计算机科学2002Vol.29№.11

基于语义相似度并运用语言学知识进行双语语句词对齐*)

Using Semantic Information and Language Knowledge in Bilingual Word Alignment

晋 薇　　黄河燕　夏云庆

(中国科学技术大学)　(中国科学院计算机语言信息工程研究中心　北京100080)

Abstract This paper claries the definition of alignment from the viewpoint of linguistic similarity. Many alignment algorithms have been proposed with very high precision. But the languages belong to occidental family. We propose a new method for alignment between languages that do not belong to the same language family. On the contrary to most of the previously proposed methods that rely heavily on statistics, our method attempts to use linguistic knowledge to overcome the problems of statistical model. Experimental results confirm that the algorithm can align over 85% of word pairs while maintaining a comparably high precision rate, even when a small corpus is used in training.

Keywords Machine translation, Word alignment, Semantic information, Bilingual dictionary

一、引言

自八十年代以来,基于统计(Statistics-Based)和基于实例(Example-Based)方法的出现及其广泛应用给机器翻译的研究工作注入了新的活力,标志着机器翻译进入了一个新时期。这两种方法共同的特点是:都需要一个双语语料库(Bilingual Corpora)直接或间接地作为翻译的知识库。这种双语语料库中包含了原文和译文相互对应的语言信息,是支持机器翻译的最为宝贵的资源。

双语语料比单语种提供了更多的信息。在近些年里,在对篇章、段落、句子对齐进行了大量的研究之后,许多工作已经涉及到词对齐研究。双语词对齐是机器翻译中实现翻译记忆复用的关键环节,因为只有确定例子中原译文之间片段的对应关系,才能够确定原语言中被改变部分在译语中对应的范围,从而确定译文中应当被修改的范围。除此之外,它还能够用于双语词典的自动构造以及构造其它有用的资源中,并且在许多的应用领域中都有广泛的应用价值。诸如:词义消歧等。

二、对齐技术概述

Brown 在1990年提出将统计的方法用于机器翻译,通过语言模型和翻译模型实现了英法双语的篇章、段落、句子、短语乃至词汇的对齐,其他文献也相继提出了基于统计的双语语句词对齐方法。这些方法均以大规模双语语料库为基础,通过建立统计模型,计算双语词汇在双语语料库中的同现概率,以此建立双语词对齐关系。

有许多统计的工具用来帮助解决相应的双语词对的关联程度。Gale 和 Church 使用 Φ^2从双语文本中识别词的对应;Fung 和 Church 在1994年提出了 K-vec 算法,基于 K-way 分割双语文本来获得双语词典,但它们都不能避免由于文本长度和出现频率所带来的的低覆盖率的问题(低频率出现的词占大多数而高频率出现的词趋向于有多种翻译方式)。这些方法假定双语语句的词汇个数遵从某种相对比例,对于同语系(例如英法)的双语词对齐处理效果较好,但对于跨语系(例如汉英)的双语词对齐,由于汉语和英语分别属于汉藏语系和印欧语系,语系的差别导致上述假定不再成立,因而方法很难奏效。

与这些主要基于统计的方法相反,Ker(1997)的英汉对齐系统在词对齐中运用了基于 class 的算法,而且系统没有使用任何的统计技术。结果显示这种新的尝试在获得很高的准确率的同时,能克服在统计学的方法中存在的低覆盖率的弊端,即使系统基于一个很小的测试集。Ker 的工作启发了我们将纯语言知识运用于解决对齐问题的可行性。

在词对齐的过程中短语级的对齐是其中一个很困难的问题。虽然一些统计学的方法被用于解决这个问题,但是它们中大部分获得的并非语言学上的词组而是一些统计的结果(Wu,1994;Shin & Choi,1996)。当双语对不属于同一语系时,短语一级的对齐尤为困难,主要因为在这种情形下两种语言的语法结构和词典信息都不尽相同,有更多的特例是不遵循多对多、一对一这种对应关联,所以传统的对齐系统中运用的信息不足以满足对齐的要求,导致覆盖率和精确性都有所下降。

针对基于统计方法的不足,许多研究者提出了将语言学知识融入统计方法的观点。在下面的部分,我们将给出一个基于双语词典获取语义信息并充分利用语言学知识来进行词对齐的尝试。

三、设计考虑

在大部分先前的研究中,关于对齐的定义被认为是将单词(短语、文本等)与它的译文相对应。看上去对齐的概念十分清晰,可能不会有人会考虑到这样一个问题:"What is the translation of an original word?"在这里,我们将如下阐述这个定义:

对齐是一项查找原语言所对应的目标语言的工作。词对齐是发现与原语词汇具有最高语义相似度的目标语。当有多于一个候选词汇的时候,系统应该对应在候选集中与原文词汇具有最高句法相似度的目标词汇。如果词一级的对齐由于语言学的一些特点而未能实现,对齐目标的范围可以延伸到短语、片段,它们应为具有与原文单词或短语片段最大的语义相似度的最小语法单位。

*)本文的研究工作得到国家自然科学基金资助。晋　薇　硕士生,研究方向为自然语言信息处理。黄河燕　研究员,博导,多年从事机器翻译领域的研究,成功实现了面向通用领域的高性能英汉机译系统。夏云庆　博士,从事多策略机译系统研究与应用开发工作。

从上面具体的定义中，我们可以很容易发现对齐的问题实质上是双语词汇相似度计算的问题。

那么什么样的信息在获取双语词汇相似度中是直接的信息呢？在单语种的处理中，字典、分类词汇词典和 WordNet 通常被使用。我们在对齐过程中需要双语信息，所以尝试使用双语词典，由它提供给我们与源词具有最高语义相似度的目标词。同时由于词一级也许不是翻译的最小单位，翻译经常是一部分一部分进行的，这一部分可能包括一个词组、一个词的搭配、一个固定表达甚至是一个句子，因此许多结构或者模式必须被作为一个翻译模板，不能被分解为更小的单元。如果我们处理的是子句级的对齐，那么我们面临的一个主要问题便是边界的界定和正确的分块。在实现中，我们的系统在处理词组一级的对齐时对双语语料的输入句采取了最优长度分割、动态抽取的策略。

四、对齐算法

我们采用了基于语言学上相似性的观点并充分利用语言学知识来进行词对齐的尝试。对于双语信息，我们尝试用双语词典、语义分类词典来解决。在实现过程中，我们还将综合考虑双语词汇在语义、词性、所属词类以及特定语法结构方面对对齐所产生的影响。对于短语词组级的界定采取了最优长度分割、动态抽取的策略。在双语语句词对齐技术上，克服了对大规模双语语料库的依赖性，有利于避免大规模双语语料库繁重的建库工作，该方法只需一部双语词典，一部语义词典和少量的语法规则和语言信息，有利于降低语言分析的复杂度，从而提高了双语词对齐的效率，增强了类比翻译处理的性能。

使用到的一些知识包括以下几个部分：一个句子对齐的双语语料库；一部双语词典；一部语义词典。第四个资源是一些特定的语言信息和少量的语法 规则(放在构造文件中)，包括一些在对齐过程中需要考虑的特殊的语法结构、短语搭配以及在对齐中被省略掉的词(例如冠词，助词)和一些可能被插入的词以及一些专有名词。少量的语法规则用来进行规约限定，协助排除有歧义的对应。双语词典和目标语言的词根/同义词表以及语义词典提供了必要的信息用来发现在被选择的语句对中原语言和目标语言单词(词组)之间的联系。这些联系被用来进行子句的对齐。一个原语言的单词被认为和一个目标语言的单词关联，当且仅当目标单词本身或者是该单词所属词根/同义词表中的任何单词出现在对原文单词的可能的翻译集中。

在对短语(词组)的识别过程中，我们将利用双语词典中提供的短语(词组)信息，采用最优长度分割、动态抽取的策略，同时借助一定的语法规则来调整具体实例中出现的情形，例如添加不定成分后的介词短语(如 devote one's life; try one's best)，在动词短语识别过程中出现的指示代词前移至动词与副词之间(如 give it up; break it down)等，使其能被正确地辨认出而不被遗漏。

对于最后对齐结果中出现的多对应情形，存在着歧义消除的问题。在这里我们使用位置信息和已建立的词对应关系来区分多词对应中哪一种最为可能。

双语词汇对齐算法如下：

算法 从一条双语对齐的语料中进行词汇对齐(英汉语料)
输入：原文句子 SrcText，译文句子 DstText。
输出：对齐的单词对或词组对。
算法流程：
BEGIN
英语分词处理与词法分析；
汉语分词处理与词法分析；
过滤掉对齐过程中不予考虑的汉语词汇(如表示语气的、状态的)以及在英语中没有对应的汉语虚词。
WHILE not at the end of SrcText.
//以词组为单位的意段抽取
1. 从 SrcText 文本中试探性地切分出词组(逐级递减单词，并根据分词词性利用语法规则调整好词序)，对其进行双语词典查询(以词典中收录的词组短语为识别基础)，如确认为词组，则取出它的所有义项；计算这些义项与目标语言词汇之间的相似性函数，如果不小于预先设定的阈值，则转向4；否则将假定词组包含的单词数减一，继续词组的识别直至剩余单个单词为止。
2. 若相似度计算的结果小于阈值，则进行基于语义分类词典(目标词汇同义词侯选)的比较计算工作，若存在大于阈值的同义词侯选项，转向4；
3. 若尚有单词序列词组抽取工作未处理完，则继续进行新一轮的以词组为单位的意段抽取，否则进行以单词为单位的意段抽取。
4. 进行词性检验。词性相似性暗示在原词和目标词之间的某种对应性。因此，从原文中分割出来的单词或词组的词性应包含(等于)译文中对应词的词性。

将当前原文词组和 DstText 中的义项作为单独的意段抽取出来进行保存，并在它们之间建立翻译对应关系；
//以单词为单位的意段抽取
当词组切分过程中仅余下单个单词时，将以单词为界标，进行单词一级的匹配，重复上述对于词组匹配的操作，建立翻译对应关系。
END WHILE
使用位置信息和已建立的词对应关系来消除有歧义的双语多对应，分离出最优解。
END

双语词汇的语义相似度度量函数：

$$\mathrm{WordSim}(E,K)=d\times\max_{K_{Ei}\in K_E}\frac{2\times|K_{Ei}\cap K|}{|K_{Ei}|+|K|}$$

其中：E＝给定源句中的英文单词(词组)；K_E＝ 双语词典中给出的对应于 E 的汉语解释词汇集；K_{Ei}＝ 集合 K_E中的元素；K＝给定译句中的汉语词汇(语义词典同义词表中给出的同义词侯选)；|K|＝K 中所含的字符的个数；$|K_{Ei}|$＝集合 K_E中第 i 个元素所含的字符个数；d＝d＜1.00，如果 K_{Ei}和 K 有不同的词性标注；否则的话 d＝1.00。

在设计考虑时，我们还将面临以下几个问题：

1)同一个词汇在句子中多次出现，但是它们在译文里的次序是未知的，方法需要明确地构造一个合理的对齐关系。

2)存在多个词汇同时对应译文中的一个词汇时，方法需要验证这种对应的合法性。

3)在词组的识别过程中，由于英文的习惯用法，有一些特殊的语法结构需要考虑。例如：当代词充当动词短语的宾语时，将代词放在两者之间，如：break it down; give it up 等。

为解决上述问题，我们提出如下设想：

1. *使用邻近关系来发现中文复合词的翻译对应* 我们能够注意到有大量的中文复合词，它们的英语翻译包括多个英文单词。如果同一个中文词汇与两个或更多的互相邻接的英文单词相关联，那么我们就能合理地推出这些英文单词是这个中文词汇的翻译。在实现中，如果两个或三个连续的英语单词部分地匹配同一个中文词汇的不同汉字，那么它们很可能是正确的翻译词对，例如：中文对于 oral exam 的翻译是口试。基于 oral 和 exam 相邻，它们的中文翻译匹配口试一词的第一和第二个汉字，那么它们很可能是口试的翻译。在下面的例子中也可以运用此类方法：

deep--深--根深蒂固	academic--学术--学术界
root--根--根深蒂固	world--界--学术界
sexual--性--性骚扰	eldest--长--长子
harassment--骚扰--性骚扰	son--子--长子

2. *特定的语法结构* 在词组的识别过程中，由于英文的习惯用法，有一些特殊的表达习惯。例如：当代词充当动词短语的宾语时，将代词放在两者之间，如：break it down; give it up 等。在这种情况下，可能导致词组识别的失败。我们将借助

词法分析过程中的词性信息用少量的规则对此类情形进行约束。例如：v. + pron. +adv. →v. +adv. + pron. 以正确地规约出 break down，give up 此类的动词词组。

再如：英语短语中还常出现 allure sb. from、all one's life、try one's best 、absent oneself from、acquaint oneself with、devote one's life 这样的形式。在具体实例中将用特定的物主代词、反身代词替代 one's、sb's、oneself，在分析词组时也需将其还原为特定的词组表示形式。

诸如此类的语法结构我们将总结在构造文件中。

3. *同一词汇在句中多次出现* 在一个句子中同一个词汇出现了多次，这产生了含混的词对应关系。例如：

if you want to go , I will go with you.

我将和你一起去如果你想去的话。

在此句中，go 和 you 均出现两次，有可能会造成错误的词对应。在对齐过程中，为了避免这类错误的发生，在算法中我们采用了局部最近匹配的方法。所谓局部最近匹配，即以最近对齐的词汇为界标，在局部范围内进行匹配，以最接近于界标的词汇作为候选词汇。基于在任何翻译过程中最基本的要求是不同语言的翻译对中的意义的保持，我们认为在局部范围内意义存在着延续性。在例句中，if 和"如果"匹配上以后，那么 if 后的 you 将从"如果"后进行匹配，因此，"如果"后的"你"将和 if 后的 you 对应，同样，go 也将正确地对应上。当指针指到句末，如果还有未处理的汉语语素的话，将继续进行匹配，在此例中，指针将循环至句首，这样，"I will go with you"也将与"我将和你一起去"正确匹配上。

4. *使用位置信息和已建立的词对应关系来消除有歧义的双语多对应* 对于最终对齐结果中出现的多对应情形，存在着合理性检验的问题。在这里我们使用位置信息和已建立的词对应关系来区分多对应中哪一种最为可能。在评价词对候选集的可行性时我们使用了一种新的 distortion 方法。为了更精确，我们采用了 Relative Distortion 来形成一个位置评价模型，这个模型将比纯粹基于绝对位置获得信息的方法规模更小。

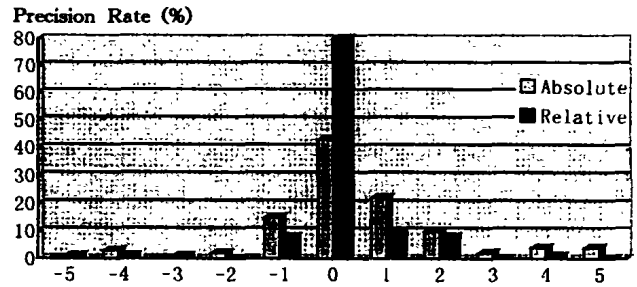

Figure1 Precision rates for candidates with different values of distortion

这种方法是基于许多语言结构在翻译过程中被予以保持，因此，一种正确对应的目标位置相对于同一结构中的其他一些对应在统计学的分布方面将会具有更小的变化。假定一些对应已经被无歧义地建立，我们将相对于这些对应来评价候选集(s,t)，其中 s 和 t 分别为 ST(源句)和 TT(译句)中的第 i 个和第 j 个词，那么将会存在两个最近的已建立的对应(i_L, j_L)和(i_R, j_R)分别在 s 的两边。Relative Distortion rd (s,t) 使用下面的公式来逼近：

$$rd(s,t)=\min(|d_L|,|d_R|)$$

where $d_L=(j-j_L)-(i-i_L)$

$d_R=(i-i_R)-(j-j_R)$

经验数据证明：具有较小的 rd 值的对应更可能是正确的。图1显示了当一个候选词对具有0 rd 值时将比一个候选词对(具有0 absolute distortion)更可能是正确的对应(0.80：0.43)。

使用这种限制，我们观察到大约有95%的正确的词对齐被保留下来，大约有80%的不正确的词对应被排除了。

五、算法运行实例及分析

在这一部分，我们将给出有关词对齐算法的结果。我们随机抽取了3类英汉双语语句，其中一类来自中英文网站，一类来自于《英语900句》，还有一类来自于金山词霸的例句库和中英文对照小说，并利用下面的公式计算对齐准确率：

$$\zeta=\frac{1}{N}\times(\Sigma\frac{N_{CR}}{N_C})\times 100\%$$

其中 ζ 表示对齐准确率，N_{CR}表示语句中正确对齐的单词(词组)个数，N_C 表示语句中单词(词组)个数，N 表示语句总数。测试结果如表所示。

No. Words	No. Matched	No. Correct	Coverage	Precision
2862	2524	2272	88.2%	90.0%

以下是对齐算法的一些运行实例：

1. 原文句子：We don't think such an abnormal phenomenon will last long.

译文句子：我们认为这样的反常现象不会持续很长。

对齐效果显示：

we：我们；think：认为；such：这样的；abnormal：反常；phenomenon：现象；

not：不；last：持续；long：长

2. 原文句子：My brother has never been abroad before, so he is finding this trip very exciting.

译文句子：我弟弟以前从未到国外，所以他觉得这次旅行十分令人兴奋。

对齐效果显示：

My：我；brother：弟弟；before：以前；never：从未；abroad：到国外；so：所以；he：他；this：这；trip：旅行；exciting：令人兴奋

3. 原文句子：Buyers have withdrawn from the market in view of the abrupt turn of the trend of prices.

译文句子：由于价格趋势的突然转变，买主已退出市场。

对齐效果显示：

in view of ：由于；prices：价格；trend：趋势；abrupt：突然；turn：转变；buyer：买主；withdraw from ：退出 market：市场

六、总结及算法的改进

双语词对齐操作是基于实例机器翻译的核心内容，是对翻译实例进行问题求解方法(翻译模式)的分析过程，是实现翻译记忆复用的关键环节。只有建立起翻译实例源译文的双语语句词对齐关系，才可能重用翻译实例的翻译模式，通过类比手段，建立输入内容的译文。

在实现过程中，对于双语信息我们用双语词典、语义分类词典来解决。在实现过程中，我们综合考虑了双语词汇在语

义、词性、所属词类以及特定语法结构方面对对齐所产生的影响。对于短语词组级的界定采取了最优长度分割、动态抽取的策略。该方法不依赖大规模双语语料库，因此非常适合在机助翻译系统中应用；算法只需要一部双语词典、一部语义词典和少量的语言模板，从而降低了语言分析的复杂度，提高了双语词对应的效率。

未来对算法的进一步改进将集中于如下几个方面：

(1)继续完善语义分类词典来提高匹配效率。由于词典收录与实际运用的差异，词典难以全面地收录每个英文单词(词组)的所有汉语解释，使得某些存在的词汇对应关系不能被发现。但是，这些背离将大量被限制在分类词典的类中。因此，用这种基于分类体系的方法，问题的复杂性将会被降低。同时它也显示了更小的存储需求和更高的系统效率的优越性。

(2)我们注意到意译是导致对齐失败的主要因素。一种很大程度的意译是由于格式或体裁的需要而对四字成语的使用。我们考虑可以用 glossary-based approach 来辨识四字成语及习语，提高对齐的正确率。

(3)加强对未登录词的识别与标注处理，增强算法的健壮性。

(4)作为以实用为目标的算法，尚缺乏大规模复杂格式及各种特殊语言现象语料的检验。系统需要加强对各类文本的识别和过滤能力，进一步提高对齐的精度和正确率。

参 考 文 献

1 Ker S J, Chang J S. Aligning More Words with High Precision for Small Bilingual Corpora. Computational Linguistics and Chinese Language Processing, 1997, 2(2): 63～96

2 Brown R D. Example-Based Machine Translation in the Pangloss System. In: Proc. of the 16 th Intl. Conf. on Computational Linguistics (COLING-96), Copenhagen, Denmark, Aug. 1996. 169～174

3 Gao Zhao-Ming. Automatic Acquisition of a High-Precision Translation Lexicon from Parallel Chinese-English Corpora. Department of Foreign Language and Literatures National Taiwan University

4 Huang Jin-Xia. Key-Sun Choi Using Bilingual Semantic Information in Chinese-Korean Word Alignment. Korterm, Computer Science Division Korean Advanced Institute of Science and Technology

5 Dagan I. Bilingual Word Alignment and Lexicon Construction. Tutorial Paper Given at the International Conference of Computational Linguistics, Copenhagen

6 黄河燕，陈肇雄，宋继平. 一种人机互动的多策略机器翻译系统 IHSMTS 的设计与实现原理. In: Proc. of the Conf. on Machine Translation & Computer Language Information Processing, June 1999. 270～276

全国搜索引擎和网上信息挖掘学术研讨会

征文通知

随着网络在全社会的普及和应用的不断发展，有关搜索引擎技术和 Web 信息挖掘的研究已成为 Internet 领域的一个新的研究热点。为了促进国内相关领域科研人员的学术和工作交流、研讨本领域的最新技术进展和发展趋势，以推动搜索引擎和 Web 挖掘技术在中国的发展。由中国计算机学会互联网专业委员会主办，北京大学信息科学技术学院承办的“全国搜索引擎和网上信息挖掘学术研讨会”于2003年3月14-15日在北京大学举行。欢迎高等院校教师、科研院所和企业的科研人员及博士生、硕士生参加。现将有关征文要求通知如下：

一、征文范围 涉及搜索引擎和海量 Web 信息挖掘领域的相关技术与方法，如：海量网络信息收集、组织与存储，网上文件的搜索与索引服务、主题搜索、网上信息语料库、信息提取、自动文本索引与分类、Web 挖掘、个性化服务等。

二、征文要求：

1. 已经发表或尚未发表的工作都欢迎，但前者需要注明已发表出处。

2. 每篇来稿篇幅不超过6000字(含图表)，论文格式参见会议主页。

3. 每篇论文请附上作者联系信息(通讯地址、电话、电子邮件)。

4. 经程序委员会评审录用的会议论文，将被收录到由国内著名出版社出版的会议论文集中。如果是已经发表的工作，由作者负责得到相关出版物的转载许可，否则可以参加会议交流，但不收录到论文集中。

5. 来稿请寄：北京大学 计算机科学技术系 **王继民** 收，邮编：100871 鼓励用电子版方式提交论文(word 或 pdf 格式)，E-mail 发至：wjm@net.cs.pku.edu.cn 联系电话：010-62758485-21

三、征文截止日期 2003年1月15日

四、会议主页 http://net.cs.pku.edu.cn/～sedb2002/

五、会议注册 请于2003年2月15日前到会议主页上注册。