软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(7):1650−1661 [doi: 10.13328/j.cnki.jos.004615]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于多策略融合 Giza++的术语对齐法*

刘胜奇, 朱东华

(北京理工大学 管理与经济学院,北京 100081)

通讯作者: 刘胜奇, E-mail: shengqiliu@126.com, http://www.bit.edu.cn

摘 要: 跨语系术语对齐质量不高,原因在于其依赖于低质量的术语抽取与对齐.提出的多策略融合 Giza++(AGiza)的术语对齐法,为提高术语抽取质量,用首尾词性规则提高召回率,用独立过滤、停用过滤提高准确率,再识别共句术语对.为提高术语对齐的对准率:基于独立度、停用度,提出独立相关度、停用相关度;由种子对相关度和单词关联度概率加组合成语义相关度;根据首尾对齐情况,提出首尾相关度,并去除值为 0 者;基于词性组成特征,构造词性相似度;由 GIZA++计算得到 g 值;经过属性的相关系数分析后,乘法组合各属性构造术语对齐度 a;最后,过滤 a 超过术语对齐阈值(由召回率设定)的术语对.实验结果表明,AGiza 术语对齐,可有效地处理跨语系术语对齐,质量高于 GIZA++,Dice, $\Phi 2$,LLR,K-VEC 及 DKVEC.

关键词: 术语对齐;多语言术语抽取;跨语言;跨语系

中图法分类号: TP391

中文引用格式: 刘胜奇,朱东华.基于多策略融合 Giza++的术语对齐法.软件学报,2015,26(7):1650−1661. http://www.jos.org.cn/1000-9825/4615.htm

英文引用格式: Liu SQ, Zhu DH. Automatic term alignment based on advanced multi-strategy and Giza++ integration. Ruan Jian Xue Bao/Journal of Software, 2015,26(7):1650−1661 (in Chinese). http://www.jos.org.cn/1000-9825/4615.htm

Automatic Term Alignment Based on Advanced Multi-Strategy and Giza++ Integration

LIU Sheng-Qi, ZHU Dong-Hua

(School of Management & Economics, Beijing Institute of Technology, Beijing 100081, China)

Abstract: The quality of cross-phylum term alignment depends on the quality of term extraction and alignment method. This paper proposes an automatic term alignment based on advanced multi-strategy and Giza++ (AGiza) integration. By analyzing the properties of the term extraction performed by using some existing methodologies in the literature, the rules of the first and the last part of speech of strings are designed to increase the recall rate. Methods that are applied for the purpose of increasing the precision of the term extraction include: (1) independence filter; (2) stopping filter; and (3) recognition of the co-occurrence of terms in the sentence pairs. The following steps are also implemmented to increase the alignment quality: (1) design the degree of the independence correspondence based on the degree of independence; (2) construct the degree of the stopping correspondence based on the degree of stopping usage; (3) propose the degree of semantic correspondence that computed by the seed pairs' correspondence and word pairs' similarity based on additivity of probability; (4) construct the alignment correspondence degree of the first part and last part between the term pairs in order to cancel the term pairs whose value is equal to zero; (5) present the similarity degree of the part of speech between the term pairs considering the patterns that define the morphosyntactic structures of terms; and (6) obtain the value of g based on GIZA++. The term-aligned degree (a) is computed by the six attributes of term pairs based on multiplication of probability after analyzing their correlations. Term pairs is extracted by select the term-aligned pairs based on the candidate term pairs whose a is more than the term-aligned threshold that make the tolerance of recall is less than 1%. The simulation results of Chinese-English term alignment show that automatic term alignment based on AGiza can be used to extract cross-phylum term pairs effectively. Furthermore, it outperforms GIZA++, the Dice coefficient, the $\Phi 2$ coefficient, the log-likelihood ratio, K-VEC and DKVEC.

* 基金项目: 国防基础科学研究计划(Q172011A001)
收稿时间: 2013-11-03; 修改时间: 2014-01-10; 定稿时间: 2014-04-09

Key words: term alignment; multilingual term extraction; cross-language; cross-phylum

日新月异的科技,导致术语井喷.术语翻译直接影响翻译质量.人工准备术语翻译对,耗时费力,术语翻译成为难题,术语对齐应运而生.

目前的术语对齐,无论是先抽取多语术语再对齐[1,2],还是先抽取单语术语,然后在另一种语言中识别对应翻译[3],均认为术语抽取和对齐是独立过程,所以其质量依赖于术语抽取及对齐方法.

在术语抽取上,经典的 C/NC-value[4]、名词计分法[5]及其衍生方法,召回率不超过 85%[5],准确率不超过 76%[4].召回率不高,主要是因为所用词法规则为词性(part of speech,又叫词类,简称 POS)固定组合,覆盖度低,若仅考虑首尾词性规则,可使召回率大为提高.准确率不高的核心原因有二:一是不完整词串的干扰,二是停用词串(非术语)的误判.C/NC-value 中的子串修正[4],实质上涉及词串独立存在的概率,虽然显著减少了干扰,但忽略了中间最短母串(左右两边各加一词构成的词串)的影响,可用中间最短母串修正母串对子串的影响,提出词串独立度.名词计分法[5]实质上是计算词串在特定领域的停用情况,但仅考虑名词术语,忽略单名词非临近词影响和复合名词外其他词影响,方法有效却极端,可综合考虑各种词性词串的停用情况,提出词串停用度.

在对齐方法上,IBM 模型 1~4[6]、IBM 模型 2 的 LLR(log-likelihood ratio)[7]改进算法[8],中英文(跨语系)词对齐错误率大于 44%[8].对齐错误率高,主要原因在于部分对齐的影响.虽然 Och[9]认为可通过训练语料调整对齐参数,英法(同语系)词对齐的对准率超过 90%[9],但是该类模型无法很好地去除中英文(跨语系)中的部分对齐.统计短语翻译[10]通过实验发现,句法分析[11,12]并不能提高短语对齐的质量,反而降低性能,所以句法分析也不好消除部分对齐的影响.

为减少部分对齐,考虑独立度、停用度在术语对齐中的影响,提出独立相关度、停用相关度;借鉴 KNOWA (knowledge intensive word aligner)[13]基于双语辞典计算词语相似度,提出单词关联度、种子对相关度,组合提出语义相关度;斟酌实验中有助提高对准率的端对齐度[3],提出首尾相关度;考察有助于减少部分对齐的 Daille 关联度[1]、条纹记录规则[2],均涉及词性模式关联度,提出词性相似度.

综上,多策略融合 Giza++(advanced multi-strategies and Giza++ integration,简称 AGiza)术语对齐法首先进行高质量候选术语对的获取,再进行术语对齐.在候选术语对获取过程中,先基于首尾相关度进行高召回率候选术语抽取,通过独立过滤、停用过滤,删除异常候选术语,然后识别共句对的候选术语,构成候选术语对集.术语对齐过程中,结合术语对的独立相关度、停用相关度、语义相关度、首尾相关度、词性相似度及 $Giza^{++}$[14](基于 IBM 模型)计算的 g 值,相关系数分析后,构造术语对齐度 a,最后识别 a 超过阈值λ的术语对,作为对齐的术语对.

1 算法假设

AGiza 术语对齐法中的候选术语对获取,以术语库为基础,要求术语库具有代表性.同时,该方法不考虑文档的先后顺序、术语的先后顺序.AGiza 术语对齐法中的术语对齐,其前提是跨语系文档中的原术语有对应翻译的译术语.这样,AGiza 术语对齐法基于如下 4 个假设.

假设 1. 术语库为领域术语集中具有代表性的样本.假设 1 要求术语库不为无代表性的小样本,其质量必须达到较高水准.

假设 2. 语料中的文档满足可交换性,即,文档的先后顺序与内容无关.文档之间除了引用关系,很少存在其他关系,而引用关系为文档层次,对于术语属性的影响较小,所以假设 2 一般成立.

假设 3. 文档中的术语满足可交换性,即,术语出现的先后顺序与重要性无关.虽然在部分总分式结构文档中,首先出现的术语更重要、更能体现文档内容,但是这种文档数量不多,为简化计算,暂不考虑.

假设 4. 跨语系文档中的原术语一般有对应翻译的译术语.假设 4 要求原术语的翻译尽量不用代词指代,尽量不省略.

据此,AGiza 术语对齐法主要应用场景为原术语有对应译术语的句对,可广泛应用于跨语系的术语对齐中.

2 候选术语对获取

候选术语对获取,首先基于首尾词性规则进行候选术语抽取,再经过独立过滤、停用过滤,删除异常候选术语,最后识别共句对的候选术语,构成候选术语对集.

2.1 候选术语抽取

定义 1(词串). 词串是指文本经过词法分析器分词后构成的字符串 $s=w_1w_2\ldots w_n,w_i(i=1,2,\ldots,n)$表示分词后形成的单词.如“静态密封往复式燃油泵”经过分词后,词串 s=静态/n 密封/vn 往复/vd 式/k 燃/vg 油泵/n.

首尾词性规则基于术语库构建,一般为出现频率较高的首尾词性组合.考虑扩展性,需要统计术语库中术语的首词性(第 1 个单词词性)重复出现情况、尾词性(最后一个单词词性)重复出现情况.

基于首尾词性规则的候选术语抽取法如下:

① 用词法分析器对文档进行分词并标注词性;

② 根据句尾符对文档进行分句;

③ 根据首尾词性规则中首词性、尾词性重复出现的情况,组合构造词串词法规则,如 *POS*=[1 或多个](首词性+任意词性)+[0 或多个](任意词性+尾词性)+[0 或多个](首词性+任意词性)+[1 或多个](任意词性+尾词性)等;

④ 若抽取时使用的首尾词性规则中首尾词性相同,则在每个句子中抽取词性为首词性的一个单词作为候选术语;否则,在每个句子中抽取词性组合为 *POS* 的多个单词组合作为候选术语.

2.2 独立过滤

定义 2(词串独立度). 词串独立度是指词串 s 在语料中独立出现的概率,记为 id.

若语料中存在以下情况之一,则 $id(s)=1$:

① s 处于文档段首、右接标点符号;

② s 处于文档段尾、左接标点符号;

③ s 左右均接标点符号.

若出现 $f_L(s)+f_R(s)-f_M(s)\leqslant 0$ 或者 $f_L(s)+f_R(s)-f_M(s)>f(s)$时,$id(s)=1-\max(f_L(s),f_R(s),f_M(s))/f(s)$;其他情况计算方法如下:

$$id(s)=1-\frac{f_L(s)+f_R(s)-f_M(s)}{f(s)} \tag{1}$$

$f(s)$即词串 s 在语料中出现的次数,$f(s)$不小于 1;$f_L(s)$即左侧最短母串频率,是指词串 s 的左边加一词构成的词串在语料中出现的次数;$f_R(s)$即右侧最短母串频率,是指词串 s 的右边加一词构成的词串在语料中出现的次数; $f_M(s)$即中间最短母串频率,是指词串 s 的左右两边各加一词构成的词串在语料中出现的次数.$f_M(s)$和 $f_L(s)$,$f_R(s)$存在重算,根据集合并运算,考虑用 $f_L(s)+f_R(s)-f_M(s)$表示 s 的最短母串出现频率.

例如:词串 s=太阳能/n 电池/n,位于段首,右接标点符号.id(太阳能/n 电池/n)=1.

又如:词串 s=消音/v,在语料中出现次数 $f(s)$=4;左侧最短母串=油泵/n 消音/v,$f_L(s)$=2;右侧最短母串=消音/v 装置/n,$f_R(s)$=3;中间最短母串=油泵/n 消音/v 装置/n,$f_M(s)$=2,则 id(消音/v)=1−(2+3−2)/4=0.25.

词串独立度与以前计算词串独立出现频率方法的差异如下所示.

① C/NC-value 中的子串修正[4]、质子串[15],根据词串出现频率与母串出现频率均值之差计算词串独立出现的频率,尚未考虑最短母串,计算较粗略;

② 子串归并[16]根据左侧最短母串、右侧最短母串出现频率中的最大值计算词串独立出现的频率,一方面忽略了中间最短母串,另一方面计算仍较粗略;

③ 词串独立度引入概念“中间最短母串”,借用集合“并”的思想,去除“中间最短母串”重复出现频率,使得词串独立出现频率的计算更加科学.

候选术语也是词串,其独立过滤法如下所示:

① 对于候选术语集中的每个词串 x,分别获取最长子串:左侧最长子串为去掉首词后形成的词串,右侧最长子串为去掉尾词后形成的词串,中间最长子串为去掉首词、尾词后形成的词串;

② 统计候选术语集中所有词串的母串情况.考察不是 x 的其他词串 s:若 s 与步骤①构造的左侧最长子串相同,则左侧最短母串频率 $f_L(s)$增加 x 的出现频率 $f(x)$;若 s 与步骤①构造的右侧最长子串相同,则右侧最短母串频率 $f_R(s)$增加 x 的出现频率 $f(x)$;若 s 与步骤①构造的中间最长子串相同,则中间最短母串频率 $f_M(s)$增加 x 的出现频率 $f(x)$;

③ 由于词法分析器的词性标注错误,出现 $f_L(s)+f_R(s)-f_M(s)\leqslant 0$ 或者 $f_L(s)+f_R(s)-f_M(s)>f(s)$时,则 $id(s)=1-\max(f_L(s),f_R(s),f_M(s))/f(s)$;否则,$id(s)=1-(f_L(s)+f_R(s)-f_M(s))/f(s)$;

④ 若 $id(s)=0$,则从候选术语集中删除 s.

2.3 停用过滤

定义 3(词串停用度). 词串停用度用于衡量语料中一个词串不参与术语构成的程度.一般地,一个单词与其他单词结合的概率越大,该单词结合的词串构成术语的概率就越小.词串停用度如下计算:

$$st(s)=1+\frac{1}{|s|}\sum_{i=1}^{|s|}(L(w_i)+R(w_i)) \tag{2}$$

$|s|$为词串 s 所含单词数;w_i 为词串 s 中分词形成的第 i 个单词;$L(w_i)$表示语料中单词 w_i 左侧单词数;$R(w_i)$表示语料中单词 w_i 右侧单词数.为防止停用度为 0,故加 1.

例如:词串 s=太阳能/n 电池/n 由两个单词构成:太阳能/n,电池/n.在候选术语集中,“太阳能/n”左侧搭配“反射/v”、“硅/n”、“吸收/v”、“辐射/v”等 9 个单词,右侧搭配“板/ng”、“采暖/vn”、“电池/n”、“充电器/n”等 24 个单词,则 L(太阳能/n)=9,R(太阳能/n)=24.同理,“电池/n”左侧搭配“太阳能/n”、“备用/vn”、“极/ng”、“充气/vn”等 32 个单词,右侧搭配“电感/n”、“电力/n”、“动力/n”等 13 个单词,则 L(装置/n)=32,R(装置/n)=13.所以,

st(太阳能/n 电池/n)=1+(9+24+32+13)/2=40.

下面给出词串停用度与以前计算词串停用程度方法的对比.

① 名词计分法[5]只考虑名词.通过分析词串所含名词,在文档中直接连接的左侧单词数、右侧单词数来测度词串的重要性(实际上是一种停用程度);

② 词语活跃度[16]只考虑非名词.通过计算词串所含非名词性单词,在语料中直接连接的左侧单词数的熵、右侧单词数的熵来测度词串的活跃度(实际上也是一种停用程度);

③ 词串停用度同时考虑名词、非名词.通过统计词串所含单词,在语料中直接连接的左侧单词数、右侧单词数来测度词串的停用程度.

候选术语也是词串,停用过滤法如下:

① 对于候选术语集中每个词串 s,分别获取词串中每个单词 w_i 左、右侧搭配的单词.其中,若单词为词串的首词,则左侧搭配的单词为””;若单词为词串的尾词,则右侧搭配的单词为””;

② 分别统计单词 w_i 在语料中左侧搭配的单词数 $L(w_i)$、右侧搭配的单词数 $R(w_i)$;

③ 计算词串 s 的停用度 $st(s)=1+\frac{1}{|s|}\sum_{i=1}^{|s|}(L(w_i)+R(w_i))$;

④ 若 $st(s)>100$,则删除.

2.4 共句术语对

现记候选原术语个数为 m,候选译术语个数为 n,若每个候选原术语和候选译术语都进行术语对齐计算,则计算的复杂度为 $A(m\times n)$,系统开销较大.为降低运算复杂度,根据假设 4,只考虑共现于句对中的共现候选术语对.若记句对数为ς,共现句对 i 中的候选原术语个数为$|S_i|$,候选译术语个数为$|T_i|$,则计算复杂度将降为

$$F((m+n)\times s)+A\left(\sum(|S_i|\times|T_i|)\right)\approx A\left(\sum(|S_i|\times|T_i|)\right)=A(\varsigma\times(|S_i|\times|T_i|))<<A(m\times n).$$

因为查找句对中的共现候选术语对的时间消耗 $F(((m+n)\times s)$,相对于术语对齐计算来说很小,基本可忽略.

据此,在已对齐的句对中,识别候选原术语与共句对的候选译术语一起构成了候选术语对集.

3 术语对齐

在候选术语对获取的基础上,可开展术语对齐.首先计算术语对齐的属性——独立相关度、停用相关度、语义相关度、首尾相关度、词性相关度以及 Giza++[14](基于 IBM 模型)计算的 g 值,分析各属性之间的相关性,构造术语对齐度 a,最后识别 a 超过阈值λ的术语对,作为对齐的术语对.

3.1 独立相关度

定义 4(独立相关度). 独立相关度表示原术语与译术语在语料中独立度的相关程度,记为 i.术语对独立度的差异可表示独立相关度,基于概率加法计算如下:

$$\begin{aligned} i(T,S)=1-(id(S)-id(T))^2-(1-id(T))^2-(1-id(S))^2+(id(S)-id(T))^2\times(1-id(T))^2+\\ (id(S)-id(T))^2\times(1-id(S))^2+(1-id(S))^2\times(1-id(T))^2-\\ (id(S)-id(T))^2\times(1-id(T))^2\times(1-id(S))^2 \end{aligned} \tag{3}$$

$id(S)$表示原术语 S 的独立度,$id(T)$表示译术语 T 的独立度.

例如:译术语 T=solar energy battery,原术语 S=太阳能电池.原术语 S 的独立度 $id(S)$=1,译术语 T 的独立度 $id(T)$=1,则 $i(T,S)$=1−(1−1)^2−(1−1)^2−(1−1)^2+(1−1)^2×(1−1)^2+(1−1)^2×(1−1)^2+(1−1)^2×(1−1)^2−(1−1)^2×(1−1)^2×(1−1)^2=1.

3.2 停用相关度

定义 5(停用相关度). 停用相关度表示原术语与译术语在语料中停用度的相关程度.术语对停用度的差异可表示停用相关度,基于概率加法计算如下:

$$\begin{aligned} s(T,S)=1-(st^{-1}(S)-st^{-1}(T))^2-(1-st^{-1}(T))^2-(1-st^{-1}(S))^2+(st^{-1}(S)-st^{-1}(T))^2\times(1-st^{-1}(T))^2+\\ (st^{-1}(S)-st^{-1}(T))^2\times(1-st^{-1}(S))^2+(1-st^{-1}(S))^2\times(1-st^{-1}(T))^2-\\ (st^{-1}(S)-st^{-1}(T))^2\times(1-st^{-1}(T))^2\times(1-st^{-1}(S))^2 \end{aligned} \tag{4}$$

$st(S)$表示原术语 S 的停用度,$st(T)$表示译术语 T 的停用度.

例如:译术语 T=solar energy battery,原术语 S=太阳能电池.原术语 S 的停用度 $st(S)$=40,译术语 T 的停用度 $st(T)$=48,则:

$$\begin{aligned} s(T,S)=1-(1/40-1/48)^2-(1-1/40)^2-(1-1/48)^2+(1/40-1/48)^2\times(1-1/40)^2+(1/40-1/48)^2\times(1-1/48)^2+\\ (1-1/40)^2\times(1-1/48)^2-(1/40-1/48)^2\times(1-1/40)^2\times(1-1/48)^2\approx 0.002036. \end{aligned}$$

3.3 语义相关度

定义 6(语义相关度). 语义相关度表示原术语与译术语在特定领域中语义内容相关的程度,记为 r.建立以术语对、种子对为基础的计算公式如下:

$$r(T,S)=sr(T,S)+wc(T,S)-sr(T,S)\cdot wc(T,S) \tag{5}$$

$sr(T,S)$表示原术语 S 和译术语 T 的种子对相关度,$wc(T,S)$表示原术语 S 和译术语 T 的单词关联度.

定义 7(种子对相关度). 种子对相关度表示在特定领域中原术语所含种子词,与译术语所含种子词的相关程度.计算如下:

$$sr(T,S)=\frac{\sum_{m=1}^{M}|SEED_m(T)|}{|T|}\cdot\frac{\left(\sum_{m=1}^{M}sc(SEED_m(T),SEED_m(S))\right)}{M} \tag{6}$$

M 表示原术语 S 和译术语 T 中有 M 个种子对,$|SEED_m(T)|$表示译种子 $SEED_m(T)$的字符数,$|T|$表示译术语 T 的字符数,$sc(SEED_m(T),SEED_m(S))$表示原术语 S 和译术语 T 中第 m 个种子对的相关度.

$$sc(SEED_m(T),SEED_m(S))=\frac{1}{\xi_m} \tag{7}$$

ξ_m 表示种子对库中译种子 $SEED_m(T)$对应的原种子数.

定义 8(单词关联度). 单词关联度表示原术语与译术语包含的单词对的相关程度.计算为

$$wc(T,S)=\frac{\sum_{i=1}^{N}|w_{T_i}|}{|T|}\cdot\frac{\left(\sum_{i=1}^{N}c(w_{T_i},w_{S_i})\right)}{N} \tag{8}$$

N 表示原术语 S 和译术语 T 中单词对的个数,$|w_{T_i}|$表示译词 w_{T_i} 的字符数,$|T|$表示译术语 T 的字符数,$c(w_{T_i},w_{S_i})$表示第 i 个单词对的译词 w_{T_i}、原词 w_{S_i} 的单词关联度.

$$c(w_{T_i},w_{S_i})=\frac{1}{\omega_i} \tag{9}$$

ω_i 表示词库中译词 w_{T_i} 对应的原词数.

例如:译术语 T=solar energy battery,原术语 S=太阳能电池.该例含有 2 个种子对:solar、太阳,energy、能,则种子对相关度为

$$sr(\text{solar energy battery,太阳能电池})=\frac{\sum_{m=1}^{2}|SEED_m(T)|}{|T|}\cdot\frac{\left(\sum_{m=1}^{2}sc(SEED_m(T),SEED_m(S))\right)}{2}$$
$$=(11/18)\times(sc(\text{solar,太阳})+sc(\text{energy,能}))/2$$
$$=(11/18)\times((1/2)+(1/1))/2$$
$$\approx 0.458333.$$

该例含有 2 个单词对:solar energy、太阳能,battery、电池,则单词关联度为

$$wc(\text{solar energy battery,太阳能电池})=\frac{\sum_{i=1}^{2}|w_{T_i}|}{|T|}\cdot\frac{\left(\sum_{i=1}^{2}c(w_{T_i},w_{S_i})\right)}{2}$$
$$=(18/18)\times(c(\text{solar energy,太阳能})+c(\text{battery,电池}))/2$$
$$=(18/18)\times(1/1+1/3)/2$$
$$\approx 0.666667.$$

如此,语义相关度为

$$r(\text{solar energy battery,太阳能电池})=sr(\text{solar energy battery,太阳能电池})+wc(\text{solar energy battery,太阳能电池})-$$
$$sr(\text{solar energy battery,太阳能电池})\times wc(\text{solar energy battery,太阳能电池})$$
$$\approx 0.458333+0.666667-0.458333\times 0.666667$$
$$\approx 0.819445.$$

下面给出语义相关度与以前相似思想的对齐法的比较情况.

① 语义相关度中的单词关联度、IBM 模型中的繁殖概率[6]均考虑将多个原词翻译成一个译词的情况.不同之处在于:繁殖概率认为,译词在译串中遵循二项分布;单词关联度假设译词在译术语中服从 Zipf 分布,而译词在译术语中出现的频率很低,大部分不超过 1,所以可简化为均值计算;

② KNOWA[13]直接用双语辞典中的词汇计算.语义相关度中的单词关联度也基于跨语言辞典计算,不过认为词对翻译的概率不一定为 1,需要根据译词对应的原词数计算,更能体现一词多义;

③ Daille 关联度[1]的单词对齐分由候选术语对中所含单词对的统计关联度求和而得,导致低关联度单词对越多,单词对齐分越高,从而影响对齐结果.语义相关度中的单词关联度一方面通过概率加组合术语对中所含单词的关联度,另一方面还用译词字符数在译术语中所占字符数比例折合而成,这样,随着低关联度单词对的增加,单词关联度相应减少,更符合实际;

④ 关联分[17]计算为候选译术语所含译词翻译成原词概率均值之积,但译词越多,关联分越小,不利于长术语对齐.语义相关度中的单词关联度取译词翻译概率之均值,与译词个数无关,对长术语对齐不产生影响;

⑤ 短语译文直译率[18]用短语对中单词对所占比例来估算,计算基础为单词数,取决于词法分析器,若切分较细,则短语译文直译率低.单词关联度与词法分析器无关,不受单词切分影响,更客观.

3.4 首尾相关度

定义 9(首尾相关度). 首尾相关度表示术语对中开头片段、结尾片段的相关程度,记为 m.对于词之间无分隔符的汉藏等语系,开头片段表示术语开头部分的单字,结尾片段表示术语结尾部分的单字.对于词之间有分隔符的印欧等语系,开头片段表示术语开头部分的单词,结尾片段表示术语结尾部分的单词.

$$m(T,S)=\frac{h(T|S)}{h(S)}+\frac{t(T|S)}{t(S)}-\frac{h(T|S)}{h(S)}\cdot\frac{t(T|S)}{t(S)} \tag{10}$$

$h(T|S)$表示跨语言术语库中与原术语 S、译术语 T 相同开头片段的术语对数,$h(S)$表示跨语言术语库中与原术语 S 相同开头片段的原术语数,$t(T|S)$表示跨语言术语库中与原术语 S、译术语 T 相同结尾片段的术语对数,$t(S)$表示跨语言术语库中与原术语 S 相同结尾片段的原术语数.实际对齐中,用词之间无分割符的语言为原语言,有利于充分利用首尾信息.

例如:译术语 T=solar energy battery,原术语 S=太阳能电池.原术语 S 词之间无分隔符,开头片段为“太”、结尾片段为“池”;译术语 T 词之间有分隔符,开头片段为“solar”、结尾片段为“battery”.术语库中开头片段为“太”、“solar”的术语对有 $h(T|S)$=85 对,开头片段为“太”的原术语有 $h(S)$=12648 条;结尾片段为“池”、“battery”的术语对有 $t(T|S)$=5848 对,结尾片段为“池”的译术语有 $t(S)$=19318 条,则首尾相关度为

m(solar energy battery,太阳能电池)=85/12648+5848/19318−(85/12648)×(5848/19318)≈0.307409.

下面给出首尾相关度与以前位置模型的区别.

① 隐喻短语[19]识别中,以原喻词结尾的短语构成训练集来获取区分词,可扩展到原喻词开头的短语.首尾相关度考虑开头、结尾的字/词对齐情况;

② 端对齐度[3],即,已有翻译对中与候选术语对有相同原词开头(结尾)和相同译词开头(结尾)的翻译对数量中较小值,忽略相同原词开头(结尾)的对应原/译术语数影响;而首尾相关度则考虑了相同原词开头(结尾)的对应原/译术语数影响,从而提高了精度.

3.5 词性相似度

定义 10(词性相似度). 词性相似度用于描述原术语和译术语在词性组成特征上的相似度,记为 p.换句话说,表示原术语词性模式翻译成译术语词性模式的概率.

$$p(T,S)=\frac{f(P_S,P_T)}{f(P_T)} \tag{11}$$

$f(P_T)$表示跨语言术语库中译术语词性模式的出现次数,$f(P_S,P_T)$表示跨语言术语库中原术语词性模式与译术语词性模式的共现次数.

例如:译术语 T=solar energy battery,T 的词性模式为/*JJ*+/*NN*+/*NN*,/*JJ*+/*NN*+/*NN* 在术语库中出现 40 110 次.原术语 S=太阳能电池,S 的词性模式为/*n*+/*n*.译术语 T 的词性模式/*JJ*+/*NN*+/*NN*,与原术语 S 的词性模式/*n*+/*n*,在术语库中共现 5 595 次,则词性相关度为

p(solar energy battery,太阳能电池)=5595/40110≈0.139491.

下面给出词性相似度与以前类似对齐法的区别.

① KNOWA[13],首先先对齐辞典中一对一的词对,作为中心词对;然后,根据双语辞典中单词的译词词性情况,选择词对词性相同部分越多、且在句对中距中心词对(前后 14 个单词内)最靠近的作为候选对齐的词对.若 KNOWA 用于术语对齐,则位置的考虑无疑会降低召回率.词性相似度不考虑位置信息,有助于提高召回率;

② Daille 关联度[1]的词性模式主要基于名词术语的词性模式计算,无疑会遗漏非名词术语的词性模式,比如动词术语.词性相似度则考虑各种词性术语的词性模式,从而提高召回率;

③ 条纹记录规则[2]阐释了英法术语对齐中的翻转现象,如 N Adj 可对齐 Adj N,N N,Adj Adj;但是 N Adj 也可能翻译成其他模式,其概率未见考虑.词性相似度则综合考虑各种词性模式的对齐概率;

④ 区间扭曲模型[20]也可计算词性模式的翻译概率,其差异在于区间扭曲模型由各个单词词性翻译概率

组合,而词性相关度基于术语库词性模式统计获得.

3.6 术语对齐度

定义 11(术语对齐度). 术语对齐度表示原术语与译术语对齐的程度,记为 a.

术语对齐度 a 由术语对的属性——独立相关度、停用相关度、语义相关度、首尾相关度、词性相关度以及 Giza^{++}计算的 g 值组合而成,需要首先确定术语对各属性之间的独立性.

由于 Giza++主要设计用于词对齐,不能直接用于术语对齐,需要对句对进行处理后才能对齐获得 g 值.在词之间无分隔符的汉藏等语系的句子中,首先导入带词性标记的抽取术语作为用户辞典,然后再用词法分析器完成单词切分.与此同时,在词之间有分隔符的印欧等语系的句子中,用"||"替换抽取的术语中的空格,这样可让 Giza++将抽取的术语作为一个单词处理.当然,完成 g 值计算后,还需要将"||"替换为空格,以保证可读性.

下面在跨语言术语库、训练语料中计算相关系数,见表 1.

Table 1 Correlations of term pairs' attributes

表 1 术语对属性的相关系数

	i	s	r	g	m	p	m^7
i	1						
s	10^{-15}	1					
r	10^{-15}	–0.210 5	1				
g	-10^{-16}	0.032 3	0.021 8	1			
m	10^{-16}	–0.255 8	0.174 1	0.049 7	1		
p	-10^{-15}	0.087 2	–0.063	0.044 3	–0.094	1	
m^7	10^{-17}	–0.115 5	0.059 2	0.053 0	0.601 3	–0.005 7	1

从表 1 的相关系数可以看出,术语对各属性之间的相关性较弱.若此,术语对各属性可认为基本独立,所以考虑采用乘法法则进行组合分析.

另外,术语对的属性值还表明:

① 语义相关度、首尾相关度、词性相关度以及 Giza^{++}计算的 g 值均存在 0 的可能,为防止 0 的极端影响,加 0.000001;

② 首尾相关度 m 经过幂变换后,虽与 Giza^{++}计算的 g 值关系稍微增强,但与其他属性的关系更弱,组合效果更佳.所以用$(m+0.000001)^{\upsilon}$代替.

这样,术语对齐度 a 计算为

$$a(S,T)=i\cdot s\cdot(r+0.000001)\cdot(g+0.000001)\cdot(m+0.000001)^{\upsilon}\cdot(p+0.000001) \tag{12}$$

υ为幂参数,取值如下:

$$\upsilon=\begin{cases}1, & m\leqslant 1\\ 7, & m>1\end{cases} \tag{13}$$

下面给出术语对齐度与以前类似方法的异同.

① 术语关联度[21]将共现模型计算的相关度乘以候选术语的最小术语值,计算只与候选术语值中较小者相关,与另一术语值无关;术语对齐度通过相关系数考证,有效组合各种属性,质量有保证;

② IBM 模型[6]中,模型 1 采用乘法法则组合串长概率、翻译概率,模型 2 增加变形概率,模型 3 增加繁殖概率,模型 4、模型 5 引入类划分优化;IBM 模型发展的 HMM 模型[9,22,23]、Giza^{++}模型[14],质量有不同程度的提高;术语对齐度在 Giza++计算结果的基础上进一步融合其他属性,考虑得更全面;

③ 双语同义术语识别法[24,25]从已对齐的专利句对中收集包含日语术语、英语术语的句对构建候选同义术语对集,且仅保留共现频率在 6~800 之间的术语对,方法简单,但召回率低;术语对齐度考虑各种共现情况,有助于召回率的提高;

④ 对偶分解[26]本质上是基于多模型参数迭代来求最优解,若用于术语对齐,每次迭代需要求解多个模型,时间开销大,质量取决于选择的模型;术语对齐度值确定,没有迭代,没有多模型产生的额外时间开

销,效率更高,质量由术语对齐度值本身决定;

⑤ 融合字符长度的翻译概率[27]基于互信息改进,融合了句子字符长度信息、共现频率、词频,忽视原词、译词均不出现的情况;术语对齐度组合各种属性,充分考虑术语对的出现情况,质量有保证.

3.7 AGiza对齐

综上,AGiza 对齐法如下:

① 候选术语对获取.识别原术语与共句对的译术语一起构成候选术语对集;

② 属性参数估算.根据定义及术语库,分别估算术语对齐属性相关的参数;

③ 计算候选术语对的首尾相关度 m.为降低计算复杂度,从候选术语对集去除 m=0 的候选术语对;

④ 记候选术语对集中当前术语对为 T_i,S_i,根据定义计算术语对齐度 $a(T_i,S_i)$.若 $a(T_i,S_i)>\lambda$(术语对齐阈值),则 T_i,S_i 为对齐的术语对;

⑤ λ(术语对齐阈值)由召回率设定.先设λ=0,若召回率低,则减小λ;若召回率高,则增大λ,直到召回率与设定的误差在 1%内,此时的λ对应召回率.

4 实证分析

4.1 平行语料准备

为完成实验,主要准备了 3 类平行语料.

① 对齐语料.从互联网中获取中英文专利摘要 9 358 个,经过人工筛选后约 8 710 个,采用句对齐器对齐后获得 25 060 句对作为对齐语料;

② 训练语料.从互联网中获得 10 000 个已对齐的中英文句对作为训练语料;

③ 术语库.从互联网中收集中英文术语 420 186 对,以此为基础建成跨语言术语库.

4.2 候选术语对获取

- 候选术语抽取.基于首尾词性获得 2 323 168 条中文候选术语,2 010 065 条英文候选术语;
- 独立过滤后,剩余 1 191 401 条中文候选术语,1 185 130 条英文候选术语;
- 停用过滤后,剩余 862 825 条中文候选术语,877 309 条英文候选术语;
- 识别中文术语与共句对的英文术语,共计 3 195 768 对术语.这些术语对可能在 25 060 句对中重复出现,去重后剩余 665 067 对候选术语,构成候选术语对集.

4.3 属性参数估算

① 独立相关度参数估算.在候选术语集中,分别计算每个候选术语的独立度;

② 停用相关度参数估算.在候选术语集、训练语料中,分别计算每个候选术语的停用度;

③ 语义相关度参数估算.在术语库中,分析英文种子对应的中文种子数、英文词对应的中文词数;

④ 首尾相关度参数估算.在术语库中,分别获取中英文术语的开头片段、结尾片段,从而获得中文术语 S、英文术语 T 对应的 $h(T|S),t(T|S),h(S),t(S)$;

⑤ 词性相似度参数估算.在术语库中,分析英文术语词性模式的出现次数、中文术语词性模式与英文术语词性模式的共现次数.

4.4 术语对齐

根据 AGiza 对齐法,实验过程如下:

① 计算候选术语对的首尾相关度 m 后,从候选术语对集中去除 m=0 的候选术语对,剩余 249 310 对术语;

② 计算术语对齐度 $a(T_i,S_i)$,取值范围为 8.91E−25~0.002959.

4.5 术语对齐法评价

为便于评价 AGiza 术语对齐法的有效性,采用以下评价方法:

- 对准率=模型正确对齐的术语对数/模型对齐的术语对数;
- 召回率=模型正确对齐的术语对数/语料中的实际术语对数.

这里,对齐语料中的实际术语对为 24 664 对.下面在候选术语对集中分别采用 Giza++、Dice 系数[28]、$\Phi 2$ 系数[29]、K-vec[30]、DKvec[31]、LLR、AGiza 进行术语对齐.对比实验结果见表 2.

Table 2 Comparative evaluation results of automatic term alignment

表 2 术语对齐对比评价结果

方法	P0.01 (%)	P0.1 (%)	P0.166 (%)	P0.25 (%)	P0.50 (%)	P0.75 (%)
Giza++	61.8	76.4	63	–	–	–
Dice	14.2	18.6	15.8	16.8	17.1	16.4
$\Phi 2$	69.6	26.1	32.6	26.8	23.8	22.4
K-VEC	14.2	12.4	11.8	12.1	11.8	11.6
DKVEC	14.2	21.1	18.2	18.4	17.8	17.1
LLR	89.1	64.6	59.9	53.4	41.2	35.9
AGiza	96	85.8	70.2	60.8	53.6	49.8
最佳方法	AGiza	AGiza	AGiza	AGiza	AGiza	AGiza

由表 2 可知:

① Giza++质量最稳定,对准率变化波动不太大,然而召回率不超过 17%.这是因为多个极值点可能引起对齐误差所致;

② Dice,$\Phi 2$,K-VEC,DKVEC 都能获得各种召回率,然而对齐效果均不好,对准率太低.原因在于:Dice 系数考察原术语、译术语在句对中的 3 种出现情况,忽视了原术语、译术语均不出现的句对情况;$\Phi 2$ 系数、K-vec、DKvec 虽考察原术语、译术语在句对中的 4 种出现情况,但公式组合不够有效;

③ LLR 用于术语对齐,也考虑原术语、译术语在句对中的 4 种出现情况,虽然能获得各种召回率,但是仅在 1%召回率时对准率优于 Giza++,随着召回率的增加,对准率下降得太快;

④ AGiza 在各种召回率时,对准率都好于 Giza++,Dice,$\Phi 2$,LLR,K-VEC 及 DKVEC,因为 AGiza 充分利用了更多的术语对属性.

5 结束语

本文提出的术语对齐方法基于多策略提高质量,在候选术语对获取中,采用首尾词性规则提高召回率,通过独立过滤、停用过滤提高准确率,通过识别共句术语对减少运算复杂度.在 AGiza 术语对齐中,考虑独立度、停用度在术语对齐中的影响,提出独立相关度、停用相关度;由单词关联度、种子对相关度,组合提出语义相关度;根据术语对首尾片段对齐情况,提出首尾相关度,并去除首尾相关度为 0 的候选术语对;基于词性组成特征,计算词性相似度;由 Giza++计算得到 g 值.经过属性的相关系数分析后,乘法组合各属性构造术语对齐度 a;最后,通过召回率设定 λ(术语对齐阈值),过滤 $a>\lambda$ 获得对齐的术语对.中英专利摘要术语对齐实验结果表明,AGiza 术语对齐法较为有效:在各种召回率时,AGiza 对准率都好于 Giza++,Dice,$\Phi 2$,LLR,K-VEC 及 DKVEC,原因在于 AGiza 充分利用了更多的术语对属性.

致谢 感谢审稿人和编辑提出的宝贵意见.

References:

[1] Daille B, Ganssier É, Langé JM. Towards automatic extraction of monolingual and bilingual terminology. In: Proc. of the 15th Conf. on Computational Linguistics. Kyoto: Association for Computational Linguistics, 1994. 515−521. [doi: 10.3115/991886.991975]

[2] Morin E, Daille B. Compositionality and lexical alignment of multi-word terms. Language Resources and Evaluation, 2010,44(1-2): 79−95. [doi: 10.1007/s10579-009-9098-8]

[3] Zhang J, Cao CG, Wang S. Web-Based term translation extraction and verification method. Computer Science, 2012,39(7):170−174 (in Chinese with English abstract). [doi: 10.3969/j.issn.1002-137X.2012.07.038]

[4] Frantzi KT, Ananiadou S, Mima H. Automatic recognition of multi-word terms: The c-value/nc-value method. Int'l Journal on Digital Libraries, 2000,3(2):115−130. [doi: 10.1007/s007999900023]

[5] Nakagawa H. Automatic term recognition based on statistics of compound nouns. Journal of Terminology, 2003,6(2):195−210. [doi: 10.1075/term.6.1.05nak]

[6] Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 1993,19(2):263−311.

[7] Dunning T. Accurate method for the statistics of surprise and coincidence. Computational Linguistics, 1993,19:61−74.

[8] Dyer C, Chahuneau V, Smith NA. A simple, fast, and effective reparameterization of IBM model 2. In: Proc. of the Human Language Technology and North American Association for Computational Linguistics Conf. (HLT-NAACL). Atlanta, 2013. 644−648. http://anthology.aclweb.org/N/N13/N13-1073.pdf

[9] Och FJ, Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003,29(l):19−52. [doi: 10.1162/089120103321337421]

[10] Koehn P, Och FJ, Marcu D. Statistical phrase-based translation. In: Proc. of the Human Language Technology and North American Association for Computational Linguistics Conf. (HLT-NAACL). Edmonton, 2003. 127−133. [doi: 10.3115/1073445.1073462]

[11] Liu SJ. Multi-Task learning for word alignment and dependency parsing. Artificial Intelligence and Computational Intelligence, 2011,7004:151−158. [doi: 10.1007/978-3-642-23896-3_18]

[12] Carpuat M, Marton Y, Habash N. Improved Arabic-to-English statistical machine translation by reordering post-verbal subjects for word alignment. Machine Translation, 2012,26(1-2):105−120. [doi: 10.1007/s10590-011-9112-y]

[13] Pianta E, Bentivogli L. Knowledge intensive word alignment with KNOWA. In: Proc. of the 20th Int'l Conf. on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2004. 1086−1092.

[14] Och FJ. Giza++: Training of statistical translation models. 2001. http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

[15] He TT, Zhang Y. Automatic Chinese term extraction based on decomposition of prime string. Computer Engineering, 2006,32(23): 188−190 (in Chinese with English abstract).

[16] Zhou L, Shi SM, Feng C, Huang HY. A Chinese term extraction system based on multi-strategies integration. Journal of the China Society for Scientific and Technical Information, 2010,29(3):460−467 (in Chinese with English abstract). [doi: 10.3772/j.issn.1000-0135.2010.03.011]

[17] Déjean H, Gaussier E, Renders JM, Sadat F. Automatic processing of multilingual medical terminology: Applications to thesaurus enrichment and cross-language information retrieval. Artificial Intelligence Medicine, 2005,33(2):111−124. [doi: 10.1016/j.artmed.2004.07.015]

[18] Zhang CX, Li S, Zhao TJ. Phrase alignment based on head-phrase extending. Computer Research and Development, 2006,43(9): 1658−1665 (in Chinese with English abstract). [doi: 10.1360/crad20060925]

[19] Fu JH, Cao CG, Wang S. Approach to recognizing Chinese metaphorical phrases based on distinction words. Computer Science, 2010,37(10):193−196 (in Chinese with English abstract).

[20] Zhang T, Yu ZT, Guo JY, Cao XB. A bilingual word alignment algorithm of Naxi-Chinese based on feature constraint models. Journal of Xi'an Jiaotong University, 2011,45(10):48−53 (in Chinese with English abstract).

[21] Zhang CZ, Wu D. Bilingual terminology extraction using multi-level termhood. Electronic Library, 2012,30(2):295−308. [doi: 10.1108/02640471211221395].

[22] Och FJ, Ney H. A comparison of alignment models for statistical machine translation. In: Proc. of the 18th Int'l Conf. on Computational Linguistics (COLING 2000). Saarbrücken: Association for Computational Linguistics, 2000. 1086−1090. [doi: 10.3115/992730.992810]

[23] Graça JV, Ganchev K, Taskar B. Learning tractable word alignment models with complex constraints. Association for Computational Linguistics, 2010,36(3):481−504. [doi: 10.1162/coli_a_00007]

[24] Liang B, Utsuro T, Yamamoto M. Identifying bilingual synonymous technical terms from phrase tables and parallel patent sentences. Procedia—Social and Behavioral Sciences, 2011,27:50−60. [doi: 10.1016/j.sbspro.2011.10.582]

[25] Liang B, Utsuro T, Yamamoto M. Semi-Automatic identification of bilingual synonymous technical terms from phrase tables and parallel patent sentences. In: Proc. of the 25th Pacific Asia Conf. on Language, Information and Computation (PACLIC 25), Information and Computation. 2011. 196−205. http://anthology.aclweb.org/Y/Y11/Y11-1021.pdf

[26] Shen SQ, Liu Y, Sun MS. Search for discriminative word alignment via dual decomposition. Journal of Chinese Information Processing, 2013,27(4):9−15 (in Chinese with English abstract).

[27] Sun L, Jin YB, Du L, Sun YF. Automatic extraction of bilingual term lexicon from parallel corpora. Journal of Chinese Information Processing, 2000,14(6):33−39 (in Chinese with English abstract).

[28] Dice LR. Measures of the amount of ecologic association between species. Ecology, 1945,26:297−302. [doi: 10.2307/1932409]

[29] Gale WA, Church KW. Identifying word correspondences in parallel texts. In: Proc. of the 4th DARPA Workshop on Speech and Natural Language. 1991. 152−157. http://www.aclweb.org/anthology/H/H91/H91-1026.pdf

[30] Fung P, Church K. K-vec: A new approach for aligning parallel texts. In: Proc. of the 15th Int'l Conf. on Computational Linguistics. Kyoto: Association for Computational Linguistics, 1994. 1096−1102. [doi: 10.3115/991250.991328]

[31] Fung P. A statistical view on bilingual lexicon extraction: From parallel corpora to nonparallel corpora. In: Proc. of the 3rd Conf. of the Association for Machine Translation in the Americas (AMTA'98). Heidelberg, Berlin: Springer-Verlag, 1998. 1−16. [doi: 10.1007/3-540-49478-2_1]

附中文参考文献:

[3] 张晶,曹存根,王石.一种基于 Web 的术语翻译获取及验证方法.计算机科学,2012,39(7):170−174. [doi: 10.3969/j.issn.1002-137X.2012.07.038]

[15] 何婷婷,张勇.基于质子串分解的中文术语自动抽取.计算机工程,2006,32(23):188−190.

[16] 周浪,史树敏,冯冲,黄河燕.基于多策略融合的中文术语抽取方法.情报学报,2010,29(3):460−467. [doi: 10.3772/j.issn.1000-0135.2010.03.011]

[18] 张春祥,李生,赵铁军.基于中心语块扩展的短语对齐.计算机研究与发展,2006,43(9):1658−1665. [doi: 10.1360/crad20060925]

[19] 符建辉,曹存根,王石.基于区分词的汉语隐喻短语识别.计算机科学,2010,37(10):193−196.

[20] 张涛,余正涛,郭剑毅,曹先彬.融合特征约束模型的纳西-汉语双语词语对齐算法.西安交通大学学报,2011,45(10):48−53.

[26] 沈世奇,刘洋,孙茂松.基于对偶分解的词语对齐搜索算法.中文信息学报,2013,27(4):9−15.

[27] 孙乐,金友兵,杜林,孙玉芳.平行语料库中双语术语词典的自动抽取.中文信息学报,2000,14(6):33−39.

刘胜奇(1978－),男,四川广安人,博士生,主要研究领域为科技数据挖掘,机器翻译,技术创新管理.

朱东华(1963－),男,教授,博士生导师,主要研究领域为技术监测,数据挖掘,技术创新管理,技术评估,两化融合.