文章编号：1003-0077(2018)08-0027-05

基于统计和词典方法相结合的韩汉双语语料库名词短语对齐

凌天斌，毕玉德

（解放军战略支援部队信息工程大学，河南 洛阳 471003）

摘　要：韩汉双语语料库短语对齐对于基于实例的韩汉机器翻译系统具有重要意义，该文从韩国语名词短语结构特点出发，在基于统计和基于词典的词对齐方法进行试验分析的基础上，提出了基于词对齐位置信息的韩汉双语语料库名词短语对齐方法。该方法通过基于统计的方法获得词对齐位置信息，在此基础上利用基于词典方法的相似度计算进行词对齐校正；根据以上结果，该文通过韩国语名词短语左右边界规则抽取名词短语及其汉语译文，利用关联度度量方法进行过滤，实现名词短语对齐。实验结果表明，在较大规模语料库情况下，该方法取得了较好的短语对齐结果。

关键词：双语语料库，词对齐，短语对齐

中图分类号：TP391　　**文献标识码**：A

Noun Phrase Alignment in the Korean-Chinese Bilingual Corpus Based on Statistics and Lexicon

LING Tianbin, BI Yude

(The PLA Strategic Support Force Information Engineering University, Luoyang, Henan 471003, China)

Abstract: Phrase alignment in a bilingual corpus is of great significance to the example-based Korean-Chinese machine translation system. This paper begins with a study of the structural features of Korean noun phrases, conducts an experimental analysis of the statistics- and lexicon-based methods of word alignment, and puts forward the method of the noun phrase alignment of Korean-Chinese bilingual corpus based on the results of the analysis. This approach resorts to statistics to obtain information of word alignment position, based on which the word alignment correction is conducted from the similarity calculation in lexicon. Then the noun phrases and their Chinese translations are extracted from the rules of left and right boundaries of the Korean noun phrases, and the method of correlation measurement is applied to filter the noun phrases and realize their alignment. The experiments show that the proposed method has achieved satisfactory results of phrase alignment in the case of a large-scale corpus.

Key words: bilingual Corpus; word alignment; phrase alignment

0　引言

在基于实例的机器翻译系统中，翻译实例获取根据粒度区分，可以分为篇章级、句子级、短语级和词语级等，其中词语对齐是基础，而短语对齐在很大一部分程度上依赖于词语对齐。本文讨论的是利用较大规模韩汉双语平行语料库，在统计和词典相结合的词对齐方法基础上，实现基于双语语料库的短语对齐。由于短语对齐比句子对齐提供了更细程度的对译信息，因此对于它的研究具有重要意义。

在短语对齐方法方面，短语级别上的对齐可以归结为双语平行语料库上的多词单元的对应。许多学者在多词单元对齐和自动构建双语翻译词典方面做了进一步的研究，基本方法有 n-gram、有限状态机、近似字符匹配、双语语法分析树等。其中 Marcu[1] 说明了单个词作为翻译基本单元的不足，并说明了在翻译中加入短语翻译对的原因，并且证明了加入短语翻译对可以提高系统性能。Zhang[2] 等人为双语句对建立一个互信息矩阵，并将矩阵中抽取的互信息值相似的区域视为短语对。Zhang 和 Stephan Vogel[3] 提出了将短语对齐视为句子分割问题的方

收稿日期：2017-10-17　**定稿日期**：2017-11-28

法,在源短语固定的情况下,寻找目标短语的最优左边界和右边界。常宝宝[4]等人提出了基于词语关联度进行词语组合方法,并利用假设—检验的方法,在汉英双语语料库中抽取翻译等价单位。程洁[5]等人采用结合阈值和关联度提取的方法获取多词单元翻译词典。屈刚[6]等人针对汉英句子候选句法分析树集中存在大量的翻译异常现象,使得源语言句法树和目标语言句法树往往不存在简单的对应关系这一问题,提出了"有效句型"概念和"翻译中相对不变准则"的短语对齐模型。

本文在现有资源的基础上,首先从韩国语名词短语结构特点出发,在统计和词典相结合的词对齐方法基础上,提出了基于词对齐位置信息的韩汉双语语料库名词短语对齐方法。该方法在较大规模语料库情况下,取得了较好的短语对齐结果。

1 韩国语名词短语结构特点

在韩国语研究方面,早期的研究都是以句子为单位,组块识别和短语结构分析是近年来关注的焦点。韩国语名词组块的研究则以基本名词短语的相关研究为主[7]。安帅飞[8]等人提出了采用左右边界判定进行名词短语获取的方法,并在此基础上总结归纳出了八类名词短语类型:

(1) 名词|代词+? +名词|名词叠加;

(2) 两个或两个以上名词(代词)混合叠加;

(3) 名词|代词+接续助词|特殊的副词+名词|代词;

(4) 冠形词+名词|代词;

(5) 数字|数词+名词;

(6) 名词|名词叠加+? +名词;

(7) 名词+名词派生接尾词+肯定指示词+冠形转成词尾+名词;

(8) 名词|代词+数词+(依存名词)。

其中,语料库中韩国语采用"世宗计划"语料库的分词标注体系进行分词标注。根据八类名词短语形式,通过定义正则表达式的方法实现语料库中名词短语的抽取。

该方法的主要原理是:根据名词短语左右相邻词出现规律,确定名词短语左右边界,实现名词短语的获取。

2 词对齐方法

2.1 词典模糊匹配词对齐方法

双语词典具有丰富的词汇对译信息,是可以充分利用的优秀资源,基于词典的词语对齐方法是利用双语电子词典来进行双语词语对齐的算法。由于真实翻译中上下文的多样性和翻译的灵活性,为了提高词典译文的覆盖率,我们引入了词典的模糊匹配。

词典的模糊匹配采用词语相似度计算的方法实现,通常用Dice系数进行两个字符串之间相似度的计算,词语相似度如式(1)所示。

$$\mathrm{Dice}(t_1,t_2)=\frac{2\times\mathrm{comm}(t_1,t_2)}{\mathrm{len}(t_1)+\mathrm{len}(t_2)} \tag{1}$$

式(1)中,$\mathrm{comm}(t_1,t_2)$是t_1和t_2中相同字符的个数,$\mathrm{len}(t_1)$是字符串t_1的长度,$\mathrm{len}(t_2)$是字符串t_2的长度,$\mathrm{Dice}(t_1,t_2)$取值在0到1之间。

在获得同一种语言中词语相似度$\mathrm{Dice}(t_1,t_2)$的基础上,则源语言词语s与目标语言词语t的相似度为,如式(2)所示。

$$\mathrm{Sim}(s,t)=\max_{d\in DTk}\mathrm{Dice}(d,t)+(\mathrm{Count}_{d\in DTk}(\mathrm{Dice}(d,t)>h)-1)\times 0.1 \tag{2}$$

式(2)中,DTk为源语言词语s的所有译文。h为定义好的相似度的阈值,Count为次数统计函数,d为源语言词语s译文中的一个。若源语言词语s存在多个译文,在计算词语相似度时,将所有译文与目标语言词语t分别两两计算,取最大值作为两个词语的相似度值。

基于词典的词语对齐方法可以得到比较可靠的非空匹配,但由于双语词典的覆盖面是有限的,在未登录词、上下文关系方面存在一定的局限性,使得该方法达到的正确率和召回率都十分有限。

2.2 基于语义相似度的词对齐方法

在真实翻译过程中,译文往往具有很强的灵活性,常常会存在同义词替代翻译词的现象。中国科学院计算技术研究所的王斌[9]等人于1999年引入了语义作为基于词典的词语对齐方法的补充。

《同义词词林》是现代汉语中比较常用的一部义类词典,哈尔滨工业大学信息检索实验室在此基础上完成了《哈工大信息检索研究室同义词词林扩展版》,它收录了了各类词语7万余条,按照树状的层次结构把所有收录的词条组织到一起,把词汇分成大、中、小三类,大类有12个,中类有97个,小类有1 400个。小类根据词义的远近和相关性原则分成若干个词群。每个词群中的词语进一步分成若干行,同一行的词语在词义方面相同或具有很强的相关性。通过词义代码可以看出、这种分类方法具有

层次性。通过抽象可以将该分类体系用一个树形图表示，则根节点的子节点就是所有大类，所有大类的子节点就是所有中类，中类的所有子节点就是所有小类。

通过《同义词词林(扩展版)》的树形结构，田久乐[10]等人提出了义项相似度算法，该算法主要思想是：利用同义词词林获得词语义项的代码，通过义项之间的语义距离计算出义项相似度。该算法基于义项代码所在分支的区别进行判断，义项代码从哪一层开始不同，就使用该层对应的系数与调节参数和控制参数相乘，得出两个义项的相似度。如式(3)所示。

若两个义项不在同一颗树上，则

$$\mathrm{Sim}(S_1,S_2)=f \tag{3}$$

若两个义项在同一颗树上，则

$$\mathrm{Sim}(S_1,S_2)=x\times\cos\left[n\times\frac{\pi}{180}\right]\times\left[\frac{n-k+1}{n}\right] \tag{4}$$

式(4)中，n 是分支层所在节点的总数量，k 为两个义项在分支层之间的距离，x 为在不同分支层对应的初值，是一个常数，根据该算法的实验结果，该值在第二层时，取值为 0.65，在第三层时，取值为 0.8，在第四层时，取值为 0.9，在第五层时，取值为 0.96，$\cos\left[n\times\frac{\pi}{180}\right]$为调节参数，该调节参数的功能是把义项相似度控制在[0，1]之间。式(3)中，f 值为义项不在同一棵树上的值，是一个常数，取值为 0.1。

由式(4)可知，两词义 S_1 与 S_2 之间的语义距离可以定义为语义树中节点 S_1 到节点 S_2 的最短路径的长度，通过比较两个词的语义编码可计算出它们的语义距离。两个词语的距离越大，其相似度越低；反之，两个词语的距离越小，其相似度越高。

在义项相似度定义的基础上，定义两个汉语词 c_1、c_2 的语义相似度公式，如式(5)所示。

$$\mathrm{ClassSim}(c_1,c_2)=\max_{\substack{S_m\in\mathrm{Senseof}(c_1)\\S_n\in\mathrm{Senseof}(c_2)}}\mathrm{Sim}(S_m,S_n) \tag{5}$$

式(5)中，Senseof(c_1)和 Senseof(c_2)函数分别返回词语 c_1 和 c_2 的词义代码集合。若词语 c_1、c_2 存在多个义项，在计算词语相似度时，将义项分别两两计算，通过式(5)取最大值作为两个词语的相似度值。

基于语义相似度的词语对齐方法，可以弥补基于词典的词语对齐方法在覆盖面方面的不足，两者结合使用可以提高对齐的召回率。

2.3　基于统计的词对齐方法

在基于统计的词对齐方法方面，本文中使用了目前比较典型的工具 GIZA++。GIZA++是 GIZA 的一个扩展，是 Och[11]等人在 GIZA 软件包基础上进一步优化得到的统计机器翻译工具。GIZA++在实现了 IBM model 1-5 和 HMM(隐马尔科夫模型)基础上，对 IBM-1、IBM-2 和 HMM 模型的概率计算算法进行了改进。

运行 GIZA++相关命令，将普通文本转化为 GIZA++ 格式，生成～.A3.final 对齐文件，包含对齐概率、目标句子、源语言句子和对齐位置信息。例如，

Sentence pair (3128) source length 14 target length 10 alignmentscore：1.55964e-17

但是$_1$在$_2$投资$_3$领域$_4$不$_5$可能$_6$一直$_7$靠$_8$运气$_9$。$_{10}$

NULL ({ 2 })하지만 ({ 1 }) 투자 ({ 3 }) 의 ({ })세계 ({ }) 에서 ({ }) 는 ({ }) 요행수 ({ 4 6 7 8 9 }) 가 ({ }) 계속 ({ }) 통하 ({ }) 지 ({ }) 않 ({ 5 }) 는다 ({ }) . ({ 10 })

2.4　统计与词典相融合的词对齐方法

通过基于词典和基于统计的词对齐实验，可以看出完全基于词典的对齐可以获得可靠的非空对齐。但是由于双语词典的覆盖面有限，得到的对齐的召回率并不理想。基于统计的方法可以弥补纯词典方法的不足，获得更多对齐，因此可以将统计的方法作为初始对齐的方法，在此基础上，使用基于词典和基于语义相似度的方法进行词对齐校正。其主要步骤为：

(1) 通过 GIZA++工具，获取词对齐文件；

(2) 通过韩汉机读辞典，获取某一韩国语词语的译文；

(3) 将该译文与汉语句子中每个汉语词语进行词语相似度计算，取相似度值大于阈值结果中的最大值，将其对应汉语词语位置加入词对齐文件；

(4) 若不存在相似度值大于阈值的结果，对韩国语所对应汉语译文与汉语句子中所有词语进行语义相似度计算，取语义相似度值大于阈值结果中的最大值，将其对应汉语词语位置加入词对齐文件。

上例中经过统计方法得到的词对齐结果再通过基于词典和基于语义相似度的方法进行词对齐校正，得到校正后的对齐文件如下所示：

Sentence pair (3128) source length 14 tar-

get length 10 alignmentscore: 1.55964e-17

但是1在2投资3领域4不5可能6一直7靠8运气9。10

NULL ({ 2 })하지만 ({ 1 }) 투자 ({ 3 }) 의 ({ }) 세계 ({ 4 }) 에서 ({ 2 }) 는 ({ }) 요행수 ({ 9 }) 가 ({ }) 계속 ({ 7 }) 통하 ({ 8 }) 지 ({ }) 않 ({ 5 }) 는다 ({ }) . ({ 10 })

通过例句可以看出,在现有资源和语料规模的情况下,综合使用基于词典和基于统计的方法可以得到更好的对齐结果。

3 名词短语对齐方法

在实现词对齐的基础上,通过正则表达式抽取出韩国语名词短语,并利用词对齐时所获得的对齐位置信息获取候选名词短语翻译对,为了从候选名词短语翻译对中获取正确的翻译等价对,本文引入 χ^2 统计值的关联度度量方法,其基本原理是:假设在 N 个双语语料句对中,X 代表源语言某一词语,Y 代表目标语言中某一词语,则两者之间的分布情况可以用联列表来描述。

表 1 X 与 Y 的联列表

	Y	$-Y$
X	a	b
$-X$	c	d

表格中 a、b、c、d 的含义为:

a:双语语料所有句对中,短语 X 和 Y 同时出现的次数;

b:双语语料所有句对中,短语 X 出现但短语 Y 不出现的次数;

c:双语语料所有句对中,短语 X 不出现但短语 Y 出现的次数;

d:双语语料所有句对中,短语 X 和 Y 均不出现的次数;

于是可以得出 χ^2 统计值的计算如式(6)所示。

$$\chi^2(x,y)=\frac{n\times(ad-bc)^2}{(a+c)(a+b)(d+b)(d+c)} \quad (6)$$

χ^2 在[0,n]之间,它的值越大,表示两个词语之间相关程度越高。

名词短语对齐方法主要利用词对齐时所获得的对齐位置信息实现名词短语对齐,其主要步骤如下:

(1) 从韩国语标注语料中通过正则表达式抽取出韩国语名词短语;

(2) 根据抽取出的名词短语,获取词对齐文件中每个韩国语词语对应的汉语位置;

(3) 将获得的汉语位置序列,按照从小到大的顺序进行排序,按照排序顺序抽取出对应的汉语词语,获得候选名词短语翻译对;

(4) 通过 χ^2 统计值的方法,进行最优过滤,得到名词短语对。

4 实验结果及分析

基于上述方法,本文初步实现了一个原型系统,并针对基于词典和语义相似度的词对齐方法、基于统计的词对齐方法和基于统计和词典相融合的方法,初步进行了一些试验,测试不同词对齐方法对本文提出的基于词对齐位置信息的名词短语对齐结果的影响。

实验中使用的韩汉双语词典包含词条 50 357 条。语义词典使用《同义词词林》。经过句子对齐并用于统计训练的双语句对 112 475 对,来自韩国《朝鲜日报》、《中央日报》和《东亚日报》发布的各类新闻,内容涵盖韩国语的政治、经济、文化、科技等方面。该语料库在内容真实的基础上,具备韩国语新闻语料最普遍的语言特点,根据这些语料进行相应研究,得出的结论也能体现出韩国语新闻语料的一般性特征,因此选用新闻语料,可使研究结果更加客观真实。其中的汉语句子经过分词处理,韩国语句子经过分词和词性标注处理。从训练语料中随机抽取 300 句对中的名词短语并做人工校对,作为标准测试语料。

在实验结果的评价方面,目前最常用的两个指标分别是准确率和召回率[12],其中,准确率和召回率的定义如式(7)、式(8)所示。

$$\text{准确率}=\frac{\text{正确对齐的数量}}{\text{所有对齐的数量}}\times 100\% \quad (7)$$

$$\text{召回率}=\frac{\text{正确对齐的数量}}{\text{所有正确对齐的数量}}\times 100\% \quad (8)$$

表 2 给出了基于词典的词对齐方法、基于统计的词对齐方法和融合的词对齐方法下的名词短语对齐结果。

表 2 名词短语对齐结果

	正确对齐数量	所有对齐数量	人工找到的数量	准确率/%	召回率/%
统计的方法	412	552	583	74.6	70.7
词典的方法	336	384	583	87.5	57.6

续表

	正确对齐数量	所有对齐数量	人工找到的数量	准确率/%	召回率/%
融合的方法	440	553	583	79.6	75.5

从表 2 可以看出，基于词典的方法中，对齐具有较高的准确率，但由于词典的覆盖能力有限，因此召回率较低。而基于统计的方法，可以提高召回率，但准确率较低。在基于统计和词典相融合的方法中，在基于统计的方法基础上，利用基于词典的方法，结合了基于统计的方法和基于词典的方法的优点，既弥补了基于统计方法中准确性的不足，使得正确的对齐数增加，保证非空对齐的正确率，又可以克服基于词典的方法中词典覆盖能力有限的问题，使得对齐的召回率有了进一步的提高，在此方法下召回率和准确率也都达到了三个实验中较为均衡的值。

分析对齐中产生的错误，一部分原因是由于资源不足引起的（词典译文缺乏、统计数据不足等）。其他错误大部分是由于汉语和韩国语之间存在固有的表达差异造成的，如韩国语中的成语、惯用搭配等在相应的汉语中通常采用意译。本文提到词对齐方法尚不能解决好这类错误，对于这些错误，有待进一步增加句法分析和语言学知识加以解决。

5　结论

本文通过对基于三种不同词对齐方法的名词短语对齐结果进行实验分析，可以得到以下结论：

（1）语言学信息在双语语料库词对齐中有着重要作用。双语词典可以提供可靠的非空对齐。基于词典和语义相似度的方法可以提高对齐的正确率。

（2）当语料库规模较大时，基于统计的方法对提高对齐的召回率具有重要作用。

（3）在资源和语料不足的情况下，基于词典和基于统计相结合的方法是进行词对齐的有效方法。

尽管本文使用了多种对齐方法，但对齐的准确率与召回率仍然不能令人满意。一个主要原因是由于韩汉双语间的语言差异，使得很多对齐问题需要在句法层面上才能得以解决。

参考文献

[1] Marcu D, Wong W. A phrase-based, joint probability model for statistical machine translation[C]//Proceedings of Emnlp 2002, 2002: 133-139.

[2] Zhang Y, Brown R, Frederking R, et al. Pre-processing of bilingual corpora for Mandarin-English EBMT [C]//Proceedings of Mt Summit VIII, 2001.

[3] Zhang Y, Vogel S. Competitive grouping in integrated phrase segmentation and alignmentmodel [C]//Proceedings of the ACL Workshop on Building and Using Parallel Texts, 2005: 99-106.

[4] 常宝宝. 基于汉英双语语料库的翻译等价单位自动获取研究[J]. 产品安全与召回，2002(2)：24-29.

[5] 程洁，杜利民. EBMT 系统中的多词单元翻译词典获取研究[J]. 中文信息学报，2004，18(1)：55-61.

[6] 屈刚，陈笑蓉. 基于有效包型的英汉双语短语对齐[J]. 计算机研究与发展，2003(2)：143-149.

[7] 毕玉德. 朝鲜语自然语言处理研究管窥[J]. 中文信息学报，2011，25(06)：166-169，182.

[8] 安帅飞，毕玉德. 韩国语名词短语结构特征分析及自动提取[J]. 中文信息学报，2013，27(5)：205-210.

[9] 王斌. 汉英双语语料库自动对齐研究[D]. 北京：中国科学院计算机技术研究所博士学位论文，1999.

[10] 田久乐，赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报（信息科学版），2010，28(6)：602-608.

[11] Franz Josef Och, Hermann Ney. Giza++: Training of statistical translation models[Z]. 2000.

[12] 俞士汶. 计算语言学概论[M]. 北京：商务印书馆，2003：322-333.

凌天斌(1978—)，讲师，主要研究领域为自然语言处理。

E-mail：ltb512@163.com

毕玉德(1967—)，通信作者，教授、博士生导师，主要研究领域为自然语言处理和韩国语句法语义学。

E-mail：biyude2005@163.com