汉英双语命名实体识别与对齐的交互式方法

陈钰枫[1)] 宗成庆[1)] 苏克毅[2)]

[1)](中国科学院自动化研究所模式识别国家重点实验室 北京 100190)

[2)](台湾致遠科技公司 台湾 新竹)

摘 要 基于汉英双语命名实体的识别与对齐特性,文中提出了一种双语命名实体交互式对齐模型,其中的修正对齐计算体现了汉英实体识别与对齐的密切结合:一方面,利用双语对齐信息帮助实体识别;另一方面,实体的对齐过程对实体的识别结果又具有一定的修正作用,两方面的结合实现了双语实体识别与对齐之间的交互式互助过程.实验证明,这种交互式对齐模型不仅显著提高了汉英实体对齐的性能(F 值从 74.4%提高到 81.2%),而且有效地提高了汉英实体识别的正确率和召回率.

关键词 命名实体;识别;双语对齐;交互;机器翻译

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2010.01688

Joint Chinese-English Named Entity Recognition and Alignment

CHEN Yu-Feng[1)] ZONG Cheng-Qing[1)] SU Keh-Yih[2)]

[1)](*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academic of Sciences, Beijing* 100190)

[2)](*Behavior Design Corporation, Hsinchu, Taiwan*)

Abstract Named entity (NE) recognition is an essential early stage and has long been a thorny problem in many natural language processing pipelines. Generally, bilingual named entity recognition and alignment are processed sequentially and independently, regardless of interactions between the two stages. Therefore, NE recognition errors are propagated and compounded in NE alignment stage. Actually, bilingual alignment information, other than monolingual information, provides further indications for NE recognition. It is necessary to capture the interactions between NE recognition and alignment. Accordingly, based on the characteristics of NE recognition and alignment, this paper proposes an interactive bilingual NE alignment model, which combines basic alignment and corrective alignment. Basic alignment is achieved as traditional alignment approach, while the corrective alignment realizes the joint of NE recognition and alignment. On the one hand, bilingual alignment information is utilized for NE recognition; on the other hand, NE recognition errors can be recovered in the NE alignment stage. Both the NE boundaries and type can be corrected in such interactive NE alignment model. The experiments show that this novel model not only achieves a significant improvement of the Chinese-English NE alignment quality (F-score from 74.4% to 81.2%), but also improves the performance of NE recognition.

Keywords named entity; recognition; bilingual alignment; interaction; machine translation

---

收稿日期:2008-03-03;最终修改稿收到日期:2010-04-26.本课题得到国家自然科学基金(60975053,60736014)、国家支撑计划项目(2006BAH03B02)和国家"八六三"高技术研究发展计划项目基金(2006AA010108-4)资助.**陈钰枫**,女,1981 年生,博士,助理研究员,研究方向为自然语言处理和机器翻译. E-mail: chenyf@nlpr.ia.ac.cn. **宗成庆**,男,1963 年生,博士,研究员,博士生导师,研究领域为机器翻译、文本分类、句法分析等. **苏克毅**,男,1955 年生,博士,教授,研究领域为统计机器学习在自然语言处理中的应用,并致力于构建高性能的技术手册的英汉翻译系统.

1 引　言

命名实体，尤其人名、地名与机构名这三类命名实体在自然语言中传递着关键信息，是信息处理的重点与难点. 对于单语序列，命名实体识别是众多自然语言处理领域的基础工作，其性能直接影响后续的信息处理步骤. 对于双语序列，双语命名实体的对齐旨在建立源语言和目标语言中命名实体之间关系，是机器翻译、跨语言信息检索等多语言处理领域的一项重要工作.

一般来说，双语命名实体对齐任务首先识别出源语言或目标语言中的命名实体，再实现它们之间的对应，这需要保证识别过程和对齐过程的准确率，因此，双语命名实体对齐成为一项挑战性很大的工作，具体体现在以下两个方面：(1) 双语命名实体的对齐结果很大程度上依赖于实体识别，而识别结果中通常存在比较多的错误，主要包括部分识别、丢失和假性识别(把非命名实体识别成命名实体). 实体对齐过程如果直接采用识别结果，必然导致识别错误的延续和扩大；(2) 双语命名实体的对齐本身难度很大，因为它不仅属于多词对应多词的对齐任务，还包括意译①和音译两种对齐形式. 因此双语命名实体对齐一般采用多特征和多语言知识融合的系统.

近几年来，双语命名实体对齐的研究主要致力于多特征对齐模型的建立，对齐本身已经取得了比较好的效果，但是它与实体的识别性能紧密相关，如果实体识别的错误没有经过修正，直接延续到对齐过程中，那么对齐性能将无法从根本上提高. 因此，双语命名实体的对齐如何同实体识别相结合，具有一定的修正识别结果的能力，是双语命名实体对齐任务的瓶颈问题. 从实体识别的角度上，双语命名实体对齐方法可以分为两大类. 一类是识别双语实体后再对齐，表示分别在源语言和目标语言中识别出命名实体，然后再建立它们之间的对齐关系. 例如，文献[1]建立了一个多特征融合的模型来抽取双语语料库中的命名实体翻译对. 这类对齐方法需要双语实体识别，因此很大程度上依赖于识别结果. 另一类是识别单语实体后再对齐，表示只识别出源语言命名实体，然后在目标语言中确定它们的翻译对应. 只在一方进行实体识别，降低了对齐过程对识别结果的依赖程度，因此大部分对齐工作[2-5]都属于这类方法. 但是这类对齐方法丢失了目标语言命名实体的识别信息，并且没有在根本上解决实体对齐过程对识别结果的依赖性问题. 上述的两类对齐方法都没有校验识别结果的置信度，因此，这样的实体对齐系统无法修正识别错误，导致识别错误延续到对齐过程中，很大程度上影响到最后的对齐结果.

针对命名实体识别，国内外已经有大量的深入研究，但识别结果还有待完善，实体识别仍然是自然语言处理领域的热点. 目前，实体识别除了利用单语序列上的局部信息(词条、词性标注、chunk 标注等)，还包括单语序列上的全局信息[6](全文标注统一性等). 此外，另一有待深入研究的可利用资源是：双语序列的对位信息对识别的辅助. 例如，文献[7]将双语语料作为反馈信息来提高源语言命名实体的识别性能. 但利用双语对齐信息来辅助实体识别，同时提高实体对齐效果，目前尚无研究.

实际上，双语命名实体对齐虽然是实体识别的后续过程，但对齐信息却能辅助命名实体的识别，修正已有的识别错误. 于是，我们的研究重点在于如何将实体对齐信息反馈给实体识别. 针对汉英双语语料，我们建立起双语实体识别与对齐相结合的整体框架. 提出一种汉英双语命名实体交互式对齐模型，使双语实体对齐具有修正识别的功能，实现二者性能的同时提高.

本文第 2 节给出汉英命名实体识别与对齐的特性分析；第 3 节提出了一种汉英命名实体交互式对齐模型，详细介绍了其中的修正对齐计算，它用于修正实体边界和类别识别的错误，从而实现了实体识别与对齐的交互；第 4 节给出实验结果和分析部分；最后一节是本文的结束语.

2 汉英命名实体识别与对齐的特性分析

命名实体识别通常包括两部分：(1) 实体边界识别；(2) 确定实体类别(人名、地名、机构名或其他). 英文命名实体具有比较明显的形式标志(即实体中的每个词的第一个字母要大写)，所以其实体边界的识别相对容易，识别任务的重点在于确定实体的类别. 与英文相比，中文命名实体的识别任务更加复杂，而且相对于实体类别的标注，中文实体边界的判断更加困难. 通过平行的汉英双语语料是否可以使汉英实体双方的识别特性互补呢？ 实际上，双语

① 指普遍含义上的翻译概念，为与音译区别，本文均指意译.

对齐信息提供了双语实体边界和类别的统一性，可以用于调整和修正汉英实体的识别.

一方面，双语实体普遍是边界统一的(也就是双语实体的内部词一一对应的①). 因此，双语实体的词对齐有利于单语实体边界的确定. 例如：

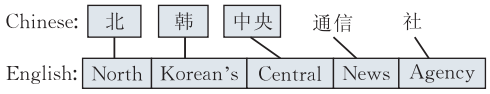
(1) 中文命名实体的识别结果：

官方的〈ORG〉北韩中央〈/ORG〉通信社引述海军声明…

(2) 英文命名实体的识别结果：

Official 〈ORG〉 North Korean's Central News Agency 〈/ORG〉 quoted the navy's statement…

(3) 双语命名实体对齐：



(4) 中文命名实体的边界修正：

官方的〈ORG〉北韩中央通信社〈/ORG〉引述海军声明…

上例中，中文实体的识别结果“北韩中央”属于部分识别的错误结果，英文实体“North Korean's Central News Agency”是正确的识别结果. 通过正确的汉英实体对齐(内部词对齐)，“News Agency”和“通信社”对齐后，中文实体可以被修正为“北韩中央通信社”.

另一方面，对齐的双语实体应该是类别统一的. 不仅类别统一性有利于确定双语实体的类别，而且不同类别的实体与对齐形式(包括意译和音译)存在很大的关联，也就是双语实体的对齐方式也有助于实体类别的判断. 例如：

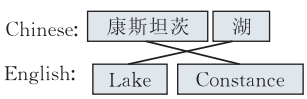
(1) 中文命名实体的识别结果：

在〈LOC〉康斯坦茨湖〈/LOC〉工作的一艘渡船船长…

(2) 英文命名实体的识别结果：

The captain of a ferry boat who works on 〈PER〉 Lake Constance 〈/PER〉 …

(3) 双语命名实体对齐：



(4) 英文命名实体的类别修正：

The captain of a ferry boat who works on 〈LOC〉 Lake Constance 〈/LOC〉 …

上例中，英文实体“Lake Constance”被错误地识别成人名，通过与中文地名“康斯坦茨湖”对齐后，有助于被修正为地名. 如果更深入地分析，上例中的词对齐，“湖”与“Lake”，属于意译对齐. 一般来说，人名普遍采用音译进行翻译，所以这个信息暗示了这个英文实体“Lake Constance”不可能是人名. 只可能是地名或机构名. 由此可见，实体的对齐形式有助于我们对命名实体类别的判断.

我们继续深入探讨双语命名实体翻译对内部的对齐形式，也就是意译和音译的组合方式. 文献[8]指出不同类别的实体倾向于不同的对齐形式. 人名对齐主要是音译形式，地名和机构名的对齐是意译和音译形式的组合，而且在机构名对齐中，意译形式占的比重更大. 针对 LDC 机构发布的汉英双语命名实体语料库(LDC2005T34)，我们通过其中意译词和音译词的频率统计，发现在人名翻译对、地名翻译对和机构名翻译对中，音译词所占的比重分别是 100%、89.4% 和 12.6%，相差的幅度非常大，这启示我们可以采用双语实体翻译对中音译对齐或意译对齐的比重来辅助实体类别的判断.

综上所述，双语实体的对齐信息为实体识别提供了边界和类别的判断信息. 在顺序处理系统中，实体对齐作为实体识别的后续过程，如果能有效利用双语对齐信息对实体识别的反馈作用，必然能避免过多的识别错误，提高识别的准确率，同时也提高对齐的正确率.

3 双语命名实体交互式对齐模型

通常情况下，命名实体识别过程和对齐过程是顺序进行的，没有考虑这两个过程之间的交互作用，因此，传统的对齐模型直接基于实体识别结果，而无法修正识别结果的错误. 为了引入对齐信息对实体识别的反馈辅助作用，同时提高对齐效果，我们建立整体推导框架，将传统的双语实体对齐模型扩展为交互式对齐模型.

汉英双语命名实体交互式对齐任务描述如下：给定汉英双语对齐的句子，分别识别出其中包含的汉英命名实体，并且实现它们之间的对齐. 一般直接采用汉英实体识别工具，可以分别找出汉英句子包含的命名实体，中文命名实体的识别结果 $\widetilde{CNE}_1^S = \widetilde{CNE}_1, \cdots, \widetilde{CNE}_s, \cdots, \widetilde{CNE}_S$ 和英文命名实体的识别结果 $\widetilde{ENE}_1^T = \widetilde{ENE}_1, \cdots, \widetilde{ENE}_t, \cdots, \widetilde{ENE}_T$，我们定义 $m_k = (\widetilde{CNE}_s, \widetilde{ENE}_t)$ 是其中的一条实体对应，表示 $\widetilde{CNE}_s$ 和 $\widetilde{ENE}_t$ 互为翻译或部分翻译(包括意译和音译形式). 因此，直接基于识别结果的基本对齐 M

① 当某一方的实体存在省略、简略等情况时，双语实体的内部词不一定一一对应.

被定义为双语实体识别结果的笛卡尔积的子集. $M\subseteq\{(\widetilde{CNE}_s,\widetilde{ENE}_t):s=1,2,\cdots,S;t=1,2,\cdots,T\}$，其中，我们不考虑对空情况.

但是，实体识别系统得到的识别结果通常存在许多错误，因此，我们的实体交互式对齐任务不仅实现实体识别结果的基本对齐，还要实现实体识别结果的修正，包括实体的边界和类别的修正. 也就是在一个基本对齐 $m_k=(\widetilde{CNE}_s,\widetilde{ENE}_t)$ 的基础上，通过汉英实体左右边界的字或词的缩放，获得 $\widetilde{CNE}_s$ 和 $\widetilde{ENE}_t$ 修正结果的候选项 CNE_k 和 ENE_k，因此，一个修正后的实体翻译对定义为 $a_k=(CNE_k,ENE_k)$，其中，CNE_k 和 ENE_k 同为一种类别 $(type_k)$（本文只考虑 3 种类别：人名、地名和机构名），因此修正后的实体对齐定义为 $A=\langle a_k,type_k\rangle_{k=1}^{K}$，表示该汉英句子中共包含 K 个实体翻译对，$\langle a_k,type_k\rangle$ 是第 k 个实体翻译对，所以汉英双语实体的交互式对齐模型定义如下：

给定汉英双语的句子翻译对：$ChnS$ 和 $EngS$，借助识别工具得到汉英实体的识别结果 $\widetilde{CNE}_1^S$，$\widetilde{ENE}_1^T$，我们的目标是实现它们之间的基本对齐 M，然后通过对齐信息修正识别结果，实现修正的实体对齐 A，包括对齐后实体翻译对的类别.

$$A^*=\arg\max_A\left[\max_M P(A,M|\widetilde{CNE}_1^S,\widetilde{ENE}_1^T,ChnS,EngS)\right] \tag{1}$$

其中，A^* 是最优的双语实体对齐结果. $P(A,M|\widetilde{CNE}_1^S,\widetilde{ENE}_1^T,ChnS,EngS)$ 的推导如下：

$$\begin{aligned}&P(A,M|\widetilde{CNE}_1^S,\widetilde{ENE}_1^T,ChnS,EngS)\\&=P(A|M,\widetilde{CNE}_1^S,\widetilde{ENE}_1^T,ChnS,EngS)\times\\&\quad P(M|\widetilde{CNE}_1^S,\widetilde{ENE}_1^T,ChnS,EngS)\\&\approx\prod_{a_k\in A}\prod_{m_k\in M}P(\langle a_k,type_k\rangle|m_k,ChnS,EngS)\times\\&\quad P(m_k|\widetilde{CNE}_s,\widetilde{ENE}_t)\end{aligned} \tag{2}$$

$P(m_k|\widetilde{CNE}_s,\widetilde{ENE}_t)$ 代表传统的基本对齐计算（直接基于识别结果），$P(a_k,type_k|m_k,ChnS,EngS)$ 是修正对齐计算，通过已有的对齐 m_k，得到修正的实体翻译对 a_k，并确定它的类别 $type_k$，体现了实体对齐和识别之间的交互. 由式(2)可见，双语实体交互式对齐包括了基本对齐计算和修正对齐计算. 基本对齐计算体现了首先实体识别，然后实体对齐的顺序过程. 而修正对齐计算作为双语实体识别和对齐之间的桥梁，使实体识别和对齐相辅相成，同时实现实体识别和对齐性能的提高.

3.1　基本对齐计算

我们直接对 $P(m_k|\widetilde{CNE}_s,\widetilde{ENE}_t)$ 采用最大熵[9]模型进行建模，在此框架下，设计一组特征函数 $h_f(m_k,\widetilde{CNE}_s,\widetilde{ENE}_t)$，其中，$f=1,2,\cdots,F$，对于每个特征函数 h_f，都有相应的模型参数 λ_f，$f=1,2,\cdots,F$. 因此，依据文献[10]建立基本对齐模型：

$$P(m_k|\widetilde{CNE}_s,\widetilde{ENE}_t)=\frac{\exp\left[\sum_{f=1}^{F}\lambda_f h_f(m_k,\widetilde{CNE}_s,\widetilde{ENE}_t)\right]}{\sum_{m_k'}\exp\left[\sum_{f=1}^{F}\lambda_f h_f(m_k',\widetilde{CNE}_s,\widetilde{ENE}_t)\right]} \tag{3}$$

我们共采用 3 个特征计算基本对齐：意译特征、音译特征、共现特征. 具体概率计算根据文献[2]，下文进行简要介绍.

3.1.1　意译特征

采用 IBM-1 模型的概率，假设 $\widetilde{CNE}_s$ 包含 I 个词 $\tilde{c}_1,\cdots,\tilde{c}_i,\cdots,\tilde{c}_I$，$\widetilde{ENE}_t$ 包含 J 个词 $\tilde{e}_1,\cdots,\tilde{e}_j,\cdots,\tilde{e}_J$. 意译特征表示如下：

$$\begin{aligned}&h(m_k,\widetilde{CNE}_s,\widetilde{ENE}_t)=\\&\quad\log P_{ts}(\widetilde{CNE}_s|\widetilde{ENE}_t)+\log P_{ts}(\widetilde{ENE}_t|\widetilde{CNE}_s)\end{aligned} \tag{4}$$

其中，$P_{ts}(\widetilde{CNE}_s|\widetilde{ENE}_t)$ 和 $P_{ts}(\widetilde{ENE}_t|\widetilde{CNE}_s)$ 分别表示英-中、中-英的翻译概率.

3.1.2　音译特征

先将英文实体 $\widetilde{ENE}_t$ 音译为 $\widetilde{ENE}_t$，$\widetilde{ENE}_{tl}=\arg\max_{\widetilde{ENE}_{tl}}P_{tl}(\widetilde{ENE}_{tl}|\widetilde{ENE}_t)$，再通过 DICE 系数 $Dice(\widetilde{CNE}_{py},\widetilde{ENE}_{tl})$ 表示和中文实体拼音 $\widetilde{CNE}_{py}$ 和英文实体音译结果 $\widetilde{ENE}_{tl}$ 的相近度. 因为存在字符直接转换为拼音的音译方式（比如中文人名翻译成英文），所以还要考虑中文实体拼音和英文实体 $\widetilde{ENE}_t$ 的相似性 $Dice(\widetilde{CNE}_{py},\widetilde{ENE}_t)$. 最后的音译特征函数取以上两种相似性的最大值：

$$\begin{aligned}h(m_k,\widetilde{CNE}_s,\widetilde{ENE}_t)=\max(&Dice(\widetilde{CNE}_{py},\widetilde{ENE}_{tl}),\\&Dice(\widetilde{CNE}_{py},\widetilde{ENE}_t))\end{aligned} \tag{5}$$

3.1.3　共现特征

共现特征表示的是汉英实体在整个语料库中的对应频率.

$$h(m_k,\widetilde{CNE}_s,\widetilde{ENE}_t)=$$

$$\frac{count(\widetilde{CNE}_s,\widetilde{ENE}_t)}{\sum count(*,\widetilde{ENE}_t)}+\frac{count(\widetilde{CNE}_s,\widetilde{ENE}_t)}{\sum count(\widetilde{CNE}_s,*)} \tag{6}$$

其中，$count(\widetilde{CNE}_s,\widetilde{ENE}_t)$表示$\widetilde{CNE}_s$和$\widetilde{ENE}_t$在一个句子翻译对中同时出现的次数，而$count(\widetilde{CNE}_s,*)$和$count(*,\widetilde{ENE}_t)$分别表示中文实体$\widetilde{CNE}_s$和英文实体$\widetilde{ENE}_t$在语料库中出现的次数.

依据以上的特征函数，根据式(3)，我们使用GIS(Generalized Iterative Scaling)算法[11]来训练基本对齐模型的模型参数λ_f. 经过适当的转换，GIS算法可以用来处理实数值特征. 我们采用由Och开发的YASMET①来执行训练.

基本对齐计算是许多文献普遍采用的方式，在本文作为我们的对齐基准系统. 通过增加更多的特征，只能使对齐本身的正确率提高，但无法修正已有的识别错误，因而不能根本上解决实体识别对实体对齐的影响.

3.2 修正对齐计算

由式(2)可以看出，在基本对齐计算后，我们再通过式$P(\langle a_k,type_k\rangle|m_k,ChnS,EngS)$进行修正对齐计算，获得修正的对齐$a_k$及其类别$type_k$. 为引入双语实体识别与对齐的结合点，我们在修正对齐中引入双语实体翻译对a_k的内部词对齐L_k.

$$\begin{aligned}
&P(\langle a_k,type_k\rangle|m,ChnS,EngS)\\
&=P(a_k|type_k,m_k,ChnS,EngS)\cdot\\
&\quad P(type_k|m_k,ChnS,EngS)\\
&=\sum_{L_k}P(a_k,L_k|type_k,m_k,ChnS,EngS)\cdot\\
&\quad P(type_k|m_k,ChnS,EngS)\\
&=\sum_{L_k}P(L_k|a_k,type_k,m_k,ChnS,EngS)\cdot\\
&\quad P(a_k|type_k,m_k,ChnS,EngS)\cdot\\
&\quad P(type_k|m_k,ChnS,EngS)\\
&\approx\sum_{L_k}P(L_k|CNE_k,ENE_k,type)\cdot\\
&\quad P(a_k|type_k,m_k,ChnS,EngS)P(type_k|m_k)
\end{aligned} \tag{7}$$

假设对于各个类别，上式中的$P(type_k|m_k)$平均分布，因此可以被忽略.

$$\begin{aligned}
&P(a_k|type_k,m_k,ChnS,EngS)\\
&=P(a_k,OtherTokens|type_k,m_k,ChnS,EngS)\\
&=P(ChnS,EngS|type_k,m_k,a_k,OtherTokens)\times\\
&\quad P(a_k,OtherTokens|type_k,m_k)/\\
&\quad P(ChnS,EngS|type_k,m_k)\\
&=P(ChnS,EngS|\\
&\quad type_k,m_k,CNE_k,ENE_k,OtherTokens)\times\\
&\quad P(CNE_k,ENE_k|OtherTokens,type_k,m_k)\times\\
&\quad P(OtherTokens|type_k,m_k)/\\
&\quad P(ChnS,EngS|type_k,m_k)
\end{aligned} \tag{8}$$

其中，$OtherTokens$表示除实体以外在汉英句子$ChnS$和$EngS$上的标注，已知$P(ChnS,EngS|type_k,m_k,CNE_s,ENE_t,OtherTokens)=1$，而且上式中的$P(OtherTokens|type_k,m_k)$和$P(ChnS,EngS|type_k,m_k)$与每个实体翻译对的候选项$\langle CNE_s,ENE_t\rangle$无关，所以在不会影响式(1)最后结果的情况下，我们将$P(a_k|type_k,m_k,ChnS,EngS)$替换为$P(CNE_k,ENE_k|OtherTokens,type_k,m_k)$，假设$CNE_s$和$ENE_t$相互独立，推导如下：

$$\begin{aligned}
&P(CNE_k,ENE_k|OtherTokens,type_k,m_k)\\
&=P(CNE_k,ENE_k|type_k)\\
&=P(CNE_k|type_k)P(ENE_k|type_k)
\end{aligned} \tag{9}$$

其中的$P(CNE_k|type_k)$和$P(ENE_k|type_k)$分别是汉英实体的类别模型，为实体识别提供了汉英单语序列上的信息，本文称其为单语序列上的实体置信度，根据文献[1]介绍的命名实体识别的隐马模型，我们可以分别计算汉英序列的实体类别概率$P(CNE_k|type_k)$和$P(ENE_k|type_k)$，相当于引入单语实体识别的概率信息.

另一方面，式(7)中的$P(L_k|CNE_k,ENE_k,type_k)$称为双语序列上的实体置信度，它表示汉英双语序列上实体翻译对的生成，为实体识别提供双语信息. 因此，双语序列上的实体置信度体现了双语实体识别和对齐的融合，是下文介绍的重点. 假定$CNE_k=c_1^I=c_1\cdots c_i\cdots c_I$，表示该中文实体包含$I$个字，$ENE_k=e_1^J=e_1\cdots e_j\cdots e_J$表示该英文实体包含$J$个词. 由于中文分词存在一定的错误，特别是包含音译词的实体，分词问题尤为困难，因此，我们考虑汉英实体包含的词对齐时从英文词e_j出发，并且只考虑英文实体所包含的实词，忽略of，for等虚词. 我们定义(c_i,e_j)表示c_i和e_j互为翻译，于是，实体翻译对a_k的内部词对齐可以表示为$L_k=\langle(c_i,e_j)\rangle_{n=1}^{N}$，实际上，$c_i$和$e_j$的对应形式包括意译(Translation，TS)和音译(Transliteration，TL)两种方式. 因此L_k进一步定义为$L_k=\{\langle(c_i,e_j)\rangle_{n=1}^{N},\delta\}=\{TS=\langle(c_x,e_x)_{ts}\rangle_{x=1}^{n_1},TL=\langle(c_y,e_y)_{tl}\rangle_{y=1}^{n_2},\delta\}$，表示该内部对齐包含$n_1$个意译对应$(c_x,e_x)_{ts}$和$n_2$个音译对应

① http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/YASMET.html

$(c_y,e_y)_{tl}$，$N=n_1+n_2$. 根据第 2 节的介绍，实体意译或音译的比重与类别密切相关，可区分度比较大. 于是我们定义内部词对齐中意译对齐个数所占的比值 $\delta=\frac{n_1}{n_1+n_2}$ 来表示意译比重. 这样我们在双语序列的实体置信度中又引入了意译比重 δ.

$$\begin{aligned}&P(L_k|CNE_k,ENE_k,type)\\&\cong P(\langle c_i,e_j\rangle_{n=1}^{N},\delta|c_1^I,e_1^J,type_k)\\&\approx\prod_{n=1}^{N}P(c_i|e_j,type_k)\times P(\delta|type_k)\end{aligned}\tag{10}$$

$P(c_i|e_j,type_k)$ 在不同类别的双语实体语料库中训练获得. 同样，在不同类别的双语实体库中根据频率统计可以获得 $P(\delta|type_k)$.

综上所述，通过单语序列上和双语序列上的实体置信度计算，完成实体翻译对的修正对齐. 修正对齐体现了双语实体识别与对齐的交互. 最后，基于 3.1 节基本对齐计算与本节修正对齐计算的综合，从所有的实体候选翻译对中搜索最优的实体对齐结果.

3.3　搜　索

实体候选项的建立：以实体识别系统获得的汉英实体结果为基准，建立一个滑动窗，可以逐次向内缩减（中文 1～4 字，英文 1～2 词）和向外扩展（中文 1～4 字，英文 1～4 词），建立一系列汉英实体的候选项. 通过组合它们，得到双语实体翻译对的候选项，根据式(2)，令每个候选实体翻译对 a_k 的分值为

$$\begin{aligned}Score(a_k)=&\log[P(\langle a_k,type_k\rangle|m_k,ChnS,EngS)]+\\&\log[P(m_k|\widetilde{CNE}_s,\widetilde{ENE}_t)]\end{aligned}\tag{11}$$

然后由一种柱搜索（beam search）[12-13] 算法获得最优的实体对齐 A^*，柱搜索算法每次保留 N 个最好的假设，N 表示柱宽度（beam width），通过调整 N，可以近似地获得全局最优对齐结果. 每一个假设的生成过程如下：

1. 针对一个汉英双语句子对，通过汉英识别系统获得识别结果后，基于识别结果（例如，$\widetilde{CNE}_s$，$\widetilde{ENE}_t$）建立滑动窗，产生实体侯选项（例如，CNE_k，ENE_k），再将所有可能的实体对位 $a_k=(CNE_k,ENE_k)$ 构成一个候选对位的集合 Aligned-Pairs，并初始化该双语句子的实体对齐假设 H 为空；

2. 根据式(11)的计算，将所有实体候选实体对按降序排列；

3. 选取一个和当前假设没有边界重叠的候选项 $a_k=(CNE_k,ENE_k)$，由式(2)可以看出，在获得实体翻译对 a_k 的同时，也获得相应的类别 $type_k$，一起放入假设 H 中，相当于对已有假设进行扩展，得到一个新的当前假设；

4. 重复步 3 直到实体对齐假设 H 不能再继续扩展为止.

每个实体对齐假设就是该汉英双语句子对的一种实体对应结果. 我们采用评价函数来估计实体对齐假设. 根据式(2)，实体对齐假设的评价函数定义为

$$Score(H)=\sum_{k=1}^{H}score(a_k)\tag{12}$$

基于一定的柱宽度，对所有假设进行评价后，我们可以获得最优的实体对齐结果，然后回溯得到最优实体对齐结果中的所有实体翻译对.

4　实验设计及分析

为了验证本文提出的双语实体交互式对齐模型的有效性，我们进行了以下实验，分别测试了它对汉英文命名实体的识别和对齐的影响. 我们从 LDC 机构发布的汉英文新闻语料库（LDC2005T06）中抽取了 300 对汉英文句子翻译对作为我们的测试集（抽取的限制条件为：每一个句对中的中文句子或英文句子至少包含一个实体）. 其中，中文句子的长度平均是 58 个字；英文句子平均包含 24 个词. 通过人工标注其中的命名实体以及汉英实体间的对应，作为实体识别和对齐实验的标准答案. 评估标准采用“正确率（Precision，P）”，“召回率（Recall，R）”和“F 值（F-score，F）”.

4.1　汉英命名实体的识别与对齐基准系统

首先，我们分别采用我们实验室开发的多知识源融合的中文实体识别系统[14] 和公开开放的基于 CRF 模型的英文实体识别系统（Mallet 工具包①）作为实体识别的基准系统. 汉英实体的识别基准系统分别识别出 685 个中文实体和 732 个英文实体，其不同类别的实体识别性能如表 1 和表 2 所示.

表 1　中文实体的识别基准系统

	性能		
	P/%	R/%	F/%
人名	84.67	90.21	87.35
地名	91.82	90.24	91.02
机构名	85.42	82.75	84.06
综合	87.52	88.75	88.13

表 2　英文实体的识别基准系统

	性能		
	P/%	R/%	F/%
人名	79.35	85.96	82.52
地名	86.17	81.66	83.85
机构名	83.34	80.12	81.70
综合	83.12	83.58	83.34

① http://mallet.cs.umass.edu/index.php/Main_Page

观察表 1 和表 2,我们发现,在所有类别中,汉英文人名的识别正确率都是最低的,主要原因在于大量地名和机构名简称被错误识别成人名.此外,由于音译词和分词的影响,一些中文人名不能被完整识别出来.根据语料分析,虽然英文具有首字母大写标志的优势,英文实体的边界易于识别,但是其类别判断存在很多错误,因而总体而言,中文实体识别的基准系统要优于英文实体识别的基准系统.

基于识别基准系统得到的汉英文实体,我们采用基本对齐计算作为我们的对齐基准系统.其中,4 个特征的训练语料来源于汉英命名实体翻译对语料(LDC2005T34)和汉英双语新闻语料(LDC2005T06),并采用 GIZA++工具包[15]训练意译和翻译概率.在搜索过程中,我们采用柱宽度 $N=5$ 进行搜索.

汉英实体对齐基准系统得到的对齐正确率是 66.24%,由于对齐基准系统直接基于识别结果,汉英实体的识别错误在对齐过程中混合扩大,很大程度上影响对齐的效果,导致对齐正确率比较低.

4.2 汉英命名实体交互式对齐系统

基于对齐基准系统,双语实体交互式对齐系统还进行了修正对齐.修正对齐包括单语序列上的实体置信度计算(简称单语修正对齐)和双语序列上的实体置信度计算(简称双语修正对齐).我们采用汉英实体识别基准系统标注汉英双语语料(LDC2005T06),然后在该标注语料上训练得到单语序列上的实体置信度概率;双语序列上的实体置信度概率的训练基于汉英文实体语料库(LDC2005T34)中的人名、地名和机构名翻译对.表 3 给出对齐基准系统和交互式对齐系统在整体对齐性能上的比较.

表 3 不考虑类别的对齐性能比较

模型	性能		
	P/%	R/%	F/%
对齐基准系统(基本对齐)	72.12	76.83	74.40
修正对齐	74.45	79.52	76.90
基本对齐+单语修正对齐	73.32	79.13	76.11
基本对齐+双语修正对齐	77.19	82.44	79.73
交互式对齐系统(基本对齐+修正对齐)	79.07	84.26	81.21

表 3 比较了不同情况下的实体对齐性能,可以看出,交互式对齐系统,即基本对齐和修正对齐的结合,可以获得最好的性能.同时,引入双语修正对齐的性能要优于引入单语修正对齐的性能,可见在双语修正对齐中引入意译比重和实体类别约束的优势.

交互式对齐系统不仅获得修正后的实体翻译对,还包括汉英实体统一的实体类别,但由于对齐基准系统不能保证每个实体翻译对的类别统一,因此,我们采用以下打分方式进行不同类别的对齐结果比较:一个实体翻译对中,如果汉英实体的类别都判断正确,给 1.0 分;如果汉英某一个实体的类别判断错误,给 0.5 分;如果汉英实体的类别都判断错误,给 0 分,然后将所有实体翻译对的得分相加,作为类别判断的分值.测试语料中实际包含 192 个人名翻译对,363 个地名翻译对和 122 个机构名翻译对.表 4 给出了对齐基准系统和交互式对齐系统在类别判断上的分值比较.

表 4 对齐性能在类别判断上的比较

	对齐基准系统	交互式对齐系统
人名	156	181
地名	335.5	348
机构名	99.5	117

交互式对齐通过单语序列上的实体置信度(见式(9))和双语序列上的实体置信度(见式(10))对实体类别 $type_k$ 进行了重新判断.相对于实体识别结果来说,实体类别的重新判断融合了单语序列信息和双语序列信息(包括意译比重和实体类别的关系 $P(\delta|type_k)$ 以及实体类别的约束),对实体类别的判断更加有效.从表 4 我们可以看出,交互式对齐与对齐基准系统相比较在类别判断上有明显的优势,也就是在对齐的同时纠正了实体类别.使每种类别判断普遍提高了十几个分值.证明了交互式对齐对实体类别的修正作用.

交互式对齐系统中的修正对齐计算有助于修正实体的边界和类别,因而在提高实体对齐性能的同时也提高了实体识别的性能.表 5 和表 6 分别给出汉英文实体交互式对齐后的实体修正结果.实验结果表明了该交互式对齐模型辅助实体识别的有效性.

表 5 双语实体交互式对齐模型修正后的中文实体

	性能		
	P/%	R/%	F/%
人名	89.12	91.93	90.50
地名	91.15	93.46	92.29
机构名	87.98	85.26	86.60
综合	89.51	90.05	89.78
识别基准系统	87.52	88.95	88.13

表 6 双语实体交互式对齐模型修正后的英文实体

	性能		
	P/%	R/%	F/%
人名	84.78	89.23	86.95
地名	86.94	85.76	86.35
机构名	85.68	82.95	84.29
综合	85.80	86.52	86.16
识别基准系统	83.12	83.58	82.84

根据表 1 和表 5 的中文实体的识别性能比较，我们发现中文人名、地名和机构名的识别性能（F 值）通过双语交互式对齐修正后分别提高了 3.15%、1.27%和 1.54%. 同样，比较表 2 和表 6 的英文实体的识别性能，英文人名、地名和机构名的识别性能（F 值）通过双语交互式对齐修正后分别提高了 3.97%、1.97%和 2.08%.

由表 5 和表 6 可以很明显地观察到汉英文实体的整体识别性能都有了比较大的提高. 虽然由于一些中文人名被错误地修正为地名，造成中文地名的正确率略有下降，但其 F 值仍有所增加. 特别要指出的是，在所有类别中，汉英文人名通过交互式对齐修正后的识别性能提高得最多，主要由于中文人名的边界和英文人名的类别得到了很好的修正. 此外，通过统计发现汉英文实体的识别错误通过交互式对齐的修正后，分别相对降低了 15.1%和 22.6%.

综上所述，交互式对齐模型和普通对齐模型最大的区别体现在：(1) 它通过实体识别和对齐的交互，具有修正识别错误的能力，在提高识别性能的同时也提高了对齐性能；(2) 如果单语序列的实体识别准确率比较低，将直接影响到双语实体基本对齐. 基本对齐的性能也将比较低，因为基本对齐是直接基于实体识别结果的. 但在交互式对齐系统中，候选项是在单语实体识别结果的边界缩放基础上建立的，而且实体类别在交互式对齐中还需要重新判断. 所以单语实体识别的准确率对交互式对齐的性能没有直接的影响. 通过 4.1 节的数据显示，汉英文实体的识别错误率还比较高，由于新词和不同领域的影响，识别问题尤为困难. 在此基础上的双语实体对齐如果无法修正识别结果，即使采用更多、更好的对齐特征，也无法从根本上提高对齐性能. 通过实体识别与实体对齐的结合，有效引入双语对齐信息对实体识别的辅助，使对齐过程具有修正实体识别的能力，为实体识别和对齐性能的提高开辟了新的途径. 本文的双语实体交互式对齐模型是双语实体识别和对齐相结合的初探，还有待进一步的研究.

5　结束语

在自然语言处理系统中，命名实体识别是一项基本任务，也是一个难以完善解决的问题. 中文实体的识别由于分词、外来词的影响，实体边界尤其难以识别；而英文实体的类别错误判断问题更为突出. 命名实体的识别性能直接影响实体的对齐性能. 已有的实体对齐系统融入多种特征主要用于提高对齐本身的性能，但却不能实质解决实体识别错误的影响. 根据汉英文命名实体的特性分析，我们发现：一方面，如果实体识别引入双语对齐信息，可以使两种语言的特性互补，而提高识别性能；另一方面，如果实体对齐引入对实体识别结果的修正能力，可以降低识别错误的影响，而提高对齐性能. 因此，本文提出一种汉英双语实体交互式对齐模型. 包括基本对齐计算和修正对齐计算. 其中修正对齐计算根据单语和双语序列上的实体置信度，对实体的边界和类别重新判断，实现了双语实体识别与对齐的交互. 实验证明，本文提出的双语实体交互式对齐模型不仅将实体对齐性能（F 值）提高了 6.8(%)，而且分别有效提高了汉英文实体的识别性能. 这充分说明了双语命名实体识别与对齐结合的有效性.

参　考　文　献

[1] Huang Fei, Vogel S, Waibel A. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization//Proceedings of the 2003 Annual Conference of the ACL, Workshop on Multilingual and Mixed-language Named Entity Recognition. Sapporo, Japan, 2003: 184-192

[2] Al-Onaizan Y, Knight K. Translating named entities using monolingual and bilingual resources//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, PA, USA, 2002: 400-408

[3] Feng Donghui, Lv Yajuan, Zhou Ming. A new approach for English-Chinese named entity alignment//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004). Barcelona, 2004: 372-379

[4] Lee Chun-Jen, Chang Jason S, Jang Jyh-Shing R. Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources. ACM Transactions on Asian Language Information Processing (TALIP), 2006, 5(2): 121-145

[5] Moore R C. Learning translations of named-entity phrases from parallel corpora//Proceedings of 10th Conference of the European Chapter of ACL. Budapest, Hungary, 2003: 456-464

[6] Krishman Vijay, Manning Christopher D. An effective two-stage model for exploiting non-local dependencies in named entity recognition//Proceedings of the 44th Annual Meeting of ACL. Sydney, 2006: 1121-1128

[7] Ji Heng, Grishman Ralph. Collaborative entity extraction and translation//Proceedings of the International Conference on Recent Advances in Natural Language Processing. Borovets, Bulgaria, 2007: 231-238

[8] Chen Hsin-His, Yang Changhua, Lin Ying. Learning formulation and transformation rules for multilingual named entities//Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition. Sapporo, Japan, 2003: 1-8

[9] Berger Adam L, Della Pietra Stephen A, Della Pietra Vincent J. A maximum entropy approach to natural language processing. Computational Linguistics, 1996, 22(1): 39-72

[10] Och Franz Josef, Ney Hermann. Discriminative training and maximum entropy models for statistical machine translation//Proceedings of the 40th Annual Meeting of the ACL. Philadelphia, PA, USA, 2002: 295-302

[11] Darroch J N, Ratcliff D. Generalized iterative scaling for log-linear models. Annals of Mathematical Statistics, 1972, 43(5): 1470-1480

[12] Tillmann, Christoph, Hermann Ney. Word reordering and dynamic programming beam search algorithm for statistical machine translation. Computational Linguistics, 2003, 29(1): 97-133

[13] Koehn, Philipp. Pharaoh: A beam search decoder for phrase-based statistical machine translation models//Proceedings of the Association for Machine Translation in the Americas (AMTA). Washington, DC, USA, 2004: 81-89

[14] Wu Youzheng, Zhao Jun, Xu Bo. Chinese named entity recognition model based on multiple features//Proceedings of the HLT/EMNLP. Vancouver, BC, Canada. 2005: 427-434

[15] Och Franz Josef, Ney Hermann. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003, 29(1): 19-51

CHEN Yu-Feng, born in 1981, Ph. D., assistant researcher. Her research interests focus on natural language processing and machine translation.

ZONG Cheng-Qing, born in 1963, Ph. D., professor, Ph. D. supervisor. His research interests include machine translation, text classification, and syntactic parsing as well.

SU Keh-Yih, born in 1955, Ph. D., professor. His current research area isstatistical machine learning with application to natural language processing. He is now focusing on developing a high quality MT system for translating English/Chinese technical manuals.

Background

Most artificial intelligence systems, such as information extraction or machine translation system, adopt pipeline architecture, in which stages of analysis are arranged sequentially, with each stage using the results of prior stages and generating a single analysis that gets enriched by each stage. Unfortunately, each stage also introduced a certain level of error to the later stage. For example, in 'traditional' pipelined systems, name entity (NE) recognition is one of the first steps in the pipeline, NE errors heavily affect subsequent stages, and error rates are often compounded by later stages, which is evident in NE alignment stage.

In most prior work, NE recognition and alignment are processed sequentially, without recovering NE recognition errors. Therefore, the tagged errors are compounded from stage to stage. It is possible to consider how interactions between the NE recognition and alignment stage can be exploited to reduce the error rate. However, little research has been devoted to a joint NE alignment model with NE recognition.

This paper establishes an alignment derivation framework in order to incorporate NE recognition, and then proposes an interactive NE alignment model, which breaks a new path for NE recognition and alignment, and makes much improvement on the performance of NE alignment, as well as NE recognition.

The research work described in this article has been partially supported by the National Natural Science Foundation of China under grant No. 60575043, and No. 60736014, National Key Technologies R&D Program of China under grant No. 2006BAH03B02, and National High Technology Research and Development Program (863 Program) of China under grant No. 2006AA010108-4.