> 丢弃	abide <> 坚持
abandon <> 放弃	abide <> 继续忍受
abandon <> 抛弃	abide <> 居住
abandon <> 弃	abide <> 居留
abandon <> 摈弃	abide <> 持久
abandon <> 遐	abide <> 蓄意等待
abandon <> 放纵	abide <> 忍受
abandon <> 遗弃	abide <> 等候
abc <> 初步	abide <> 守候
abc <> 入门	abide <> 遵守
abc <> 字母	abide <> 滞留
abc <> 照	abc <> 基本

在应用于专利领域的句对齐时，由于领域词典通常存在大量的多词表达式，去除词典中

的多词表达式会极大地影响词典资源的利用效率。本文将改进 Champollion 的对齐算法，使其在词典中能够使用多词表达式。改进后的 Champollion 将可以使用如表 2 所示形式的词典。

表 2 改进的 Champollion 词典示例

angle junction <> 直角 转 接 器
ground fault <> 接地 短路
multi - resolution algorithm <> 多 分辨 算法
radio remote unit and base band unit <> 基 带 拉 远 设备
two - way radio <> 双向 无线电 通信
bandpass sample <> 带 通信 号 采样 定理
unscented particle filter <> Unscented 粒子 滤波
statistic the frequency of the word <> 词 频 统计
surface noise <> 音 纹 噪声
pseudo - object - language <> 伪 对象 语言
service monitor broker <> 服务 监测 中介
crumb <> 屑 用
semantic similarity <> 语 义 相似 度
j2ee <> Java2 企业 版 平台
beer pump <> 啤酒 泵
public computer room <> 公共 计算 机房

Ma 在[5,6]中描述和实现了 Champollion 的对齐算法。使 Champollion 的词典可以使用多词表达式的关键在于需要修改其句对相似度计算方法。在导入词典时，在输入句的词列表中增加输入句子中所有 n-gram 形式的词串。(为了提高算法效率，n 将被限制小于一定的阈值 *thr*。)例如对于下面的输入：

a. Marketplace bombing kills 23 in Iraq
b. 伊拉克 集市 爆炸 造成 23 人 死亡

在句对齐算法中，当 *thr*=3 时，要考虑表 3 中中英文相应的 n-gram 间的对齐权重。

表 3 输入句的 n-gram 列表

English n-grams	*Chinese n-grams*
unigram	*unigram*
Marketplace	伊拉克
bombing	集市
kills	爆炸
23	造成
in	23
Iraq	人
	死亡
bigram	
Marketplace bombing	*bigram*
bombing kills	伊拉克 集市
kills 23	集市 爆炸
23 in	爆炸 造成

in Iraq *trigram* Marketplace bombing kills bombing kills 23 kills 23 in 23 in Iraq	造成 23 23 人 人 死亡 *trigram* 伊拉克 集市 爆炸 集市 爆炸 造成 爆炸 造成 23 造成 23 人 23 人 死亡

词串间的对齐分数的计算方法（tf-idf score）与 Champollion 的计算方法相同。表 4 是修改后的 Champollion 对齐算法的伪代码。（Champllion 原有的对齐算法实现可以参考 champollion-1.2[6]目录下的 bin/champollion_kernel 中 Match_sentences_lex 函数。）

表 4 改进的 Champollion 句子对齐打分算法

```
function   match_sentences_lex_with_mwe（s1, s2）{                                    (1)
    for n in [1..thr] {                                                               (2)
        add all n-gram in s1 to xtokens;                                              (3)
    }

    for n in [1..thr] {                                                               (4)
        add all n-gram in s2 to ytokens;                                              (5)
    }

    for xtoken in xtokens {                                                           (6)
        if xtoken in s2 and xtoken is not a stop-word {                               (7)
            minpairs = min(number of xtoken in xtokens, number of xtoken in ytokens); (8)
            score += tf-idf score * minpairs;                                         (9)
        }
        else   {
            for xtoken_trans in all-xtoken-translation {                              (10)
                minpairs=min(number of xtoken in xtokens, number of ytoken in ytokens); (11)
                score += tf-idf score * minpairs;                                     (12)
                update xtokens and ytokens;                                           (13)
            }
        }
    }
    Return score                                                                      (14)
}
```

步骤 1：函数输入句子 s1 和 s2；
步骤 2-3：将输入 s1 中的所有长度小于某一阈值 *thr* 的 n-gram 加入词表 xtokens；
步骤 4-5：将输入 s2 中的所有长度小于某一阈值 *thr* 的 n-gram 加入词表 ytokens；
步骤 6：对 xtokens 中每一个词执行 6-10 的操作；

步骤 7-9：对输入 s1 和 s2 中相同（不包括停用词）的 n-gram 对进行打分，分数的计算公式与[5]中描述的相同；
步骤 10-12：对输入 s1 和 s2 中不相同的 n-gram 进行基于词典的对齐，并对其打分，分数的计算公式与[5,6]中描述的相同。
步骤 13：更新 xtokens 与 ytokens，这一部分与[6]中的实现方法相同。
步骤 14：返回 s1 和 s2 的对齐分数。

在得到 s1 和 s2 的对齐分数后，Champollion 会在整篇文档上使用动态规划算法计算出句对齐的最优路径。这一部分与[5,6]中的实现方法相同，因此在上面的伪代码中没有给出这一部分。

4 实验

实验使用的语料来自中国专利信息中心的专利语料库。测试语料共包含 500 篇中英对齐专利摘要。英文摘要的分句使用了标点作为切分标记。中文摘要的分句除了使用了标点作为切分标记外，还使用了一些简单的规则（例如状语“在…之后，”不可以分句）。测试语料的句对齐答案集先由 Champollion 自动生成，再经过人工校对。表 5 是测试语料的具体信息。

表 5 测试语料统计信息

领域	数量	句子数（中文/英文）
IPC E 部	500 篇摘要	12572/14536

我们先测试了 Champollion（基线系统）在上述 500 篇摘要上的句对齐性能。实验结果如表 6 所示。

表 6 Champollion 句对齐实验结果

正确率	召回率	F 值
0.772	0.851	0.810

从实验结果可以看出，Champollion 在专利语料上的性能明显低于 Ma 在[5]中在常规语料上得到的结果。

如前所述，Champollion 使用了基于词典的对齐方法。因此，我们预期专利领域术语繁多是造成句对齐性能下降的重要原因。在下面的实验中我们通过向 Champollion 不断增加专利术语词典来测试其句对齐性能。在此实验中我们使用了未作改进的 Champollion，因此我们滤掉了词典中的多词表达式。测试中使用的专利词典共 200000 词条。

表 7 Champollion 增加领域词典的句对齐结果

测试集	#单词	正确率	召回率	F 值
20K	480	0.772	0.851	0.810
40K	952	0.773	0.853	0.811
60K	1432	0.773	0.854	0.812
80K	1919	0.778	0.858	0.816
100K	2398	0.778	0.859	0.817
120K	2886	0.778	0.859	0.817
140K	3393	0.786	0.864	0.823
160K	3872	0.787	0.865	0.824
180K	4367	0.788	0.866	0.825
200K	4857	**0.788**	**0.865**	**0.825**

表 7 中的第一列（测试集）的数字表示选取的词条总数（包括单词和多词表达式）。第二列是单词词条的数量，也是实验中有效的部分。例如第一组实验中共有 20K 的词条，其中有 480 个单词词条。从上面的数字可以看出，单词形式的词条大约占词条总数的 2.4%。和上面的基线系统相比，正确率提高了 1.6 个百分点。实验结果证实了预期的假设，即在使用基于词典的对齐过程中，增加领域词典可以有效提高专利文本句对齐的性能。

从上面的实验结果可以看出，在典型的专利领域词典中，单词所占的比例仅为 2.4%，这意味着词典中 97.6%的部分无法被 Champollion 使用。在下面的实验中，我们尝试使用术语词典中所有的词条（包括单词和多词表达式）。这一部分需要本文第三节中提出的改进的句对齐算法。表 8 是实验结果。

表 8 改进的 Champollion 句对齐实验结果

测试集	正确率	召回率	F 值
20K	0.774	0.853	0.811
40K	0.775	0.855	0.813
60K	0.777	0.857	0.815
80K	0.781	0.861	0.819
100K	0.782	0.862	0.820
120K	0.782	0.863	0.821
140K	0.789	0.868	0.827
160K	0.792	0.870	0.829
180K	0.793	0.871	0.830
200K	**0.793**	**0.872**	**0.831**

从实验结果可以看出，在加入所有 200K 词条后，句对齐的正确率和基线系统相比提高了 2.1 个百分点；和加入了相同的术语词典，但未作修改的 Champollion 系统相比，正确率提高了 0.5 个百分点。

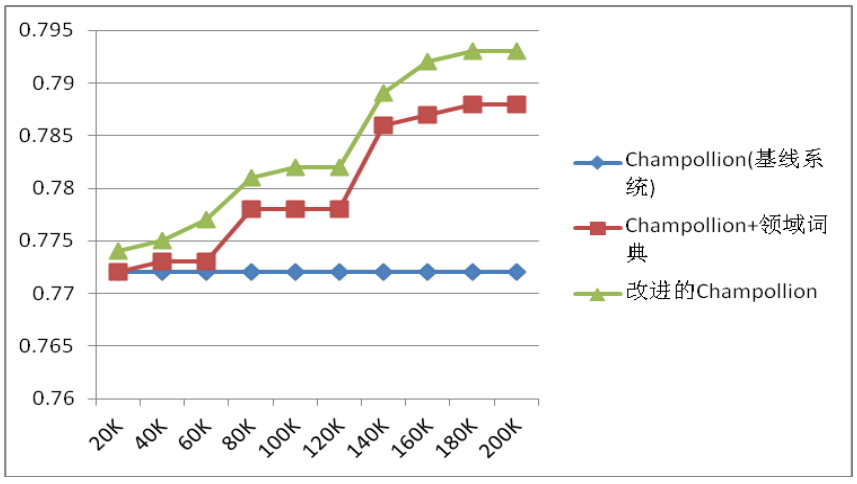

图 1 句对齐性能比较

在应用于统计机器翻译任务时，语料库的规模通常在数百万句对以上，正确率提高 0.5 个百分点意味着将增加数万个正确的句对。因此本文第三节提出的改进算法在实际应用中具有重要的价值。

我们还通过实验研究了本文中提出的方法在统计机器翻译上的应用。在实验中，我们使用了中国专利信息中心专利语料库中 47000 余篇专利摘要作为训练语料。语料的具体信息如表 9 所示。

表 9 统计机器翻译训练语料

领域	数量	句子数（中文/英文）
IPC E 部	47562 摘要	1170188/1313684

训练语料经过前述三个系统生成的句对齐语料信息如表10所示。其中Baseline表示使用基线系统得到的句对齐结果；+Dict.表示使用Champollion并增加专利领域词典（仅单词有效）得到的句对齐结果；Ext.表示使用改进对齐算法的Champollion并增加专利领域词典得到的句对齐结果。

表 10 句对齐语料信息

	Baseline	+Dict.	Ext.
句对数	486354	517433	487797

上面三组句对齐结果分别用来训练统计翻译模型，全部的1313684英文句子用来训练语言模型。实验中我们使用了统计机器翻译的开源工具链：GIZA++[10]用来训练词对齐模型，Moses toolkit[12]用来训练翻译模型，IRSTLM[11]用来训练语言模型。同时我们还使用了MERT[15]来优化模型参数，开发集包含1000个专利句对。表11是上述三个系统在开发集上的BLEU[14]结果。

表 11 开发集上的统计翻译结果

	Baseline	+Dict.	Ext.
未优化	25.06	25.17	25.12
MERT	27.87	28.09	28.31

测试集来自于中国专利信息中心专利语料库 IPC E 部的 1000 个句子。表 12 是上面三个系统在测试集上的 BLEU 结果。

表 12 测试集上的统计翻译结果

	Baseline	+Dict.	Ext.
BLEU	23.24	**23.39**	**23.53**

从表 12 的实验结果可以看出，和基线系统相比，本文提出的句对齐改进方法均提高了实验结果的 BLEU 值，从而证明了在统计机器翻译应用中的有效性。

5 总结

专利文本句对齐是专利研究与应用的基础任务。本文以开源的句对齐工具 Champollion 为基础，利用中国专利信息中心的专利语料库，研究了基于词典的句对齐方法在专利文本上的应用。通过实验证明了使用领域词典可以有效的提高句对齐的正确率。同时，本文还改进了 Champollion 的对齐算法，使其能够在词典中使用多词表达式。从本文给出的实验结果可以看出，在典型的专利领域词典中，多词表达式所占的比例在 90%以上。因此本文提出的改进算法对提高词典资源的利用有重要的意义。本文还研究了上述改进对统计机器翻译性能的影响。实验数据证明了改进的句对齐算法可以显著的提高平行语料库的质量，从而提高统计机器翻译的译文质量。

参考文献

[1] Peter F. Brown, Jennifer C. Lai and Robert L. Mercer. Aligning Sentences in Parallel Corpora[C]. Proceedings of ACL. 1991:pp.169-176.

[2] Peter F. Brown, Stephen A. Della Pietra, Vincent J.Della Pietra and Robert L. Mercer. Mathematics of Statistical Machine Translation: Parameter Estimation[J]. Computational Linguistics, 1993, 19(2):263-311.

[3] Stanley F. Chen. Aligning Sentences in Bilingual Corpora Using Lexical Information[C]. Proceedings of ACL. Columbus, OH. 1993:pp. 9-16..

[4] William A. Gale and Kenneth W. Church. A Program for Aligning Sentences in Bilingual Corpora[C]. Proceedings of ACL. 1991:pp.79-85.

[5] Xiaoyi Ma. Champollion: A Robust Parallel Text Sentence Aligner[C]. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). Genova, Italy. 2006.

[6] Xiaoyi Ma. Champollion Toolkit[OL]. http://champollion.sourceforge.net/. 2006.

[7] Robert C. Moore. Fast and Accurate Sentence Alignment of Bilingual Corpora[C]. Proceedings of AMTA. 2002:pp.135-144.

[8] Franz Joseph Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models[J]. Computational Linguistics, 2003, 29(1):19-51.

[9] Michel Simard and Pierre Plamondon. Bilingual Sentence Alignment: Balancing Robustness and Accuracy[J]. Machine Translation, 1998, 13(1):59-80.

[10] Franz Josef Och, Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models"[J]. Computational Linguistics, 2003, 29(1):pp.19-51.

[11] M. Federico, N. Bertoldi, M. Cettolo, IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models[C]. Proceedings of Interspeech, Brisbane, Australia, 2008.

[12] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation[C]. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, 2007.

[13] Bin Lu, Benjamin K. Tsou, Jingbo Zhu, Tao Jiang and Oi Yee Kwong. The Construction of A Chinese-English Patent Parallel Corpus[C]. MT Summit XII 3rd Workshop on Patent Translation, Ottawa, Canada. 2009.

[14] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. BLEU: a method for automatic evaluation of machine translation[C]. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. 2002:pp.311–318.

[15] F. J. Och. Minimum Error Rate Training in Statistical Machine Translation[C]. In 41st Annual Meeting of the Association for Computational Linguistics (ACL). Sapporo, Japan. 2003:pp.160–167.

作者联系方式:姜涛 北京海淀区北太平庄路 25 号,中国专利信息中心 100088 010-82092340
jiangtao_1@cnpat.com.cn