中英命名实体识别及对齐中的中文分词优化

尹存燕,黄书剑,戴新宇,陈家骏
(南京大学计算机软件新技术国家重点实验室,江苏南京 210023;南京大学计算机科学与技术系,江苏南京 210023)

摘 要: 中文分词结果对中英命名实体识别及对齐有着直接的影响,本文提出了一种命名实体识别及对齐中的中文分词优化方法.该方法利用实体词汇的对齐信息,首先修正命名实体识别结果,然后根据实体对齐结果调整分词粒度、修正错误分词.分词优化后的结果使得双语命名实体尽可能多地实现一一对应,进而提高中英命名实体翻译抽取和统计机器翻译的效果.实验结果表明了本文优化方法的有效性.

关键词: 分词;命名实体识别;双语对齐;机器翻译
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2015)08-1481-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.08.003

Optimization of Chinese Word Segmentation in Named Entity Recognition and Word Alignment

YIN Cun-yan, HUANG Shu-jian, DAI Xin-yu, CHEN Jia-jun
(*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu* 210023, *China*;
Department of Computer Science and Technology, Nanjing University, Nanjing, Jiangsu 210023, *China*)

Abstract: Bilingual named entity recognition and alignment are important for many natural language processing. Named entity translation can improve a lot the performance of the system like statistical machine translation or cross-language information retrieval. Quality of Chinese word segmentation does have a big impact over named entity (NE) recognition and bilingual NE extraction. Bilingual alignment information provides indications for NE recognition and word segmentation. Accordingly, based on the characteristics of NE recognition, NE alignment, and word segmentation, this paper proposes an optimization algorithm of Chinese word segmentation. By correcting word segmentation error and adjusting word segmentation granularity, the optimization algorithm can enhance extraction effect of Chinese-English NE translation and performance of statistical machine translation. The experimental result on Chinese-English news corpus shows the efficiency of our algorithm.

Key words: word segmentation; named-entity recognition; alignment; machine translation

1 引言

在自然语言中,命名实体(人名、地名、组织机构名等)传递着重要信息,命名实体的识别是自然语言处理中的一项重要工作.对于跨语言的自然语言处理应用而言,除了命名实体的识别,命名实体的翻译对于提升机器翻译质量、跨语言信息检索效果等也有着重要的作用.命名实体随着时代变化,层出不穷,因此,大多数的命名实体都属于词典的未登录词(Out of Vocabulary, OOV),这为命名实体识别和翻译带来了困难.很多研究都试图借助于双语平行语料来提升命名实体的识别和翻译效果.

基于双语平行语料进行命名实体的翻译获取往往是在双语命名实体对齐的基础上进行的.目前,双语命名实体对齐的方法主要有两种.一种是在双语实体识别的基础上,利用对齐模型寻找两者的对齐关系,这类研究中比较有影响的有 Huang 等人提出了一种基于多特征代价最小的自动抽取命名实体翻译等价对的方法,他们在对齐的过程中融合了多个特征[1]:词典特征、音译特征、实体标签特征以及词汇翻译概率特征等.由于双语命名实体识别的错误会延续到对齐过程中,因此,另一种方法是仅在一种语言中识别命名实体(该语言命名实体识别率较高,如英语),然后利用融合多特征的对齐模型,在另一个语言中寻找它们对应的翻译[2~4].这类方法虽然减少了双语命名实体识别错误的影响,但是却丢失了另一种语言中有用的命名实体信息.有学者研

收稿日期:2014-03-12;修回日期:2014-05-23;责任编辑:蓝红杰
基金项目:国家自然科学基金委优秀国家重点实验室研究项目(No.61223003);国家自然科学基金(青年)(No.61300158)

究发现,双语信息有助于命名实体识别[5~7].因此,Chen 等人[8]提出了一种双语命名实体识别和对齐进行交互的模型,利用双语对齐信息对识别结果和对齐结果进行修正,这种方法不仅提高了实体对齐的效果,而且也有效地提高了实体识别的正确率和召回率.

然而,中文命名实体识别往往是在分词之后进行的,分词错误会影响实体识别结果[9],进而影响双语实体对齐和翻译结果.虽然现有的中文分词系统有较高的水平(F-score 可以达到 95%以上),但是在命名实体词汇的切分中常常出现错误,这主要是因为命名实体词汇往往是未登录词(OOV),而 OOV 造成的分词精度失落至少比分词歧义大 5 倍以上[10].此外,中文分词并没有统一的标准,现有的分词系统基本上是采用"结合紧密,使用稳定"作为分词粒度的界定标准,而实际上不同的应用,对于分词粒度有着不同的需求.已有研究表明,对单语而言最优的分词结果对于双语对齐和机器翻译而言未必是最合适的[11].Chen 等人提出的双语命名实体识别和对齐的交互式模型并没有修正中文分词对命名实体识别和对齐的影响.因此,本文对双语命名实体词汇进行置信度评估,根据置信度高的实体对齐,对命名实体相关的分词错误进行修正从而提高命名实体识别的效果,调整命名实体词汇的分词粒度,优化双语命名实体对齐,提高双语命名实体翻译抽取的效果,进而提高统计机器翻译质量.

2 中文分词对中英命名实体识别和对齐的影响

双语命名实体在没有省略翻译、或者使用简称的情况下,往往边界是统一的.因此利用双语词对齐信息可以修正命名实体识别的错误[8].例如,在图 1 的例子中,中文句子中的命名实体起初没有识别出来,根据识别出的与之对齐的英文命名实体,可以将中文实体识别结果修正为"国际 \ ORG 原子能 \ ORG 机构 \ ORG".

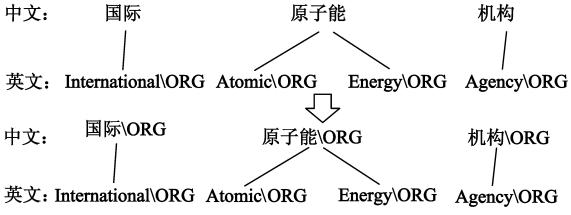

图1 根据对齐信息修正命名实体识别结果

然而,由于中文命名实体识别往往是在分词之后进行,而分词系统往往在命名实体词汇上出现分词错误,特别是音译词汇.分词错误会直接造成命名实体的部分识别或者未识别.例如,在图 2 的例子中,初始的中文分词结果为"基 德曼",基于它的命名实体识别结果为"基 德曼 \ PER",根据双语对齐信息,可以修正命名实体识别的边界,得到"基 \ PER 德曼 \ PER".由于一对一的词对齐结果有利于提升统计机器翻译的质量,因此,把 "基"和"德曼"合并为"基德曼 \ PER"效果会更好.

再例如,在图 3 的例子中,"发展局 \ ORG"和"训练局 \ ORG"初始分词结果为一个词,从对齐结果可以看到,虽然"局"属于中文组织机构名的特征词尾[12],但它本身有独立的英文翻译(Council),因此,将分词调整为"发展 \ ORG 局 \ ORG"和"训练 \ ORG 局 \ ORG",使得对齐结果优化为一对一对齐,这不仅有利于命名实体的翻译知识抽取,而且也有助于机器翻译中命名实体的自动翻译.

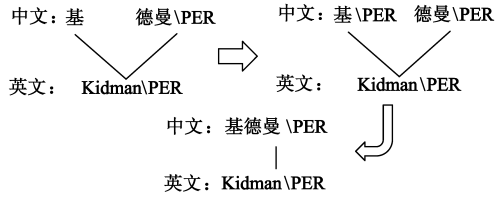

图2 根据对齐修正实体词汇的分词结果

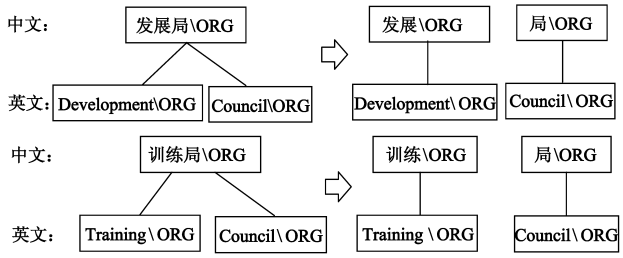

图3 根据对齐调整实体词汇的分词粒度

现在的问题是,依据什么来合并/分开实体词汇,以保证在词对齐中尽可能多地实现中英文实体词汇一一对应? 本文通过命名实体词汇对齐的置信度评估来解决这个问题,下一节我们将首先介绍对齐置信度的评估方法,然后详细说明在置信度高的对齐基础上如何进行分词优化.

3 基于对齐置信度的命名实体分词优化

3.1 对齐置信度的评估方法

命名实体词汇对齐的置信度可以通过下面的方法进行计算.

对于一个中英双语句对,中文在初始分词的基础上进行命名实体识别,结果为 I 个实体词集合 $NE_c = \{(ne_{ci}, type_{ci}): i = 1,2,\cdots,I\}$,其中,$ne_{ci}$ 为中文实体词,$type_{ci}$ 为中文实体类别;英文命名实体的识别结果为 J 个实体词集合 $NE_e = \{(ne_{ej}, type_{ej}): j = 1,2,\cdots,J\}$,其中,

ne_{ej}为英文实体单词,$type_{ej}$为英文实体类别.这里的实体类别只考虑人名(PER)、地名(LOC、GPE)、组织机构名(ORG).在最大熵模型[13]的基础上,根据中英命名实体翻译的一些特点,加入多个特征函数进行对齐置信度的计算,计算模型如下:

$$P(ne_{ej}|ne_{ci}) = p_{\lambda_1^M}(ne_{ej}|ne_{ci}) = \frac{\exp[\sum_{m=1}^{M}\lambda_m h_m(ne_{ej}, ne_{ci})]}{\sum_{ne'_{ej}}\exp[\sum_{m=1}^{M}\lambda_m h_m(ne_{ej}, ne_{ci})]} \quad (1)$$

中文命名实体词ne_{ci}和英文命名实体词ne_{ej}对齐置信度的评估公式为[14]:

$$\hat{ne_e} = \arg\max_{ne_{ej}}\{P(ne_{ej}|ne_{ci})\} = \arg\max_{ne_{ej}}\{\sum_{m=1}^{M}\lambda_m h_m(ne_{ej}, ne_{ci})\} \quad (2)$$

这里采用了 4 个特征:音译特征、意译特征、字对齐特征、同现特征.

(1)音译特征

假设中文命名实体词ne_{ci}有 n 个汉字组成,每个汉字转换为拼音,表示为$\{py_1, py_2, \cdots, py_n\}$,英文命名实体词$ne_{ej}$有 m 个字母组成,表示为$\{e_1, e_2, \cdots, e_m\}$,则$ne_{ci}$和$ne_{ej}$的音译特征函数$h_1(ne_{ej}, ne_{ci})$可以表示为:

$$h_1(ne_{ej}, ne_{ci}) = \begin{cases} 1, & \text{if } py_1\cdots.py_n \supseteq e_1\cdots e_m \\ \mathrm{Sim}(py_1\cdots.py_n, e_1\cdots e_m), & \text{else} \end{cases} \quad (3)$$

式(3)条件$py_1\cdots.py_n \supseteq e_1\cdots e_m$,表示英文单词是中文汉字拼音字符串的子串或两者完全一致.考虑到在人名中使用汉字的亚洲国家或地区中采用不同的拼音体系,因此我们在汉字拼音转换时,会将中文汉字分别转换成普通话拼音、台湾通用拼音、粤语拼音、韩国汉字拼音.如果中文词包含日文常用姓名汉字(500 个日文常用姓名),还将转换成日文汉字的罗马拼音.

对于不符合条件 $py_1\cdots.py_n \supseteq e_1\cdots e_m$ 的中英实体词ne_{ci}和ne_{ej},则只计算普通话拼音字符串和英文单词的相似度.相似度的计算方法是:

$$\mathrm{Sim}(py_1\cdots.py_n, e_1\cdots e_m) = \prod_{j=1}^{J}\sum_{i=1}^{n}p(unit_j|py_i) \quad (4)$$

其中$unit_j(j=1,\cdots,J)$表示英文单词的音译单元.音译过程中,在保证源语言和目标语言发音接近的基础上,还会根据目标语言的发音习惯调整音译单元.受到赵明明等人工作[15]的启发,我们对英文单词音译单元的切分,是在英文单词音节划分的基础上,根据中文音译习惯进行基于音节首字母匹配的调整,具体调整方法在文献[15]中有详细描述.比如"Clinton \ 克林顿"的音译单元划分为"C \ 克","lin \ 林","ton \ 顿","Disney \ 迪士尼"的音译单元划分为"Di \ 迪","s \ 士","ney \ 尼".英文音译单元$unit_j$和中文拼音字符串的py_i的对齐概率$p(unit_j|py_i)$采用 IBM Model 1[16]在中英音译语料库中训练得到,我们采用的语料库是来自于《世界人名翻译大辞典》[17],其中包含了 59 万多条欧美人名的中文音译词汇.

(2)意译特征

我们用中文命名实体 ne_{ci}和英文命名实体 ne_{ej}的翻译概率来表示意译特征,利用 GIZA + + 工具产生的 IBM Model 1 、Model 4 以及 HMM 三个模型的双向词汇翻译概率来计算意译特征,计算公式为:

$$h_2(ne_{ej}, ne_{ci}) = \frac{p_{\mathrm{Model1}}(ne_{ej}, ne_{ci}) + p_{\mathrm{Model4}}(ne_{ej}, ne_{ci}) + p_{HMM}(ne_{ej}, ne_{ci})}{3} \quad (5)$$

$$p_{\mathrm{Model1}}(ne_{ej}, ne_{ci}) = p_{\mathrm{Model1}}(ne_{ej}|ne_{ci}) + p_{\mathrm{Model1}}(ne_{ci}|ne_{ej}) \quad (6)$$

$$p_{\mathrm{Model4}}(ne_{ej}, ne_{ci}) = p_{\mathrm{Model4}}(ne_{ej}|ne_{ci}) + p_{\mathrm{Model4}}(ne_{ci}|ne_{ej}) \quad (7)$$

$$p_{HMM}(ne_{ej}, ne_{ci}) = p_{HMM}(ne_{ej}|ne_{ci}) + p_{HMM}(ne_{ci}|ne_{ej}) \quad (8)$$

(3)字对齐特征

受 Xi, et al[18]工作的启发,中文"字对齐"的信息可以用来评估词汇对齐的置信度.假设中文命名实体词汇ne_{ci}由 K 个汉字组成,表示为$ne_{ci} = \{c_1, c_2, \cdots, c_K\}$,则$ne_{ci}$和英文实体词汇$ne_{ej}$的字对齐特征可以用下面的公式计算:

$$h_3(ne_{ej}, ne_{ci}) = \sum_{k=1}^{K}p(ne_{ej}|c_k) \quad (9)$$

将中英平行语料中的中文句子按字切开,利用 IBM Model 1 可以计算得到中文汉字 c_k 和英文单词ne_{ej}的翻译概率$p(ne_{ej}|c_k)$.

(4)共现特征

如果中文命名实体ne_{ci}和英文命名实体ne_{ej}在整个语料中总是在平行句对中出现,那么这两个词汇在一定程度上有互译的可能性,因此我们在计算对齐置信度时采用共现特征作为特征之一.共现特征的计算公式如下:

$$h_4(ne_{ej}, ne_{ci}) = \frac{\mathrm{count}(ne_{ej}, ne_{ci})}{\mathrm{count}(ne_{ci})} + \frac{\mathrm{count}(ne_{ej}, ne_{ci})}{\mathrm{count}(ne_{ej})} \quad (10)$$

其中,$\mathrm{count}(ne_{ej}, ne_{ci})$表示中英实体词汇$ne_{ci}$和$ne_{ej}$同在一个平行句对出现的次数,$\mathrm{count}(ne_{ci})$表示中文实体词汇$ne_{ci}$在语料库中出现的次数,$\mathrm{count}(ne_{ej})$表示英文实体词汇$ne_{ej}$在语料库中出现的次数.我们在统计命名实体相关词汇共现信息时,为了降低运算复杂度,在统计中进行了初步的剪枝,去除了数字、标点符号以及首字母是小写的英文非实体单词,并且统计结果中只保

留共现特征较强的词汇对,词对数从最初的 183 万多个减少到 110 万多个.

本文使用开源工具包 OpenNLP[19] 中 Maxent 工具来训练公式(2)中的 4 个特征函数的参数.命名实体词汇对齐的置信度评估是本文优化方法的基础,下面我们详细说明中文分词的优化方法.

3.2 分词优化

通过公式(2)的计算,我们可以获得平行语料中对齐置信度较高的双语命名实体词汇.根据本文第 2 节对中文分词、命名实体识别和对齐的相互影响分析,首先利用对齐信息优化命名实体识别结果,在此基础上优化命名实体相关的分词结果,通过优化分词粒度和修正分词错误,实现中英文实体词汇一一对应.整个过程可以用下面 4 个算法描述:

算法 1 根据对齐修正命名实体边界.在对齐的实体词汇前后,寻找是否存在非实体词汇和实体词汇对齐,并且对齐置信度较高.根据双语命名实体边界统一的规则,修正初始命名实体识别边界上的错误.

Function: ReviseNEBoudary(C_0, E_0, $Align_0$)
Inputs: C_0:初始中文句子单词数组
E_0:初始英文句子单词数组
$Align_0$:初始对齐
Outputs: C_1:算法 1 执行后的中文单词数组
E_1:算法 1 执行后的英文单词数组
$Align_1$:算法 1 执行后的中英文单词对齐
begin
$C_1 = \phi$; $E_1 = E_0$; $Align_1 = Align_0$;
for each 中文词 c_i in C_0 do
if c_i 不是实体词 and 在 $Align_0$ 中 c_i 对齐到英文实体词汇 e
if c_i 和 e 的对齐置信度较高
根据 e 的实体类型将 c_i 修正为该类型的实体词 c_i^*;
$C_1 = C_1 \cup c_i^*$;
end if
else
$C_1 = C_1 \cup c_i$;
end if
end for
end begin

算法 2 合并对齐到一个英文 *NE* 的中文 *NE*.在实体对齐结果中,寻找是否存在多个中文实体词汇对应一个英文单词,如果存在,则将这些中文词合并为新的命名实体词.

Function: CombineChineseNE(C_1, E_1, $Align_1$)
Inputs: C_1:算法 1 执行后的中文单词数组
E_1:算法 1 执行后的英文单词数组
$Align_1$:算法 1 执行后的中英文单词对齐
Outputs: C_2:算法 2 执行后的中文单词数组
E_2:算法 2 执行后的英文单词数组
$Align_2$:算法 2 执行后的中英文单词对齐
begin
$C_2 = C_1$; $E_2 = E_1$; $Align_2 = \phi$;
for each 英文实体词 e_m in E_1 do
根据 $Align_1$,计算 e_m 和对齐的中文实体词个数 x;
if $x > 1$ and 对齐的中文实体词序号连续
将 C_2 中这些中文词为一个新词 c,实体类型同 e_m,将 c 和 e_m 的对齐加入 $Align_2$;
end if
end for
根据 $Align_1$,将没有变动的词对齐加入 $Align_2$;
end begin

算法 3 根据音译拼音对齐,切分中文 *NE*.在对齐结果中,根据音译特征,减小中文音译实体词的分词粒度,使得新切分出的中文实体词和英文实体词一一对应.

Function: SegmentTransliterateNE(C_2, E_2, $Align_2$)
Inputs: C_2:算法 2 执行后的中文句子单词数组
E_2:算法 2 执行后的英文句子单词数组
$Align_2$:算法 2 执行后的中英文单词对齐
Outputs: C_3:算法 3 执行后的中文单词数组
E_3:算法 3 执行后的英文单词数组
$Align_3$:算法 3 执行后的中英文单词对齐
begin
$C_3 = C_2$; $E_3 = E_2$; $Align_3 = \phi$;
for each 英文实体词 e_m in E_2 do
计算 e_m 和对齐的中文实体词 c 的音译特征值 h;
if $h = 1$ and e_m 是 c 拼音字符串的子串
在 C_3 中,将 c 切分为 c' 和 c'' 两个实体词,类型和 c 相同;
c' 的拼音字符串和 e_m 相同,将 c' 和 e_m 的对齐加入 $Align_3$;
if 英文实体词 e_{m-1} 和 c'' 的拼音字符串相同,
将 c'' 和 e_{m-1} 的对齐加入 $Align_3$;
end if
if 英文实体词 e_{m+1} 和 c'' 的拼音字符串相同
将 c'' 和 e_{m+1} 的对齐加入 $Align_3$;
end if
end if
end for
根据 $Align_2$,将没有变动的词对齐加入 $Align_3$;
end begin

算法 4 分离中文实体特征词尾.在对齐结果中,将中文地名、组织结构名的特征词尾切分为独立的词,在英文实体中寻找是否存在特征词尾的翻译,如果存在,则将其和特征词尾对齐,否则将特征词尾对齐到空.

Function: SplitNESuffix(C_3, E_3, $Align_3$)
Inputs: C_3:算法 3 执行后的中文句子单词数组
E_3:算法 3 执行后的英文句子单词数组
$Align_3$:算法 3 执行后的中英文单词对齐
Outputs: C_4:算法 4 执行后的中文单词数组
E_4:算法 4 执行后的英文单词数组
$Align_4$:算法 4 执行后的中英文单词对齐

```
begin
  C4 = C3; E4 = E3; Align4 = ϕ;
  for each 中文实体词 ci in C3 do
    if ci 是地名或机构名 and ci 包含特征字尾 w
      计算 ci 对齐的英文实体单词集合 E;
      if E 的元素个数 = 1
        在 C4 中,将 ci 切分为 c′ 和 w 两个实体词,类型和 ci 相同;
        将 c′ 和 E 中元素的对齐加入 Align4;
      end if
      if E 的元素个数 > 1
        if ∃ em in 特征字尾 w 互译英文单词集合 ew and em in E
          C4 中,将 ci 切分为 c′ 和 w 两个实体词,类型和 ci 相同;
          将 w 和 em 的对齐加入 Align4;
          将 Align3 中,c′ 和除 em 之外的对齐加入 Align4;
        else
          在 C4 中,将 ci 切分为 c′ 和 w 两个实体词,类型和 ci 相同;
          将 c′ 和 E 中元素的对齐加入 Align4;
        end if
      end if
    end if
  end for
  根据 Align3,将没有变动的词对齐加入 Align4;
end begin
```

上述 4 个算法的优化效果是:(1)利用对齐修正实体对齐的边界,从而优化实体识别结果,进而得到更多的实体对齐;(2)将因为分词粒度过大或错误切分造成的一对多或多对一的对齐,通过分词优化,实现一对一的对齐,从而达到优化实体对齐的效果.

下一节我们通过实验来说明优化算法对中英双语命名实体短语翻译的抽取以及统计机器翻译的影响.

4　实验和分析

4.1　中英命名实体短语翻译抽取实验

为了验证本文优化算法的有效性,我们对优化前后中英命名实体短语翻译抽取和统计机器翻译的效果进行对比.本文的算法是针对命名实体词汇相关的分词优化,命名实体在语料中的比例会直接影响到实验效果,新闻语料包含命名实体较多,因此我们采用新闻语料检测本文算法的有效性.命名实体短语翻译抽取实验中,我们从 LDC 发布的中英新闻语料库(LDC2002E18;LDC2005T06)中抽取 300 个中英长句对作为我们的测试集,因为长句可能包含多个命名实体词汇,其中中文句子平均长度约为 61 个汉字,英文句子平均长度约为 26 个单词.由于优化前后中文分词结果有变动,因此我们不对双语命名实体词对齐进行评估,而是对双语命名实体短语的抽取进行评估,本文的优化算法没有考虑修正实体类型的识别错误,并且对齐双语实体类型的识别可能出现不一致,因此将人名、地名以及组织结构名的命名实体短语翻译综合在一起评估,评估采用正确率(Precision,P)、召回率(Recall,R)和 F-score(F).

我们采用中科院分词工具 ICTCLAS、Stanford 分词工具对中文句子进行分词,其中 Stanford 分词工具分别采用基于(Chinese Treebank,CTB)和(Peking university,PKU)两种数据集上训练的模型.对上述三种中文分词结果,均采用 Stanford 命名实体识别工具(分类器:chinese.misc.distsim.crf.ser)进行初始中文命名实体识别.英文句子采用 Stanford 命名实体识别工具(分类器:english.conll.4class.distsim.crf.ser)进行英文命名实体识别.

我们首先采用 GIZA++ 进行词对齐,得到 IBM Model 1、IBM Model 4 和 HMM 三种词对齐结果(grow.model1.align、grow.model4.align 和 grow.hmm.align),然后,使用 Koehn 等人提出的 grow-diag-final 启发式规则[20]对每个模型的词对齐进行合并,加入命名实体词汇的对齐置信度评估,得到词对齐结果(grow.baseline.align),该结果相对于原始对齐结果具有较低的错误率.在这个词对齐结果上利用基于层次短语的翻译系统[21]进行短语抽取,从中得到中英命名实体短语翻译,其中包括人名、地名、组织机构名,以此作为 baseline.在 grow.baseline.align 的词对齐基础上,利用对齐,修正命名实体识别边界和分词错误,优化分词粒度.在优化后的词对齐结果上,再次进行短语抽取,得到优化后的中英命名实体短语翻译.表 1 给出了优化结果与 baseline 的对比.

表 1　中英命名实体短语翻译抽取优化前后对比

	baseline			优化后的结果		
	Precision	Recall	F-score	Precision	Recall	F-score
ICTCLAS	80.91%	82.87%	81.88%	87.02%	89.10%	88.05%
Stanford-PKU	77.12%	79.51%	78.30%	85.61%	88.34%	86.95%
Stanford-CTB	79.44%	83.16%	81.26%	86.89%	89.48%	88.17%

从实验结果可以看出,在不同的初始中文分词结果基础上,本文的优化算法都使得双语实体短语的抽取结果有了提高.我们对实验结果进行进一步分析发现:本文的优化算法对于采用音译翻译方式的命名实

体有比较明显效果,而对英文简称相关的双语短语翻译在准确率上不是特别理想,因为简称相关的词汇对齐置信度不高,本文优化算法的前提是初始命名实体对齐具有较高的置信度.

4.2 中英统计机器翻译实验

统计机器翻译实验中,中英训练语料有 469 万多句对,语言模型采用 SRILM 对于训练语料的英文句子进行大小写不敏感的训练,得到一个五元语法模型作为机器翻译系统的语言模型.我们采用 CWMT2013 的开发集作为实验开发集,SSMT-2007 的测试集(1002 个句对)作为实验的测试集,这两个数据集是中英双语新闻为主,包含较多的命名实体词汇.统计机器翻译的结果用大小写不敏感的 NIST-BLEU 值(4-gram)进行评价.

我们在中英新闻项目上一共做了 3 个词对齐结果,这 3 个词对齐结果是:(1)4.1 节中提到的 grow-diag-final 启发式规则将 GIZA++ 获得的 3 种词对齐结果(IBM Model 1、IBM Model 4 和 HMM),加入命名实体对齐的置信度评估,得到词对齐结果 grow.baseline.align 词对齐(记为 baseline).(2)在 grow.baseline.align 基础上,类似于 chen 等人提出方法[8],以英文实体词汇为参照,在中文实体词汇的前后,采用滑动窗口的方法,根据对齐信息,只修正中文实体识别的边界,不对分词结果进行修改,得到词对齐结果 grow.boundary.align(记为 boundary);(3)在 grow.baseline.align 基础上,采用本文的优化算法得到的词对齐结果 grow.final.align(记为 final).我们使用基于层次短语的翻译系统在这三个词对齐基础上,将命名实体翻译知识加入统计机器翻译,表 2 给出了这三个实验中 NIST-BLEU 值的变化.

从实验结果可以看出,仅仅对命名实体边界进行修正对提升 BLEU 值的作用不明显,而在修正命名实体边界的基础上,进一步修正分词错误和优化分词粒度可以提升 BLEU 值,这一点在不同的词对齐结果上都有体现.不过,由于本文分词优化方法是仅针对命名实体词汇进行的,因此 BLEU 的提升幅度较小.

表 2 三种词对齐结果的 SMT NIST-BLEU 变化

	baseline	boundary	final
ICTCLAS	27.61	27.70	27.99
Stanford_PKU	28.09	28.11	28.65
Stanford_CTB	27.84	27.73	28.27

5 结束语

中文分词结果对命名实体识别以及对齐有着直接的影响,本文在命名实体词汇的对齐置信度评估的基础上,通过对分词粒度的优化,错误分词的修正,优化了中英命名实体识别和对齐的结果,进而提高了中英命名实体翻译的抽取效果,以及中英新闻领域的统计机器翻译的效果.实验结果表明了本文算法是有效的.

本文的优化算法只考虑了命名实体识别结果中标记为人名、地名、组织机构名的实体词汇,未考虑标记为"MISC(杂项)"的实体,在这类实体识别结果中,存在着一定数量被错误标记的人名、地名、组织结构名,我们将在今后的工作中考虑如何修正错误的命名实体类型标记,进一步提高双语命名实体识别效果.此外,在今后的工作中我们也将考虑如何利用双语对齐信息优化命名实体以外的词汇分词结果,进一步提高双语词对齐效果.

参考文献

[1] Huang F, Vogel S, Waibel A. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization[A]. Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition-Volume 15[C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003.9 – 16.

[2] Feng D, Lü Y, Zhou M. A new approach for English-Chinese named entity alignment[A]. EMNLP. 2004[C]. Barcelona: Association for Computational Linguistics, 2004.372 – 379.

[3] Lee C J, Chang J S, Jang J S R. Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources[J]. ACM Transactions on Asian Language Information Processing(TALIP), 2006, 5(2): 121 – 145.

[4] Moore R C. Learning translations of named-entity phrases from parallel corpora[A]. Proceedings of the tenth Conference on European Chapter of the Association for Computational Linguistics-Volume 1[C]. Budapest: Association for Computational Linguistics, 2003.259 – 266.

[5] Ji H, Grishman R. Collaborative entity extraction and translation [J]. Recent Advances in Natural Language Processing V: Selected Papers from RANLP 2007, 2009, 309: 73 – 84.

[6] Che W, Wang M, Manning C D, et al. Named entity recognition with bilingual constraints[A]. Proceedings of HLT-NAACL[C]. Atlanta: Association for Computational Linguistics, 2013.52 – 62.

[7] Wang M, Che W, Manning C D. Joint word alignment and bilingual named entity recognition using dual decomposition [A]. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics[C]. Sofia: Association for Computational Linguistics, 2013.1073 – 1082.

[8] Chen Y, Zong C, Su K Y. A joint model to identify and align bilingual named entities[J]. Computational Linguistics, 2013, 39(2): 229 – 266.

[9] Zhou J, Qu W, Zhang F. Chinese named entity recognition via joint Identification and categorization[J]. Chinese Journal of

Electronics,2013,22(2):225 – 230

[10] 黄昌宁,赵海.中文分词十年回顾[J].中文信息学报,2007,21(3):8 – 19.
Huang C,Zhao H. Chinese word segmentation:A decade review[J]. Journal of Chinese Information Processing,2007,21(3):8 – 20.(in Chinese)

[11] Chang P C,Galley M,Manning C D. Optimizing Chinese word segmentation for machine translation performance[A]. Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics[C]. 2008.224 – 232.

[12] 冯元勇,孙乐,张大鲲,等.基于小规模尾字特征的中文命名实体识别研究[J].电子学报,2008,36(9):1833 – 1838.
Feng Y Y,Sun L,Zhang D K. Study on the Chinese named entity recognition using small scale character tail hints[J]. ACTA Electronica Sinica,2008,36(9):1833 – 2838.(in Chinese)

[13] BergerA L,Pietra V J D,Pietra S A D. A maximum entropy approach to natural language processing[J]. Computational Linguistics,1996,22(1):39 – 71.

[14] Och F J,Ney H. Discriminative training and maximum entropy models for statistical machine translation[A]. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics[C]. Philadelphia:Association for Computational Linguistics,2002.295 – 302.

[15] 赵明明,梁颖红,周美玲,等.基于音节首字母匹配的音译单元对齐方法[J].江南大学学报(自然科学版),2009,8(6):639 – 642.
Zhao MM,Liang Y H,Zhou M L. Transliteration unit alignment method based on the first syllable letter mapping[J]. Journal of Jiangnan University(Natural Science Edition),2009,8(6):639 – 642.(in Chinese)

[16] Brown P F,Pietra V J D,Pietra S A D,et al. The mathematics of statistical machine translation:Parameter estimation[J]. Computational linguistics,1993,19(2):263 – 311.

[17] 世界人名翻译大辞典[M].北京:中国对外翻译出版公司,2007.
Names of the World's Peoples:a Comprehensive Dictionary of Names in Roman-Chinese[M]. Beijing:China Translation & Publishing Corporation,2007.(in Chinese)

[18] Xi N,Tang G,Dai X,et al. Enhancing statistical machine translation with character alignment[A]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics:Short Papers-Volume 2[C]. Jeju Island:Association for Computational Linguistics,2012.285 – 290.

[19] Bierner G,Baldridge J,Morton T. OpenNLP:A collection of natural language processing tools[Z]. URL http://opennlp.sourceforge.net,2007.

[20] Koehn P,Och F J,Marcu D. Statistical phrase-based translation[A]. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1[C]. Budapest:Association for Computational Linguistics,2003.48 – 54.

[21] Chiang D. A hierarchical phrase-based model for statistical machine translation[A]. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics[C]. Ann Arbor:Association for Computational Linguistics,2005.263 – 270.

作者简介

尹存燕　女,1976 年生于江苏南京.南京大学计算机科学与技术系讲师.研究方向为自然语言处理.
E-mail:yincy@nju.edu.cn

黄书剑　男,1984 年生于江苏南京.南京大学计算机科学与技术系助理研究员.研究方向为统计机器翻译和自然语言处理.
E-mail:huangsj@nju.edu.cn

戴新宇　男,1979 年生于江苏盱眙.南京大学计算机科学与技术系副教授.研究方向为自然语言处理和文本挖掘.
E-mail:daixinyu@nju.edu.cn

陈家骏　男,1963 年生于江苏南京.南京大学计算机科学与技术系教授、博士生导师.研究方向为自然语言处理和软件工程.
E-mail:chenjj@nju.edu.cn