Sentence Alignment Study of English-Chinese Military Corpus

Sentence Alignment Study

of English-Chinese Military Corpus

英汉军事语料句子对齐研究

严灿勋 ◎ 著

国防工业出版社
National Defense Industry Press

英汉军事语料句子对齐研究

Sentence Alignment Study of English-Chinese Military Corpus

◎严灿勋 著

国防工业出版社

内容简介

本书系统阐述了基于双语词典的二分图顶点最大权重配对句子对齐方法，其中包括句子对齐所用的双语词典的自建方法、英语句子边界识别方法、英语单词形态还原方法及利用双语词典对汉语句子进行分词的方法，解决了普通长度的英汉平行军事文本的句子对齐问题，对实现其他领域英汉平行语料句子对齐具有借鉴意义。

本书可供自然语言处理领域研究人员、计算语言学研究人员、语料库建设和应用研究人员、双语词典编纂平台设计使用者、机辅翻译平台设计使用者阅读参考。

图书在版编目(CIP)数据

英汉军事语料句子对齐研究/严灿勋著. —北京：国防工业出版社，2015.6

ISBN 978-7-118-10283-3

Ⅰ.①英… Ⅱ.①严… Ⅲ.①军事－英语－翻译－研究 Ⅳ.①H315.9

中国版本图书馆 CIP 数据核字(2015)第 116722 号

※

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号 邮政编码 100048)

国防工业出版社印刷厂印刷

新华书店经售

*

开本 880×1230 1/32 **印张** 5 **字数** 107 千字

2015 年 6 月第 1 版第 1 次印刷 **印数** 1—2000 册 **定价** 48.00 元

(本书如有印装错误，我社负责调换)

国防书店：(010)88540777 发行邮购：(010)88540776

发行传真：(010)88540755 发行业务：(010)88540717

PREFACE 前言

句子对齐英汉军事语料库的建设有利于军事领域的英汉翻译、军事英语教学、英汉军事词典编纂以及围绕英汉军事语料进行的各项自然语言处理工作的发展。然而，在我国，句子对齐英汉军事语料库的建设严重不足，亟待大力发展。

目前，双语句子对齐处理技术已经相对成熟。句子对齐方法大体上可以分三种：(1)基于句长的方法，该方法通常也基于概率统计；(2)基于双语词汇互译信息的方法，该方法既可以基于概率统计获得词汇互译信息，也可以基于双语词典获得词汇互译信息，但都是基于词汇互译信息计算句对相关性，根据句对相关性分值判断句子对齐结果；(3)句长和双语词汇互译信息混合的方法。

句子对齐英汉军事语料库建设的现实情况是：(1)军事领域范围广，子领域众多；(2)针对子领域的英汉平行语料少，不适合基于统计的句子对齐方法；(3)尽管基于英汉词典的句子对齐方法能够提供可靠的词汇互译信息，比其他方法更加适合英汉军事语料的句子对齐处理，但是，目前基于英汉词典的句子对齐方法所使用的双语词典容量普遍较小，需要增加单词及释义，还应该根据需要增加英汉军事术语，对齐算法也有进一步研究的必要。

本书结合句子对齐英汉军事语料库建设所面临的困难，在分析平行语料句子对齐技术的基础上，根据平行语料句子对齐的二分图模型，在相关统计数据支持下，提出了基于双语词典的二分图顶点最大权重配对句子对齐方法，并基于该方法设计实现了一个实用的英汉平行语料句子对齐处理平台。具体内容包括相关语言知识库的建设、英汉平行文本拆分及段落对齐、英汉句子边界识别、英语单词形态还原、基于双语词典的汉语分词及顶点最大权重配对句子对齐算法。

本书的意义在于：提出基于双语词典的二分图顶点最大权重配对句子对齐方法，为句子对齐英汉军事语料库建设和其他小规模双语平行语料的句子对齐自动处理找到了一个正确率很高的方法。基于此方法建设的英汉平行语料句子对齐处理平台，能够高质量实现各种规模的英汉军事语料的句子对齐处理。因为该方法中所用的英汉词汇互译信息库容量很大，所以该英汉平行语料句子对齐处理平台也适用于其他领域的英汉平行语料的自动句子对齐处理。

本书在写作过程中，得到了许多专家、学者无私的帮助和指导。在此向程工、何莲珍、李文中、濮建忠、李经伟、王岚、张淑静、赵翠莲、易绵竹、张克亮、毕玉德、高航、赵蔚彬、韩子满等教授和李景泉、邢富坤、周光磊等博士表示深深的感谢！也向好友姬增军和熊建国、赵峰、李峰博士表示衷心的感谢！

我更要感谢我的亲人，特别是我的父母、妻子和儿子！他们默默的付出、耐心的期盼和美好的祝愿是我随时可以寻求的慰藉，是我不断进取的力量源泉！

由于水平有限，书中疏漏之处在所难免，敬请读者、同仁批评斧正！

我的 Email 地址是 yancanxun@126.com。

严灿勋

2015 年 3 月 15 日于襄城

CONTENTS 目 录

第1章

»» 平行语料句子对齐方法研究概述

1.1 句子对齐的重要性

不管是在语言学、翻译学、语言教学、双语词典编纂领域，还是在自然语言处理领域，平行语料库的重要性已经勿庸置疑。句子、短语、词汇的对齐能够进一步提升平行语料在上述各领域中的研究价值和使用价值（Nagao，1984；Brown et al.，1990；Klavans & Tzoukemann，1990；孙乐等，2000；刘冬明等，2005；冯敏萱，2006）。进行短语、词汇对齐前一般需要首先实现句子对齐。句子对齐的平行语料是效用最大的平行语料（Ma，2006）。因此，句子对齐是平行语料库建设的重要一环。

机器翻译的发展促进了句子对齐方法的研究。句子对齐双语语料库对提高机器翻译质量非常有用。不管是统计机器翻译、基于实例的机器翻译，还是机器辅助翻译，都需要大量高质量的句子对齐双语语料（王飞，2004；修驰，2009；赵小曼，2010）。原始的双语语料通常只是篇章对齐，要经过加工才能实现句子对齐。很显然，人工进行双语语料的句子对齐处理费时费力，不容易大

规模实现。因此,计算机辅助实现自动句子对齐处理得到了普遍关注。

1.2 常见的句子对齐方法

目前,双语句子对齐处理技术已经相对成熟。句子对齐方法主要有三种(Moore, 2002;Ma,2006;吕学强等,2004;黄俊红等,2007;热西旦·塔依、吐尔根·依布拉音,2009):(1)基于句长的方法(Brown et al. ,1991;Gale & Church, 1991);(2)基于双语词汇互译信息的方法(Kay & Roscheisen,1993;Ma,2006);(3)句长和双语词汇互译信息混合的方法(Wu,1994;Tan & Nagan, 1995;Moor,2002)。有的研究者只认可前两种基本方法(熊伟等,2007);也有研究者将基于双语词典的句子对齐方法作为第四种方法,与前面的三种方法并列陈述(毕雪华,2006),而从本质来看,双语词典的作用就是提供双语词汇互译信息,只是其词汇互译信息不是从双语语料中提取,而是直接来源于编纂好的双语词典。

1.2.1 基于长度的句子对齐方法

20世纪80年代末至90年代初,Catizone et al(1989),Kay(1991),Kay & Röcheisen(1993)等学者开始研究利用相关的语言信息处理技术实现句子自动对齐的方法,但是由于效果不佳,没有得到推广。

在90年代初,Brown, Lai & Mercer(1991),以及Gale & Church(1991,1993)在同一时间分别提出了基于平行语料双语句子长度的句子对齐计算方法,Gale & Church(1993)还在其研究报

告后面附上了他们的句子对齐方法的核心 C 语言代码。这些方法利用双语句对之间的长度关系，使用概率统计手段，通过动态规划算法，寻找可能性最大的句子对齐结果。他们的研究为英法、英德等大规模平行语料库的高效句子对齐找到了解决方法。

在 Brown 等人的方法中，句长是基于单词数计算的，而在 Gale & Church 的方法中，句长是基于字符数计算的。Gale & Church (1991，1993)在他们的实验中利用同样的语料和程序分别测试了基于字符数的句长计算方法和基于单词数的句长计算方法，得出结论是基于字符数的长度计算方法在句子对齐中更加有效。

在基于长度的句子对齐方法中，虽然 Gale 和 Church 指出，适用于大多数欧洲语言之间的平均字符数对应比值($C \approx 1$)不适合英汉对齐，但是他们认为基于长度的方法适用于任何语言对，只是对应的参数需要根据具体的语言对进行修改。不过，对这个观点，有的学者并不赞同，Ma(2006)指出，基于长度的句子对齐方法在像英语、法语这样关系近的语言之间对齐效果可以，在像英语、汉语这样关系远的语言之间对齐效果就迅速下降了。

基于句长的句子对齐方法以这样一个简单事实为依据：源语言中较长的句子翻译成目标语时一般还会是较长的句子。但是，这里存在两个明显问题：(1)当句长相对接近，中间只要漏掉一句，余下的内容则很难对齐；(2)在长度对应比值跨度较大、离散程度较大(可以通过标准差、离均差等反映)的情况下，就需要在句子间取长补短，但是，只要有一句配对错误，又会影响余下的句子配对。我们曾经利用一种基于锚点和句长的方法在标准差(英语句子字符长度比汉语句子字符长度)低于 0.70 的双语语料中获得了总体高于 99% 的正确率的对齐效果，但是同样的方法在标准差接近甚至超过 0.85 时，正确率就受到了很大的影响。另外，

当平行语料中非1:1的句对类型(即1:2、2:1、1:n、n:1、m:n等句对(m>1 & n>1))较多时,其对齐结果也必然会有较大下降,原因是1:2、2:1、1:n、n:1、m:n等类型的句对的正确率比1:1句对的正确率低得多。尽管复杂句对类型判断正确率低的问题是句子对齐的普遍问题,但是,这个问题在基于长度的句子对齐中更加突出,原因还是前面提到的第1个典型问题。下面通过例1.1中的"带古诗的汉英平行语料"①来看基于句长的句子对齐处理的第2个典型问题。

例1.1 带古诗的汉英平行语料

汉语句子:

1. 人间四月芳菲尽,山寺桃花始盛开。(长度:18)

2. 这两句诗准确描述了当前刚刚迈进危机后时代所面临的局面:(长度:27)

3. 经过2007年至2009年三年时间里接连不断的危机,我们终于看到了复苏的曙光。(长度:39)

4. 根据国际金融机构近期发布的数据显示:(长度:18)

5. 自2010年年中开始,主要经济体开始像桃花萌发一样复苏,让我们感受到迟来的春天的气息。(长度:43)

英语句子:

1. "When spring flowers wither and fall in the early summer wind, you can still find peach blossoms deep in the mountains in a temple's backyard." (长度:144)

2. The artistic conception of these two lines of poetry exactly de-

① 语料取自《中国军事科学学会第三届香山论坛论文集》,我们按顶点最大权重配对句子对齐方法中所采用的句子边界识别方法切分了句子。为方便比较,在例子所涉及的文本中,每句前面添加了序号,后面添加了按字符长度计算的句长信息。

scribes the current situation at the beginning of this post – crisis era:（长度:134）

3. after three consecutive years of crisis from 2007 to 2009, we finally see the dawn of recovery.（长度:95）

4. Data from the first half of 2010, published recently by international financial organizations, position the major economies just like the peach blossoms, bringing us true feelings of a late spring.（长度:198）

这个“带古诗的汉英平行语料”用基于句长的句子对齐方法处理后得到的对齐结果显示在表 1.1“基于句长的句子对齐方法实例”中:

表 1.1　基于句长的句子对齐方法实例

“When spring flowers wither and fall in the early summer wind, you can still find peach blossoms deep in the mountains in a temple's backyard.”	人间四月芳菲尽,山寺桃花始盛开。~~~这两句诗准确描述了当前刚刚迈进危机后时代所面临的局面:	-0.00945
The artistic conception of these two lines of poetry exactly describes the current situation at the beginning of this post – crisis era:	经过 2007 年至 2009 年三年时间里接连不断的危机,我们终于看到了复苏的曙光。	0.261765
after three consecutive years of crisis from 2007 to 2009, we finally see the dawn of recovery.	根据国际金融机构近期发布的数据显示:	0.163636
Data from the first half of 2010, published recently by international financial organizations, position the major economies just like the peach blossoms, bringing us true feelings of a late spring.	自 2010 年年中开始,主要经济体开始像桃花萌发一样复苏,让我们感受到迟来的春天的气息。	0.184426
		0.3

对例 1.1 所用的句子对齐程序是 hunalign[①]，不加词典。表 1.1中的句子对齐结果是 hunalign 给出的全部结果。该对齐程序在不加词典时首先按 Gale & Church 的基于句长的句子对齐方法实现句子对齐，然后根据统计结果生成对齐词典，最后再根据对齐词典对结果进行调整[②]（Varga et al.，2005）。本例中的文本特别短，并且没有对汉语进行分词，不能生成有效的对齐词典，对齐结果的主要依据是基于句长的句子对齐方法处理所得的结果。

认真阅读上面的句子对齐结果可以发现，表 1.1 中的句子配对全部错了，原因主要是在例 1.1 的“带古诗的汉英平行语料”中，汉语句子 1 是一句古诗，它的英译太长，英语句子 1 本来应该与汉语句子 1 相配对，但由于汉语句子 1 太短，程序将汉语句子 2 加到了汉语句子 1 上，一起与英语句子 1 配对。汉语中的习惯用语、成语经常会有类似的翻译现象，英译很长。

从上面的例子可以看出，完全基于长度的句子对齐方法不适合情况比较复杂的汉英平行语料。钱丽萍等（2001）根据他们的实践指出，完全基于长度的方法由于使用根据一定语料训练出来的参数进行概率统计，当语料发生变化时，参数不一致，就会造成很多错误。还有实验证明，在基于长度的句子对齐方法中，用根据技术手册语料训练出来的参数去对齐通用杂志语料，F 值[③]能够从 98.2% 下降到 85.6%（Kueng & Su.，2002）。

显然，英汉平行语料不适合使用完全基于句长的句子对齐方法进行句子对齐处理。

---

① hunalign－1.2 在不加双语词典时首先按 Gale & Church 的基于句长的句子对齐方法实现句子对齐，然后根据统计结果生成包含词汇互译信息的对齐词典，最后再根据对齐词典对结果进行调整（Varga et al，2005）。

② 根据 Hunalign－1.2 软件包中的 readme. html 介绍。

③ F 值是句子对齐的正确率和召回率的调和平均值。

1.2.2　基于词汇互译信息的句子对齐方法

基于双语词汇互译信息的方法，既可以是根据语料，用统计的方法找出的词汇互译信息，也可以是基于双语词典，直接从源语词目及与其对等的目标语释义中获取词汇互译信息。下面首先讨论基于语料统计所得词汇互译信息的句子对齐方法，然后介绍基于双语词典的句子对齐方法。

1. 基于语料统计所得词汇互译信息的句子对齐方法

Chen(1993)提出了以概率统计为基础的、基于语料统计所得词汇互译信息的句子对齐计算方法。这种方法不需要双语词典，而是先利用100对已经对齐了的平行句对得出相关的词汇互译信息，在此基础上，再对同一领域的其他对齐语料进行基于词汇互译信息的句子对齐处理，并在对齐过程中继续基于新获取的句对扩充词汇互译信息，使得对齐效果继续提高。该方法在实验中总体上获得了很好的对齐效果，同时证明，随着词汇互译信息的增加，句子对齐效果会越来越好。Chen声称，他提出的句子对齐方法可以适用于任何语种。

2. 基于双语词典的句子对齐方法

基于词典的句子对齐方法一般是利用词典从待对齐的句子中查找出对译的词或词组，然后根据对译的词或词组的多少或长短计算待对齐句子之间的相关性分值，计算方法多种多样，多数在评价函数中结合了句长。下面举两个例子：

例1.2　钱丽萍等(2001)提出的评价函数(1)：

$$\text{distance}(s,t) = \left(\sum_{e \in S} \text{d\_match}(e,T) + \sum_{e' \in S} \text{l\_match}(e', T) \right) / (\text{len}(S) + \text{len}(T))$$

在评价函数(1)中,S、T 分别是英语和汉语句子,len(x)是字符串 x 的长度,e 和 e′是英语句子 S 中的单词,当 e′不是英语词的原形的时候,需要先进行形态还原。当 e 在词典中的任何一个释义与 T 中某一字符串匹配时,相关性分值增加 1 分,e′类似。该评价函数以对齐的段落为单位,利用动态规划算法,求得从段落头到段落尾的总评价值最大的句对路径,该句对路径即为该段的句子对齐结果。根据他们的实验报告,在使用释义比较全面的词典时能够取得较好的实验效果。

例 1.3 杨沐昀(2002)提出的评价函数(2):

$$H = \frac{(Length(|MatchDic(Se)|) + Length(|MatchDic(Sc)|))}{(Length(|Se|) + Length(|Sc|))}$$

在评价函数(2)中,MatchDic(Se)表示匹配到的英语文本,MatchDic(Sc)表示匹配到的汉语文本,Se 表示待对齐的英语句子,Sc 表示待对齐的汉语句子。Length(X)表示字符串 X 全部元素的字节长度之和。从上面的公式可以看出,H 是匹配到的英汉文本字节长度在英汉句子总字节长度中所占的比值。在理想状态下,该比值越大,匹配度越高,英语句子 Se 与汉语句子 Sc 之间存在对齐关系的可能性越大;反之则匹配度越低,对齐的可能性越小。杨沐昀利用上述方法在实验中取得了很好的句子对齐效果。熊伟等(2007)在利用上述方法进行句子对齐研究时指出,当词典较小时,上述公式计算出的匹配度会急剧下降,以至不能正确识别存在匹配关系的句对。这种现象在基于双语词典的句子对齐方法中普遍存在,它从反面证明:在基于双语词典的句子对齐方法中,词典的容量越大,提供的词汇互译信息越充分,对齐效果越好。

Ma(2006)介绍了基于词典的开源句子对齐工具 Champol-