基于双语平行语料的分层次命名实体抽取[1]

庞薇[1]，徐波[1,2]

（1. 中科院自动化所 数字内容技术研究中心，北京 100080；Email：wpang@hitic.ia.ac.cn

2. 中科院自动化所 模式识别国家重点实验室， 北京 100080；Email：xubo@hitic.ia.ac.cn）

文 摘： 本文设计实现了一种基于多模型分层次的从双语语料库中抽取命名实体对的方法。我们首先对命名实体识别。然后分层次抽取命名实体。第一层通过双语识别信息和对齐技术，利用意译模型和音译模型打分得到短命名实体。第二层用规则的方法合并短命名实体生成长命名实体对。实验显示，双语识别信息和对齐技术对于短命名实体的抽取效果很好，针对长命名实体抽取问题的合并规则也能在一定程度上抽取出长命名实体。

关键词： 命名实体，对齐，抽取，分层，音译

Name Entity Extraction Based on Multi-layer and Bilingual Comparable Corpora

PANG Wei[1], XU Bo[1,2]

(1. Digital Content Technology Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing, 100080; 2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100080)

Abstract: We proposed a multi-model and multi-layer based method of extracting Chinese-English named entities from bilingual comparable corpora. Firstly, the named entities are recognized. Then we use multi-layer based method to extract Chinese-English named entities. On the first layer, by using bilingual recognition information and alignment technique, many short named entities are recognized with translation cost and transliteration cost. On the second layer, we combine short named entities to long named entity according to different rules. The experiments show that bilingual recognition information and alignment technique are useful for short named entities extraction, while the combining rules can extract long named entities at a certain extent.

Key words: Name Entity; Alignment; Extraction; Multi-layer; Transliteration

1 引言

命名实体(Named Entity，NE)是标识某一特定实体的词或词组，主要包括人名、地名和组织机构名。由于中英文命名实体常常会出现集外词，因此用一个固定的命名实体词典很难满足命名实体的翻译。这样就提出了通过大规模语料提取中英文双语命名实体词典来辅助命名实体翻译的方法。由于新的命名实体不断涌现，使得收集双语命名实体词典比较困难，因此用对齐技术自动建立双语词典就显得特别重要。这种通过双语对齐提取命名实体对的技术不仅可以用于从海量的信息中直接提取命名实体的翻译。还可以为命名实体的直接翻译提供训练语料。同时命名实体对齐，对于自然语言处理(NLP)领域中的命名实体识别、多语言信息检索、问答系统等都非常有用。

[1]**基金项目：** 国家 863 项目资助（2006AA01Z194）面向网络应用环境的口语翻译关键技术与系统研究

作者简介： 庞薇（1980——），女，博士研究生，主要研究方向为统计机器翻译。徐波（1966—），男，研究员，博士生导师，研究方向为语言识别，机器翻译，中文信息处理等。

通常命名实体对齐先要分别在两种语言中进行命名实体识别，然后在识别结果的基础上产生候选翻译对，再过滤得到对齐结果，比如：Huang[1]提出基于一个线形组合的多特征代价函数的最小化来抽取中英命名实体翻译等价对。有时识别和对齐也会结合在一起进行，比如：Mooer [2]把命名实体识别和对齐（英语-法语）同时进行，使用了一系列逐步递进的代价模型。这种方法很大程度上依赖于语言信息，如在两种语言中都出现的字串和大写字母标志，这对于不属于同一语系的两种语言是不适用的。当然也有不进行切分直接在原句子中得到候选翻译的。比如：Feng [3]使用了最大熵方法进行中英命名实体对齐。首先自动识别出了英语命名实体，为了避免错误传播，直接在中文文本上抽取候选的中文命名实体，没有对中文进行词切分和标注，然后利用最大熵模型计算对齐。

我们采用了最常用的在命名实体识别结果之上进行对齐的方法，这种抽取双语命名实体翻译对的方法分为如下 4 个主要步骤：

1) 分别对两种双语文本进行命名实体识别。

中文我们采用了自动化所模式实验室的一套命名实体识别系统[4]。针对汉语命名实体识别的难点，这套命名实体识别系统提出了基于多特征相融合的汉语命名实体识别模型。英文识别系统训练和测试采用的是网上公开的工具包 CRF++2。这是一个基于条件随机场(Conditional Random Fields, CRFs)的英文命名实体识别工具包。

2) 生成意译对齐候选，用意译打分过滤得到意译命名实体对

3) 生成音译对齐候选，用音译打分过滤得到音译命名实体对

4) 合并意译和音译命名实体，得到命名实体双语词典

本文的内容组织如下：第 2 节介绍了命名实体对齐和抽取的具体步骤及打分方法；第 3 节描述了命名实体合并的方法；第 4 节对分层次的命名实体抽取进行实验验证并分析；第 5 节是本文的总结和对以后工作的展望。

2 命名实体对齐和抽取

2.1生成意译命名实体对

对于大规模的双语语料，在分词标注和命名实体识别之后我们采取的策略是通过词对齐提供锚点，通过滑动窗寻找命名实体候选对再生成意译命名实体对。步骤如下：

1) 设双语语料含有 m 个中文单词和 n 个英文单词。中英文词交叉生成 m*n 个候选词对。
2) 计算这 m*n 个词对的词对齐得分，高于阈值的词对中，含有命名实体的词对加入命名实体候选中，不含有命名实体的词对加入词对齐词典中。
3) 如果命名实体候选为空则结束，否则在命名实体候选对中选择意译得分最高的候选对。
4) 以当前命名实体词对中的英文词为锚，滑动窗宽度为 2，得到 x 个一对多命名实体候选，以当前命名实体词对中的中文词为锚，滑动窗宽度为 2，得到 y 个多对一命名实体候选。在这 x+y 个命名实体候选对中得到得分最高的命名实体对（NEe, NEc）。如果它的打分高于设定的阈值则移除命名实体候选队列中的中英文词（NEe,*）和（*, NEc），并转 3。否则结束。

[2] http://chasen.org/~taku/software/CRF++/

2.2 生成音译命名实体对

在得到意译命名实体对和词对齐结果后，对于未对齐的词生成音译命名实体对。

1) 利用词性标注结果和词对齐结果，得到可能为音译对齐的中英文候选词。这些词满足两个条件：1、这些词不在意译命名实体对中 2、这些词不在词对齐词典中。
2) 设双语语料含有 a 个未对齐中文单词和 b 个未对齐英文单词。以这些候选词为锚点，滑动窗为 3，得到 x 个中文音译候选词组和 y 个英文音译候选词组，交叉生成 x*y 个候选词组对。
3) 计算这 x*y 个候选音译得分，选择意译得分最高的候选对（NEe, NEc）。如果它的打分高于设定的阈值则移除命名实体候选队列中的中英文词（NEe,*）和（*, NEc），并转 3。否则结束

在这里我们将打分分为了两个层次，首先用意译模型为候选命名实体短语对打分，如果得分超过阈值则表示命名实体之间是意译，如果得分很低则很有可能命名实体之间是音译。由于意译和音译中命名实体的形式差别很大，因此我们采用不同的滑动窗范围。我们可以看到一般命名实体的意译中一对多或者多对一的情况相对较少，而且即使存在一对多或者多对一的情况，它们的对应短语长度差距也不大，因此把滑动窗的长度设的比较小，来减少噪声，降低计算量。而在音译中，由于中英文切分的不同出现一对多、多对一或多对多的现象较多，因此我们把滑动窗的长度设为 3。希望能包含更为完整的音译命名实体对。

2.3 意译打分

本节我们用最大熵模型对候选命名实体短语对打分来得到命名实体对齐。首先选取合适的特征函数建立最大熵模型，并进行模型训练。然后用训练好的模型对每对候选翻译对进行预测。再用模型预测的概率值对所有候选翻译对打分。

对于命名实体词对齐候选的打分，我们采用了当今比较通用的四个特征[5] [6]即：两个方向的基于频率的短语概率（见公式(1)）和两个方向的词汇化概率（见公式(2)）：

$$p(f \mid e) = \frac{N(f,e)}{\sum_{f} N(f,e)} \tag{1}$$

$$lex(f \mid e,a) = \prod_{i=i_1}^{i_2} \frac{1}{\left|\{j \mid (i,j) \in a\}\right|} \sum_{\forall (i,j) \in a} p(f_i \mid e_j) \tag{2}$$

第一个是基于频率的短语概率，N(f,e)表示 f，e 共现的频率。第二个词汇化概率是用 IBM 模型 1 进行计算。其中的翻译分数 $p(f_i \mid e_j)$ 由 GIZA++[3]工具在训练语料上训练得到。

我们用最小错误率的方法来调整各个特征之间的权重，用 Venugopal [7]的基于 C++的训练工具进行最小错误率的训练。

2.4 音译打分

音译模型的打分与翻译模型打分类似。它是将拼音序列看作中文短语，将英文字母序列看作英文短语；将字母看作短语中的单词。这样就可以利用 GIZA++训练出两个方向的音译概率。

[3] http://www.fjoch.com/GIZA++.html

1) 字母音译打分

定义英文人名串含有 L 个字母 $e_1e_2\cdots\cdots e_l$，中文拼音串含有 J 个字母 $f_1f_2\cdots\cdots f_j$，通过 IBM_1 模型估计人名对的字母翻译概率。由于在音译中中英文字母的对应多为顺序对应，因此对应字母在词中的位置相对比较接近。在字母对应概率中加入顺序特征强调位置相近的字母之间的对应得分。其中 F 和 E 表示当前字母在中英文字母串中的位置。

$$\mathrm{Pstrsliorder}(e\mid y)=\frac{1}{L^{J}}\prod_{j=1}^{J}\sum_{l=1}^{L}\left[p(f\mid e)*\frac{1}{|F-E|+1}\right] \tag{3}$$

字母音译打分:

$$\mathrm{Cstrsliorder}=-\left[\log(\mathrm{Pstrsliorder}(e\mid y))+\log(\mathrm{Pstrsliorder}(y\mid e))\right] \tag{4}$$

2) 辅音音译打分

从实际的中英文词典中我们可以观察到英文中的辅音和中文中的声母在读音上起到的区分作用是相当明显的。英文辅音包括除去五个元音（a,o,u,i,e）之外的所有音节。中文声母则包括（z,x,c,b,s,d,f,g,h,j,l,q,w,r,t,y,p）。我们将训练词典中的这些字母组成一个声母辅音词典,训练一个声母辅音模型加入音译模型。通过 IBM1 模型估计人名对的声母辅音翻译概率, 其中 S 和 C 表示当前字母在中英文字母串中的位置。:

$$\mathrm{Pshy}(c\mid s)=\frac{1}{M^{N}}\prod_{n=1}^{N}\sum_{m=1}^{M}\left[p(s\mid c)*\frac{1}{|S-C|+1}\right] \tag{5}$$

辅音音译打分:

$$\mathrm{Cshy}=-\left[\log(\mathrm{Pshy}(c\mid s))+\log(\mathrm{Pshy}(s\mid c))\right] \tag{6}$$

我们将这两个打分 Cstrsliorder，Cshy,结合成音译打分 Ctrsli。

3 合并命名实体

之前的对齐都是以词为锚点，窗口设定较小，并且翻译方法统一为音译或者意译，因此得到的命名实体词对比较短。而很多机构名等命名实体是由多个词通过音译和意译混合翻译而成的。为了得到这些命名实体，我们加入了合并模块。合并主要通过一些规则实现。通过分析双语命名实体识别结果,我们发现长命名实体多出现在机构名中。英文语料中的机构名通常都区分大小写，所以英文识别结果相对较好。识别结果中有很多连续的命名实体标注，它们可以组成一个长命名实体串。我们将合并规则定义为命名实体与长命名实体中出现频率高的非命名实体的合并。然后通过合并后的命名实体的对齐范围确定长命名实体对，如果对应的命名实体满足我们的抽取条件则作为一对长命名实体抽取出来。同时我们注意到命名实体对齐过程中非常容易产生混淆的是一些虚词比如中文的“的，之”等，和英文的“of, the”等。由于它们并不表示确实的意思，我们把这些虚词归纳为一个虚词词典，把它们忽略不计。我们的合并过程如下：

1) 找到中英文句中符合合并规则的命名实体词串，（$x_m\cdots x_n$），得到它们中所有实词的对齐点($y_i\cdots y_j$)。y_i 是左边界，y_j 是右边界。

2) 当 y_i 和 y_j 之间所有非虚词的对齐点都在 x_m 到 x_n 之间时，把（$x_m \cdots x_n$）和（$y_i \cdots y_j$）作为一对长命名实体抽取出来。

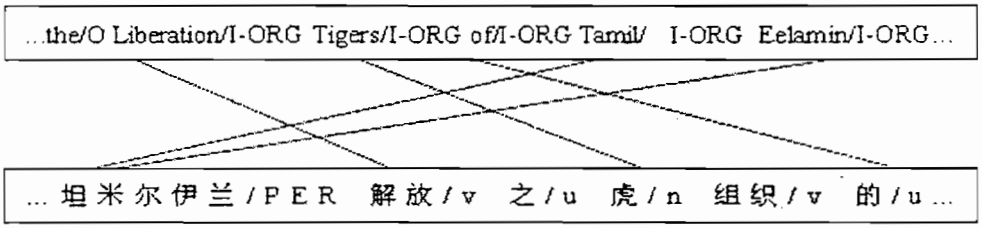

图 1：中英文机构名识别后的对齐实例

中英文机构名识别后的对齐结果如图 1 所示。在忽略了虚词 of 的对齐之后我们得到了一对短语“Liberation/I-ORG Tigers/I-ORG of/I-ORG Tamil/ I-ORG Eelamin/I-ORG”和“坦米尔伊兰/PER 解放/v 之/u 虎/n”。这对短语满足我们抽取的条件，因此被抽取出来作为一条长命名实体对加入命名实体词典中。

4 实验

4.1 实验设置

1)我们通过 LDC 命名实体词典训练得到意译打分和音译打分的概率模型。本文共用到 LDC 命名实体词典中的 6 个文件，分别是：“ldc_propernames_people_ce_v1.beta”和“ldc_whoswho_international_ce_v1”，这两个文件中是人名对，含有 523093 个人名对；“ldc_propernames_place_ce_v1.beta”文件中是地名对，含有 276382 个地名对；“ldc_orgs_intl_ce_v1.beta”，“ldc_propernames_org_ce_v1.beta”和“ldc_propernames_industry_ce_v1.beta”文件中是机构名对，共含有 92587 个机构名对。

由于我们的命名实体翻译主要是应用于 NIST 评测中的翻译，因此测试集由 NIST 评测的测试集中抽取。我们随机抽取出含命名实体较多的 NIST 评测句对 100 句，在自动标注和命名实体识别的基础上人工纠正识别结果并得到命名实体对。从这 100 个测试句对中共抽取出命名实体对 302 个，其中人名 93 个，地名 137 个，机构名 72 个。

实验采用传统的精确度、召回率和 F 值作为评测指标。

4.2 命名实体的抽取及合并

本节我们比较加入合并结果后的命名实体词典与直接以词对齐为锚点扩展得到命名实体词典的结果。结果如表 1 所示。

表 1：命名实体的抽取和合并

	精确度	召回率	F 值
人名	85.67%	59.26%	71.92%
合并后人名	87.55%	66.45%	74.90%
地名	82.67%	49.56%	61.97%
合并后地名	89.35%	55.78%	68.68%
机构名	41.22%	32.98%	36.64%

合并后机构名	70.87%	49.23%	56.62%

从表…中可以看到，对于人名抽取，合并后和合并前提高并不大，说明抽取出的人名地名长度较短，可以直接通过音译或者意译对齐得到。合并规则对于机构名抽取的影响最大，机构名中长词比例非常高，我们统计发现机构名的平均长度是 2.6 词。而且其中很大一部分都是由音译和意译混合翻译得到的，因此直接通过短词和一种对齐打分很难找到长的对齐对，只有通过后期的合并处理后才能抽取出比较正确的机构名。

4.3 音译模型的影响

为了研究命名实体对齐中音译模型的影响，我们将几个音译模型加入系统，观察系统性能的变化。实验结果罗列在表 2 中，意译四个概率用 P4 表示，字母音译打分用 Cstrsli 表示，加入顺序特征的字母音译打分用 Cstrsliorder 表示，辅音音译打分用 Cshy 表示。

表 2：多模型打分对命名实体抽取的影响

	正确率	召回率	F 值
P4	57.39%	48.74%	52.71%
P4+Cstrsli	68.26%	60.64%	64.22%
P4+Cstrsliorder	70.56%	62.45%	66.26%
P4+Cstrsli+Cshy	73.25%	68.47%	70.78%

从上表中可以看出多个音译模型的加入对于抽取效果都有一定的贡献。

4.4 单语命名实体识别与双语修正后命名实体识别结果的比较

为了研究双语修正后命名实体识别结果对单语命名实体识别结果的修正效果，我们比较了中文命名实体识别结果，英文命名实体识别结果和双语命名实体识别结果。实验结果罗列在表 3 中。由于命名实体识别的标注都是以词为单位，因此这里的评分以词为单位，不进行命名实体合并。比如：“Liberation/I-ORG Tigers/I-ORG of/0 Tamil/I-ORG Eelamin/I-ORG”中 5 个词分别标注，他的正确率为 0.8。

表 3：单语与双语命名实体识别结果比较

	正确率	召回率	F 值
中文识别结果	92.74%	59.50%	72.79%
英文识别结果	69.37%	73.50%	71.38%
对齐后中文识别结果	85.92%	83.15%	84.51%
对齐后英文识别结果	89.71%	83.77%	86.64%

从上表中可以看出，当用对齐结果修正单语命名实体识别结果后中英文的 f 值都取得了一定程度的增长。为了使翻译时覆盖到更多的命名实体，本文设置的意译和音译的阈值都较低，因此召回率提高较多但是正确率有所下降。但是总体来看 f 值还是有了很大程度的提高，说明命名实体识别的结果有了一定的改进。

4.5 与混合打分方法的比较

这个实验将我们的方法和 [Feng 2004] 中直接对长命名实体计算混合翻译得分的方法相比

较，测试一下我们的方法在分开考虑音译意译翻译对齐后再合并的方法是否有改善。

表 4：与混合打分方法的比较

	正确率	召回率	F 值
Feng	45.64%	41.28%	43.35%
本文方法	73.25%	68.47%	70.78%

实验结果如表 4 所示，从表中容易看出，本文的方法不论是在精确度、召回率还是在 F 值上都比[Feng 2004] 的方法有明显提高，在 F 值上本文的方法高出 27 个百分点。我们分析[Feng 2004]的方法中对于短命名实体比如人名，单词地名等由于采用混合模型，主要是融合了音译和意译两种评分，它们之间的相互影响在一定程度上降低了各自的区分度。而对于需要用到混合模型的多词地名和机构名，其长度较长，因此对于候选长度的限定较小，就产生了非常多的候选，为相似打分增加了很多的噪声，也很难得到完全准确双语命名实体对。因此我们本章采用的基于短命名实体对齐，再通过合并生成长命名实体的方法取得的结果就比较好。

5 总结

本文提出了一种从双语语料库中自动抽取命名实体的方法。我们的抽取主要分成两个部分：命名实体识别和命名实体对齐与抽取。针对命名实体识别的问题，我们提出了不同的解决方案。通过双语识别信息和对齐技术得到短命名实体，然后用规则的方法合并短命名实体生成长命名实体对。实验显示，双语识别信息和对齐技术对于短命名实体的抽取效果很好，针对长命名实体抽取问题的合并规则也能在一定程度上抽取出长命名实体。

参考文献

[1] Fei Huang, Stephan Vogel and Alex Waibel. 2003. Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-feature Cost Minimization. In Proceedings of the 2003 Annual Conference of the Association for Computational Linguistics (ACL'03), Workshop on Multilingual and Mixed-language Named Entity Recognition, July, 2003

[2] R.C.Mooer. Learning Translations of Named-Entity Phrases from Parallel Corpora, EACL — 2003.Budpaest, Hungary 2003.

[3] Dong-Hui Feng, Ya-Juan Lv, Ming Zhou. 2004 . A New Approach for English-Chinese Named Entity Alignment. Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, Jul. 2004.

[4] 吴友政。汉语问答系统关键技术研究，中国科学院自动化研究所博士毕业论文，2006。

[5] Philipp Koehn. Pharaoh: A Hierarchical Phrase-Based Model for Statistical Machine Translation. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. 2005.

[6] Och, F.J., Ney, H,: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, Vol. 29, No. 1 (2003) 19-51.

[7] Ashish Venugopal, Stephan Vogel, Alex Waibel, "Effective Phrase Extraction from Alignment Models", In the Proceedings of ACL 2003, Sapporo, Japan.