语料对齐工具的性能比较与选择

蔡辉 中央财经大学

摘要：本文利用实验研究的方法，以文学、财经和科技三种文体为样本，对6款常见的语料对齐工具进行了比较研究。研究发现：（1）除Déjà Vu X3之外，相同文本使用docx和txt格式对对齐结果没有影响；（2）Transmate、ABBYY Aligner 2.0和memoQ 2015的对齐准确率位居前列，表现稳定；（3）使用不同体裁的文本，对齐质量也会不同。科技文本的对齐效果最佳，其次是财经和文学；（4）对齐准确率是评测对齐质量的主要指标，但不是唯一指标；（5）距离完美对齐的距离、句段长短、标签数量也影响对齐质量。本文还提出了对齐准确率的概念和计算公式。本研究对对齐工具的选择和改进具有一定参考作用。

关键词：语料；对齐；对齐准确率

中图分类号：H059 **文献标识码：**A **文章编号：**1000-873X（2019）03-0150-06

对齐既可表示寻找不同语言文本之间互译片断的过程（align），也可用于表示该过程产生的结果（alignment）。根据互译片段的长短或单位，可以分为词语对齐、短语对齐、句子对齐、段落对齐等。为研究方便，本文仅研究句子单位的对齐。通过对齐既可以生成双语平行语料库，也可以生成翻译记忆库。两者在自然语言处理的许多领域都具有较高的研究和实用价值，在机器翻译、词典编纂、信息检索、词义排歧和辅助翻译等方面等有较大的应用价值。

在现实生活中，很多文本信息都存在双语和多语版本，如果将这些语料对齐，将能产生巨大的经济社会效益。但人工对齐显然不现实。许多学者在对齐领域开展相关研究。在对齐算法方面，Brown et al.（1991）和Gale & Church（1991）提出了基于长度的方法（length-based）；Kay & Roscheisen（1993）提出了基于词汇的方法，Tan & Nagao（1995）和Wu（1994）则主张混合法。俞劲松等（2015）提出了“提出基于单词间粘合度与松弛度的语块划分评分方法以及双语语块划分的双向约束算法”。但这些研究并没有分析比较各款工具的性能和参数。也有一部分学者研究了不同长度单位的对齐，例如王斌等（2010）提出了借助锚点词所在句子的匹配获得锚点句子对来进行段落对齐的方法。Ker（1997）提出了根据语义类实现词对齐的方法。陈钰枫等（2011）提出了一种汉英实体名称的对齐模型。这些研究主要关注对齐的算法和途径，而对于对齐工具的对其质量缺乏比较和研究。无法给对齐工具的用户提供参考意见。

实际上，目前市场上对齐工具林林总总，十分繁杂，例如Abbyy Aligner、Tmxmall等，许多计算机辅助翻译工具也内置有对齐模块，例如SDL Trados、Déjà Vu、memoQ、Transmate等。各款工具都能实现语料对齐的功能，但表现各有不同。可惜，迄今为止，还没有人对这些对齐工具做系统分析。各款工具有什么优势和特点？学界对此尚未进行横向比较。对齐质量如何？学界也缺乏评价标准。正因为如此，译者在面临不同题材、不同格式的文本对齐任务时，在对齐工具选择方面缺乏明确的参考标准，在一定程度上也影响了对齐的质量和效率。有鉴于此，加之句对

齐生成的双语语料库的对机器翻译、计算机辅助翻译均有较高的价值，因此，本文选择了六款常见的对齐工具，对其句对齐功能展开比较分析，具体而言，将重点探索以下几个问题：（1）各种工具的对齐准确率和质量有什么不同？（2）不同格式的语料是否对对齐有什么影响？（3）不同题材的文档是否会影响对齐的准确率？（4）如果译文质量合格，如何评价各款对齐工具的对齐质量？本研究将有利于用户在执行句对齐任务时选择合适的工具，有助于对齐工具的进一步改进和优化。

一、研究方法

本文的研究对象为六款常见对齐工具，见表1。

表1 六款对齐工具

工具名称	原产国	是否收费
Abbyy Aligner 2.0	俄罗斯	是
Déjà Vu X3	法国	是
memoQ 2015	匈牙利	是
SDL Trados 2017	英国	是
Tmxmall[①]	中国	否
Transmate 7.3	中国	否

在表1所列六款工具中，ABBYY Aligner是一款由ABBYY集团开发的专业对齐工具，目前最高版本为2.0。ABBYY Aligner通过使用词典库，不仅可以将切分句段按句序进行匹配，而且可检查原文和译文的语法相似度，从而准确识别匹配句段，提高文本对齐质量[②]。Déjà Vu是一款常见的CAT工具，目前最高版本为X3。Déjà Vu内置有alignment模块，可以实现语料对齐功能。memoQ是一款常见的CAT工具，内置有livedocs模块，可以实现语料对齐功能。SDL Trados是著名的CAT工具，目前最高版本为2019，其内置有WinAlign模块，可以实现语料对齐功能。Tmxmall是是由上海一者科技有限公司研发的产品，其主营业务是语料商城和语料共享，它较早在国内推出了网页版在线对齐。Transmate是由成都优译信息技术有限公司研发的单机版工具，免费使用，其内置有双语对齐的模块，可实现语料对齐功能。市面上还有很多类似的对齐工具，例如雪人CAT、Bilingual Sentence Aligner、LF Aligner、Corpus Sort、Wordfisher等等，由于以上六款软件较为常用，加之笔者购置条件限制，因此本研究仅选择了以上六款工具作比较研究。

本研究选取文学、财经、科技三种文体，原文为英文，译文为中文，均为笔者翻译。篇幅长度在1000汉字左右，样本为纯文字内容，排版规范，不含图表。每个样本都分别保存为docx和txt两种格式。各样本的其它文本特征见表2。

具体研究步骤如图1所示。首先，将每对样本按照先docx文档后txt文档的顺序进行对齐，对齐过程按照软件的默认流程和默认设置，弹出对齐界面后，不进行人工干预，保存对齐结果。然后，对对齐结果进行统计分析。分析分两步：第一步先进行工具内比较（即将同一款软件同一样本的docx和txt两种文本格式的对齐结果进行比较）；第二步，进行工具间比较（即将同一样本使用不同软件的对齐结果进行比较）。

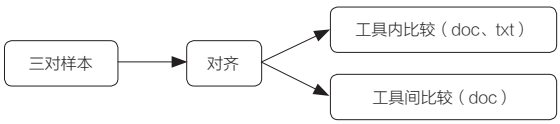

图1 研究流程图

二、结果分析

（一）工具内比较：word格式与记事本格式

随机抽取一种文体的文本，将其docx格式和txt格式分别导入上述六款对齐工具中，并记录对齐结果。实验结果发现：ABBYY、memoQ、Tmxmall、Trados、Transmate对于docx和txt

两种格式的相同文本进行对齐，对句段切分、准确率、格式标记、段首标记、中文符号识别不会产生差异。但是在 Déjà Vu X3 中，使用不同格式的文本对对齐的句段切分、对齐准确率、格式标记、文本识别都有所影响。

如表 3 所示，在 Déjà Vu X3 对齐中，使用 word 对齐率略高，没有段首标记，中文符号识别没有乱码，但是格式标签较多。使用 txt 文档对齐率略低，会保留段首标记、中文符号会有乱码，但是没有格式标记。

（二）工具间比较

第一，基本技术指标。对六款工具所支持的语种、格式以及导出格式等基本技术指标进行横向对比，见表 4。

如表 4 所示，在支持语种的数量上，Déjà Vu、memoQ 和 SDL Trados 支持的语种数量均超过了 100 个，具有明显的优势。从所支持的文本格式的种类数量来看，ABBYY、Déjà Vu、memoQ 和 SDL Trados 也处于领先水平。从导出格式的种类来看，Déjà Vu、memoQ 和 SDL Trados 均不能直接导出为 tmx 格式③。只有 ABBYY、Transmate 和 Tmxmall 支持导出为 tmx 格式，具有更好的兼容性。

第二，断句准确率。将三种文体的文本的 docx 格式，按照研究步骤逐次导入上述六款工具中，并记录下原文断句结果，并计算断句准确率，见表 5。

断句（segmentation）是对齐的基础。原文断句准确率越高，对齐准确率越高。从表 4 可见，在文学体裁中，Transmate 和 SDL Trados 断句准确率表现突出，分别为 98% 和 94%；在财经体裁中，Transmate 完全正确，表现最佳，ABBYY 和 Déjà Vu 并列第二，准确率为 96%；在科技体裁中，ABBYY、Transmate 和 SDL Trados 均 100% 断句正确。由是观之，在断句准确率方面，国产软件 transmate 表现最为突出，在三种不同题材中，均取得了最佳成绩。从六款工具在三种体裁中的平均值来看，断句准确率从高至低依次为科技、财经和文学，分别为 88.3%、82% 和 76.2%。

第三，对齐准确率。将三个文本的 docx 格式，导入上述六款工具之后，按照默认设置对齐，并记录对齐准确率，实验结果见表 6、7、8。

从表 6 可以看出，在文学体裁的对齐试验中，表现最好的三款工具是 memoQ、ABBYY 和 TMXmall，分别对齐了 28 句、20 句和 11 句。其中 TMXmall 只实现了段落对齐④。

如表 7 所示，在财经体裁的对齐试验中，表现最好的前两款工具是 transmate、ABBYY，分别对齐了 25 句、21 句；memoQ 和 SDL Trados 并列第三，均对齐了 9 句。值得一提的是，国产软件 Transmate 实现了完美对齐，不仅句段切分合理，对齐准确率也是 100%。TMXmall 虽然对齐准确率是 100%，但同样只实现了段落对齐。

如表 8 所示，在科技体裁的对齐试验中，表现最好的前三款工具是 transmate、ABBYY 和 Déjà Vu X3，分别对齐了 37 句、37 句和 19 句。值得一提的是，Transmate 和 ABBYY 两款工具均实现了完美对齐，不仅句段切分合理，对齐准确率也是百分之一百。TMXmall 在 21 个段落的对齐中，仅对准 1 段。

结合上述实验结果，为便于比较，笔者提出对齐准确率的概念。所谓对齐准确率，是指对齐句段数量与原文句段数⑤之比。对齐准确率是衡量对齐质量的重要指标。由于精准匹配的原文句段和译文句段的数量总是一致，但由于句段切分的规则和算法不同，原文句段和译文句段在数量上常常有出入。有鉴于此，为综合考虑句段切分因素，本文提出以下对齐准确率

的计算公式：

$$对齐准确率 = \frac{对齐句段数量}{原文句段数} \times \%$$

根据以上公式，六款工具的对齐准确率计算如下：

如表 9 所示，在文学体裁实验中，memoQ 和 ABBYY 表现最佳，对齐准确率分别为 57.1%、40.8%；在财经体裁实验中，transmate 和 ABBYY 表现最佳，对齐准确率分别为 100%、84%；在科技体裁实验中，transmate 和 ABBYY 表现最佳，对齐准确率均为 100%。从三种体裁平均对齐准确率来看，ABBYY 和 Transmate 表现最优，分别为 75% 和 73.3%。国产软件 transmate 财经和科技体裁实验中，均获得满分，可惜在文学体裁的对齐中差强人意。ABBYY Aligner 在三种题材中均表现突出、稳定。memoQ 在三种体裁的对齐中也表现突出。从六款软件在三种体裁的对齐平均值来看，平均值从高至低依次是科技、财经和文学，分别为 56.3%、48.7% 和 28.2%。这表明，这六款工具对科技体裁的对齐效果最佳，其次是财经和文学。这也表明，使用不同题材的文本，对齐效果也会不同。

第四，格式标签（tag）。格式标签用于表示文字特征（例如字体）、或文字流动特征（例如分页符），它可以分为：行内或结构标签、独立标签或标签对中的一个、可译或不可译标签等[⑥]。对齐产生的标签将会带入到记忆库，如不清除，不仅会影响句段的匹配，也会影响语言资产的重复利用。由于标签对记忆库的重复利用会产生消极影响，有无标签以及标签的多寡将会直接影响到记忆库的质量。因此，对齐标签的有无多寡，也是衡量对齐质量的一个重要指标。

从表 10 可见，在文学体裁对齐中，Déjà Vu 和 memoQ 均产生了大量标签，前者英中文中分别出现了是 15 和 25 个标签，后者分别产生了 6 和 33 个标签。其他工具均没有出现标签。在财经体裁对齐中，Déjà Vu 和 memoQ 均产生了标签，前者英汉文中分别出现了 34 和 53 个标签，后者英汉文中各产生了 1 个标签。其他工具均没有出现标签。在科技体裁对齐中，Déjà Vu、memoQ 和 Trados 均产生了标签，前者英汉文中分别出现了是 13 和 119 个标签，memoQ 分别产生了 9 和 11 个标签，Trados 只在英语文本中产生了 6 个标签。其他工具均没有出现标签。由是观之，从标签有无来看，ABBYY、Tmxmall 和 Transmate 表现最佳，均没有产生标签。

第五，纠错能力。ABBYY、memoQ、Transmate 具有纠错功能，当某一句段对齐紊乱之后，其后也能发现对齐的句段。其它三款工具则不具备这种功能，即当某一句段对齐紊乱之后，其后的句段会全部紊乱。

三、结论

本文通过对三种文体（文学、财经、科技）、两种格式（docx、txt）的样本实验，比较了 ABBYY Aligner2.0、Déjà Vu X3、memoQ 2015、SDL Trados 2017、Tmxmall 和 Transmate 等六款工具的对齐功能，结论如下：

（1）从支持语种、格式等基础技术指标来看，Déjà Vu、memoQ 和 SDL Trados 占据优势；从导出格式来看，只有 ABBYY、Transmate 和 Tmxmall 可以直接导出为 tmx 格式，具有更好的兼容性。

（2）在断句准确率方面，国产软件 transmate 表现最为突出，在三种不同题材中，均取得了最佳成绩。

（3）从对齐准确率来看，国产软件 Transmate 财经和科技体裁实验中，均获得满分，可惜在文学体裁的对齐中差强人意。

表2 样本特征

特征 / 文体	原文来源	总字数	文内标题	段落	句数[7]	首行缩进
文学	Excerpt from *In a Class by Himself*[8]	英文 570 词 中文 1002 字	3	7	49	中文有 英文无
财经	Excerpt from ASEAN, PRC, India: The Great Transformation[9]	英文 487 词 中文 874 字	2	5	25	中文有 英文无
科技类	Excerpt from Pre-Chamber of Internal Combustion Engine[10]	英文 675 词 中文 983 字	7	14	37	中文有 英文无

表3 Déjà Vu X3的对齐表现

项目 / 格式	句段切分		准确率		格式标记		段首标记		中文符号识别	
	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN
Docx	26	27	2	2	34 个	53 个	无	无	无	无
Txt	25	26	1	1	无	无	有 iiii	空格	有乱码	无

表4 基本技术指标对比

工具 / 项目	ABBYY	Déjà Vu	memoQ	SDL Trados	Tmxmall	Transmate
支持语种	24	>100	>100	246	18	13
支持格式	21	39	40	20[11]	13	3
导出格式	TMX, RTF	dvmdb	mqxlz	sdltm、SDLXLiff、SDLalign	tmx	tmx、uetm

表5 原文断句准确率

	ABBYY	Déjà Vu	memoQ	SDL Trados	Tmxmall	Transmate	平均值
文学	83.7% (41/49)	83.7% (57/49)[12]	75.6% (37/49)	94% (52/49)	22.4% (11/49)	98% 50/49	76.2%
财经	96% (24/25)	96% (26/25)	80% (20/25)	92% (23/25)	28% (7/25)	100% (25/25)	82%
科技	100% (37/37)	86.5% 42/37	86.5% (32/37)	100% (37/37)	56.8% (21/37)	100% (37/37)	88.3%
平均	93.2%	88.7%[13]	80.7%	95.3%	35.7%	99.3%	82.2%
排名	3	4	5	2	6	1	-

表6 文学文本对齐结果

软件 / 项目	ABBYY		Déjà Vu		memoQ		SDL Trados		Tmxmall		Transmate	
	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN
句段数	41	44	57	39	37	37	52	38	11	11	50	45
对齐句段数量	20	20	8	8	28	28	6	6	11	11	10	10

表7 财经文本对齐结果

软件 / 项目	ABBYY		Déjà Vu		memoQ		SDL Trados		Tmxmall		Transmate	
	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN
句段数	24	24	26	27	20	20	23	21	7	7	25	25
对齐句段数量	21	21	2	2	9	9	9	9	7	7	25	25

表8 科技文本对齐结果

软件 / 项目	ABBYY		Déjà Vu		memoQ		SDL Trados		Tmxmall		Transmate	
	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN
句段数	37	37	42	37	32	31	37	37	21	21	37	37
对齐句段数量	37	37	19	19	18	18	13	13	1	1	37	37

表9 六款工具的对齐准确率

	ABBYY	Déjà Vu	memoQ	SDL Trados	Tmxmall	Transmate	平均值
文学	40.8% (20/49)[14]	16.3% (8/49)	57.1% (28/49)	12.2% (6/49)	22.4% 11/49	20.4% 10/49	28.2%
财经	84% (21/25)	8% (2/25)	36% (9/25)	36% (9/25)	28% (7/25)	100% (25/25)	48.7%
科技	100% (37/37)	51.4% (19/37)	48.6% 18/37	35.1% 13/37	2.7% 1/37	100% (37/37)	56.3%
平均对齐准确率	75%	25.3%	47.3%	27.7%	17.7%	73.3%	44.4%
排名	1	5	3	4	6	2	-

表10 对齐产生的标签数量统计

软件 / 体裁	ABBYY		Déjà Vu		memoQ		SDL Trados		Tmxmall		Transmate	
	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN
文学	0	0	15	45	6	33	0	0	0	0	0	0
财经	0	0	34	53	1	1	0	0	0	0	0	0
科技	0	0	13	119	9	10	6	0	0	0	0	0
总计	0		279		60		6		0		0	

ABBYY Aligner 在三种题材中均表现突出、稳定。memoQ 在三种体裁的对齐中也表现突出。

（4）使用不同体裁的文本，对齐质量也会不同。科技文本的对齐效果最佳，其次是财经体裁，文学体裁的对齐效果最差。

（5）从标签有无来看，ABBYY、Tmxmal 和 Transmate 表现最佳，均没有产生格式标签。SDL Trados 有少量标签。Déjà vu 和 memoQ 标签较多。格式标签和对齐质量之间存在负相关关系，格式标签越多，对齐质量越差。

（6）在译文质量合格的前提下，评价一款对齐工具的对齐质量主要看其对齐准确率。对齐准确率和对齐质量之间存在正相关关系，即对齐准确率越高，对齐质量越高。但对齐准确率不是评测对齐质量的唯一指标，距离完美对齐的距离、有无标签，句段长短等因素也是评

四、研究的局限性和建议

本研究的局限有四个方面。一是文体的局限，仅作了文学、财经和科技类三种文本的比较，且每一文体只选取了一篇短文；二是文本格式的局限：仅 word 和 txt 两种文本格式的对齐实验；三是语言对的局限，本研究仅限中英文语对的实验。四是文本长度的局限，原文仅在千字以内；五是测试工具的局限性，由于条件限制，有些工具没有获取，有的工具不是最新版本。以上局限均对测试的准确性、有效性产生一定影响。研究者可以从其它文体、其它文本格式、其他语言对文件展开更深入研究，亦可以测试更多、更长、更复杂的文本，以及采用更新的软件版本或其它对齐软件，还可以从人工干预程度、操作便捷性等方面进行更加深入的研究。

注释

① 2015 年 11 月 18 日，TMXMall 发布了在线版对齐工具，2016 年 7 月 31 日，发布单机版对齐工具 Tmxmall Aligner。由于后者是付费工具，本研究采用的是在线版。

② 见 https://abbyy-ls.com/about。

③ Tmx（translation memory exchange）是记忆库的标准格式，它可以便捷地导入到各种 CAT 工具的记忆库中。Déjà Vu、memoQ 和 SDL Trados 在对齐后，需要经过更多的操作，才能将对齐文件导出为 tmx 格式。

④ Tmxmall 生成段落对齐，这降低了对齐的难度，但同时也降低了对齐的质量，因为段落对齐的复用率很低，而从段落对齐生成句对齐的记忆库，还需要大量人工干预。

⑤ 此处的原文句段数是指由人工计算的原文句段数，见表 2。

⑥ 见 http://producthelp.sdl.com/SDL%20TM%20Server%202009%20SP3/en/mergedProjects/glossary/TMSGlossary.htm。

⑦ 以句号、分号、感叹号以及段落回车（不计算以句号、分号、感叹号结尾的段落回车）为标志计算句段数量。

⑧ 见 https://www.rd.com/advice/parenting/teacher-inspires-harlem-children/。

⑨ 见 https://www.adb.org/sites/default/files/publication/159310/adbi-asean-prc-india-transformation.pdf，第 23-24 页。

⑩ 见 http://www.freepatentsonline.com/y2017/0167358.html。

⑪ SDL Trados2017 版没有查到所支持的格式数量，20 是根据 SDL Trados 2007 版统计得来的数据。

⑫ 句段切分多于原文句段数量的计算方法为：

$$\frac{\text{原文句段数}-(\text{切分句段数}-\text{原文句段数})}{\text{原文句段数}}\times\%$$

⑬ 取小数点后一位，四舍五入。

⑭ 括号内分数为对齐句段数和原文句段数之比。

参考文献

[1] 陈钰枫、宗成庆、苏克毅 . 汉英双语命名实体识别与对齐的交互式方法 [J]. 计算机学报，2011（9）：1689-1695.

[2] 王斌、刘群、张祥 . 汉英双语库自动分段对齐研究 [J]. 软件学报，2000（11）：1548-1554.

[3] 俞劲松、王惠临、吴胜兰 . 高正确率的双语语块对齐算法研究 [J]. 中文信息学报，2015（1）：67-74.

[4] Brown, P.F., Lai, H.C. & Mercer, R.L. Aligning sentences in parallel corpora [A]. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* [C]. 1991: 169 - 176.

[5] Gale, W. & Church, K. A program for aligning sentences in bilingual corpora [J]. *Computational Linguistics,* 1991（1）: 75 - 89.

[6] Kay, M. & Roscheisen, M. Text-translation alignment [J]. *Computational Linguistics*, 1993（1）:121, 142.

[7] Ker, S. J. & Chang J. S. A class-based approach to word alignment [J]. *Computational Linguistics,* 1997（2）: 313 - 341.

[8] Tan, C.L. & Nagao, M. Automatic alignment of Japanese-Chinese bilingual texts [J]. *IEICE Transactions on Information and Systems*, 1995(1): 481 – 485.

[9] Wu, D. Aligning a parallel English-Chinese corpus statistically with lexical criteria [A]. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics* [C]. Las Cruces, New Mexico, 1994: 80-87.

作者简介 蔡辉，中国社会科学院博士，中央财经大学外国语学院副教授。研究方向：语言学、语言经济学、翻译学。
作者电子信箱 lfcaihui@aliyun.com