基于单语语料的面向日语假名的日汉人名翻译对抽取方法 *

王东明，徐金安，陈钰枫，张玉洁

（北京交通大学计算机与信息技术学院，北京 100044）

摘要：命名实体的翻译等价对在跨语言信息处理中非常重要。传统抽取方法通常使用平行语料库或可比语料库，此类方法受到语料库资源的质量和规模的限制。在日汉翻译领域，一方面，双语资源相对匮乏；另一方面，对于汉字命名实体，通常使用汉字对照表；对于日语纯假名的命名实体，通常采用统计翻译模型，此类方法受到平行语料库的质量和规模的限制，且精度低下。针对此问题，本文提出了一种基于单语语料的面向日语假名的日汉人名翻译对自动抽取方法。该方法首先使用条件随机场模型，分别从日语和汉语语料库中抽取日语和汉语人名；然后，采用基于实例的归纳学习法自动获取人名实体的日汉音译规则库，并通过反馈学习来迭代重构音译规则库。使用音译规则库计算日汉人名实体之间的相似度，给定阈值判定人名实体翻译等价对。实验结果表明，提出方法简单高效，在实现系统高精度的同时，克服了传统方法对双语资源的依赖性。

关键词：机器翻译；命名实体；日语假名；归纳学习法；音译

中图分类号：TP391　　**文献标识码：**A

Research on Extracting Kana Names Translation Equivalents Based on The Monolingual Corpora

Abstract: Named entity translation equivalents play a critical role in cross-language information processing.The traditional method often based on large-scale parallel or comparable corpus, but it is limited to the size and quality of the corpus resources.In Japanese-Chinese translation field, the bilingual corpora resources relatively scarce and it is usually use the Chinese Hanzi and Japanese Kanji comparison table to deal with Chinese named entity.It is usually use statistical machine translation model to deal with the pure kana named entities.But the method is limited to the size and quality of the corpus resources and is inefficient.We propose a method based on the monolingual corpora. Firstly,using conditional random field model to extract Japanese and Chinese names from monolingual corpus.Japanese-Chinese transliteration role base is construct by using inductive learning method based on the instance.The rule base is iteratively reconstructed through feedback learning.The method used the role base to calculate the Chinese and Japanese named entity similarity and got the named entity translation equivalents.Experimental results shoe that the proposed method is simple and efficient ,which overcome the shortcoming that the traditional method have a severely dependency on bilingual resource.

Key words: machine translation;named entities;Japanese kana;inductive learning method;transliteration

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金 (61370130 和 61473294)；中央高校基本科研业务费专项资金（2015JBM033）；国家国际科技合作专项资助（2014DFA11350）

1 引言

命名实体是标识某一特定实体的词或词组，其主要包括人名、地名和组织机构名等[1]，是自然语言的重要信息载体，在机器翻译、信息检索、问答系统以及跨语言信息处理等研究领域至关重要。

双语命名实体翻译等价对是指来自两种不同语言的具有互译关系的命名实体对。在日语中，假名是表音文字。“假”即“借”，“名”即“字”。意即只借用汉字的音和形，而不用它的意义。在日汉机器翻译研究领域，由于日语纯假名人名所占比例约为 27%[2]，且日语纯假名多为欧美等国的外来语，此类实体对汉语而言，同属于外来语。比如：“伊莎贝拉”和“イサベラ”，来源于西班牙语的外来词“Isabel”。因此，面向日语纯假名的命名实体的日汉翻译对的自动获取，蕴含了外来语到日语和汉语的音译规则，而由于日语和汉语在构词规则和发音上的差异，增加了日语假名的日汉双语实体对自动抽取的技术难度。

双语命名实体翻译等价对的自动获取方法很多[3]，其中，最直接的方法是用机器翻译系统直接进行翻译，即利用已知的源语言命名实体，通过翻译直接得到目标语言对应的命名实体。邹波等[4]对英汉人名的音译方法进行了研究，详细比较了两种机器学习方法和两种统计机器翻译模型在英汉人名音译上的应用效果，实验结果表明虽然这些方法取得了一定的效果，但仍然有较大的改善空间。体现在基于纯统计的方法对英汉人名进行音译是远远不够的，需要其他技术手段来获取更好的音译结果。李婷婷等[2]对日本人名的识别和翻译做了研究，针对日本人名的翻译，将日本人名分为假名人名和汉字人名，汉字人名的翻译通过建立日本人名常用汉字翻译词典实现，而假名人名的翻译通过 Moses 系统训练翻译模型实现，实验结果表明对于汉字人名部分的翻译正确率达到了 100%，这是因为其所建立的“日本人名常用汉字翻译字典”质量很好，对日本人名常用汉字的覆盖率很高，而对于假名人名部分的翻译正确率只有 47.34%，这说明仅使用统计机器翻译的方法来处理假名人名效果是不理想的。

另一种方法是给定源语言的命名实体，通过网络挖掘辅助翻译的方法得到目标语言对应的命名实体[5]。此方法属于直译方法的一种扩展形式。近年来，互联网高速发展，其中的语料资源越来越丰富，很多研究者都在利用互联网的语料资源来提取命名实体等价对，Jiang L 等[6][7]利用音译模型和网络挖掘来得到目标语言对应的命名实体，首先利用音译模型生成一个候选翻译，继而利用音译信息配合网络挖掘获得更多的候选翻译，最后使用最大熵模型综合考虑源语言和候选的目标语言命名实体的各种特征，得到最终的结果。实验结果显示该方法取得了一定的成效。

第三种方法是从平行语料库或可比语料库里批量抽取命名实体翻译等价对，Huang 等[8]提出了一种基于多特征的最小代价的命名实体翻译对自动抽取方法，实验结果表明该方法对命名实体翻译等价对的抽取得了较好地效果，但该方法对平行语料库有较大依赖，大规模的双语资源相对匮乏，构建成本高。

第四种方法是利用汉字对照表和归纳学习方法从单语语料库中抽取命名实体翻译等价对[9][10]，此类方法对日汉汉字命名实体翻译等价对的抽取简单高效，有效解决了对日汉双语资源的依赖性。但是，该方法对日语纯假名的日汉实体等价对的抽取具有一定的局限性。

综上所述，传统方法中，基于机器翻译的方法对翻译系统的性能具有依赖性；基于双语语料库或可比语料库的方法，其性能受限于语料库的质量和规模。而基于日语和汉语汉字对照表或词典的方法，无法有效解决日语纯假名的实体的日汉翻译等价对的自动抽取。

为了解决上述问题，本文提出了一种基于单语语料的面向日语假名的日汉人名翻译等价对自动抽取方法。首先，该方法使用条件随机场模型，分别从日语和汉语语料库中抽取日语和汉语人名；然后，采用基于实例的归纳学习法[11]自动获取人名实体的日汉音译规则库，通过反馈学习来迭代重构音译规则库。然后，使用音译规则库计算日汉人名实体之间的相似度，给定阈值判定人名实体翻译等价对。实验结果表明，提出方法简单高效，抽取的假名人名翻译等价对正确率高，可以达 86%以上。本方法在实现系统高精度的同时，克服了传统方法对双语资源的依赖性。

本文的组织结构如下：第二章介绍归纳学习法；第三章详细描述本文提出的方法；包括基于条件随机场的单语命名实体识别、基于归纳学习法的规则获取、以及反馈学习和校正处理等。第四章，实验部分，先给出一种基于统计机器翻译模型的日语纯假名日汉翻译等价对的抽取方法，作为本论文的基线系统，然后给出实验结果和分析讨论。最后，给出结论和未来工作。

2 归纳学习法

归纳学习法由日本学者荒木健治等[11]提出，其基本思路主要包括两个方面，其一是对两个具有相似性的实例中的相同部分和差异部分进行递归式抽取以获取规则；其二是通过校正和反馈处理，对抽取的规则进行筛选，更新规则库。该方法通过归纳学习获取实例间的内在规则，确定字符串之间的对应关系，表 1 为从未知字符串抽取对应关系规则的的例子。

表 1. 从未知字符串抽取对应关系

输入 1		$\alpha\theta\sigma\underline{\psi\delta}\lambda\nu$
输入 2		$\beta\gamma\underline{\psi\delta}\mu\pi$
段 1	$\alpha\theta\sigma$	$\beta\gamma$
段 2	$\psi\delta$	$\psi\delta$
段 3	$\lambda\nu$	$\mu\pi$

表 1 的输入 1 和输入 2 存在着对应关系，以下划线的形式将其标出。随后，将两边的不同部分按照先后顺序对齐。其结果如表 1 所示，段 1、段 2、段 3 分别构成对应关系。两个字符串间不同部分的对应关系，除表 1 所示的顺序对应外还有可能是逆序对应关系。至于采用顺序对应还是逆序对应，将取决于所研究的具体问题，在本文中，基于如上所述的假名人名的特点，我们采用顺序对应。

按照同样的方法可以从段抽取出共同部分并将段分解为基元。从段中抽取基元的例子如表 2 所示。将段 1,2 中用下划线标注的共同部分作为基元 2 抽取出来，并将其两侧的不同部分分别看成基元 1 和基元 3。如此，通过分离共同部分和不同部分，可以得到 3 个基元。

表 2 从段中抽取基元

段 1	$\alpha\underline{\theta\sigma}$	$\tau\underline{\psi\Theta}$
段 2	$\underline{\theta\sigma}\pi\nu$	$\underline{\psi\Theta}\prod$
基元 1	α	τ
基元 2	$\theta\sigma$	$\psi\Theta$
基元 3	$\pi\nu$	$\prod$

因为可以通过组合的方式将基元还原成段，所以这 3 个基元就成了两个段的完全替代品。这种抽取方式通常还需要借助确定对应关系的经验法则。本手法基于实例分阶段地抽取异同部分，从而获取知识，是一种归纳学习的方法。

3 日语假名和汉语人名翻译等价对自动抽取方法

现有的命名实体翻译等价对抽取方法，通常使用平行语料库或可比语料库，因而，受限于双语语料库的质量和规模。本文提出的方法旨在突破此限制，并有效提高日语假名实体等价对的抽取精度。提出的方法的系统架构如图 1 所示。

首先，我们使用条件随机场模型（CRFs），分别从日语和汉语单语语料库中抽取日语和汉语人名实体集合，再将其转换成罗马字[12]音节列表和汉语拼音列表；然后，使用音译规则库计算日汉人名实体之间的相似度，得到相似度列表。针对相似度高的人名实体对实例，筛选出来，利用归纳学习法，通过反馈学习来获取新的人名实体的日汉音译规则，经过数次迭代重构，得到最终的音译规则库。并根据规则库，通过相似度计算获取双语实体等价对。

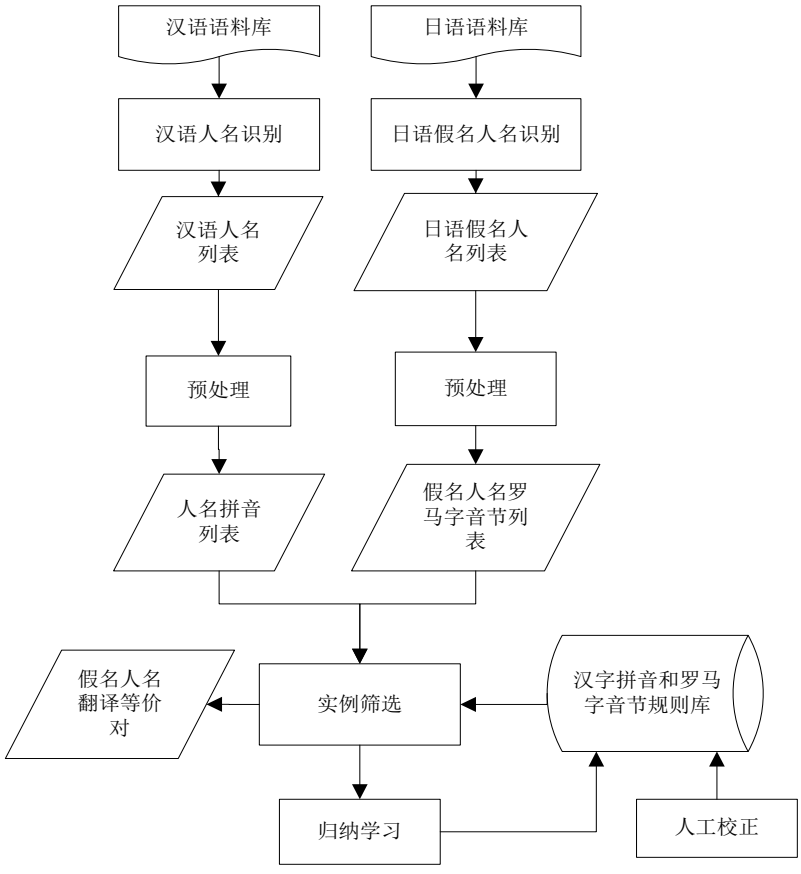
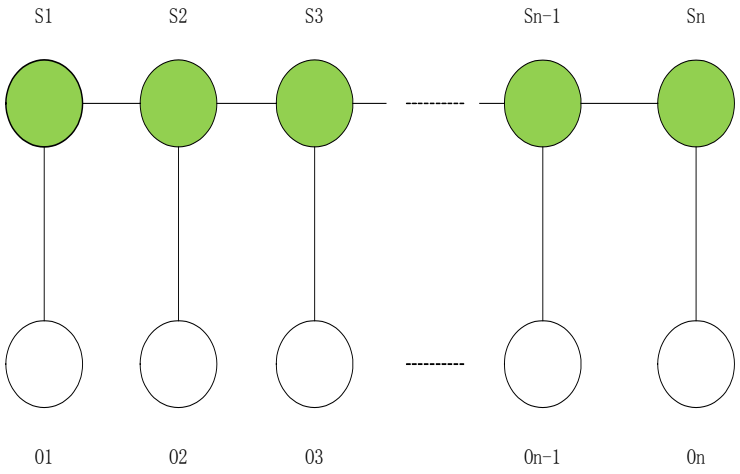


图 1 系统架构图

3.1 单语命名实体抽取

传统的单语命名实体抽取方法，主要包括基于规则、基于词典和基于统计的方法[13]。由于所需额外知识少，移植性好，基于统计的方法正成为研究者所使用的主流方法。基于统计的方法大量使用了机器学习领域的各种算法，主要包括，隐马尔可夫模型(HMM)、最大熵马尔科夫模型(MEMM)、条件随机场模型(CRFs)等。

条件随机场(Conditional Random Fields，简称 CRFs) 是一种用于序列数据标注的条件概率模型，由 Lafferty 等[14]于 2001 年提出，它是通过定义标记序列和观察序列的条件概率来预测最可能的标记序列的。条件随机场模型(CRFs)是近年来在序列标注问题中应用的比较多，也是效果最好的一种模型。它没有隐马尔可夫模型那样严格的独立性假设，因而可以容纳任意的上下文信息。同时，由于 CRFs 计算全局最优输出节点的条件概率，克服了最大熵马尔科夫模型和其它非生成的有向图模型所固有的标记偏置的缺点。CRFs 是在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率分布，而不是在给定当前状态条件下，定义下一个状态的状态分布。



□

□ 图 2 线链 CRFs 结构图

□ 条件随机场是以给定的观察值为条件，从而计算输出状态的概率的条件概率模型。其中最简单的 CRFs 是一个称为链图或线图的无向图(如图 2 所示)，称为线链 CRFs(linear-chain CRFs),也是最常用的一种条件随机场模型。

□ 假设 $O=o_1,o_2,.....,o_n$ 是一个长度为 n 的观察序列，线链 CRFs 的参数 $\Lambda=\{\lambda_1,\lambda_2,.....,\lambda_k\}$,则此模型输出词位序列 $S=s_1,s_2,.....,s_T$ 的条件概率为：

□

□ $$P_{\Delta}(S \mid O) = \frac{1}{Z_o} \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t)) \quad (1)$$

□ 其中,Z_o是归一化因子，作用是确保所有可能的词位标记序列的条件概率和为 1，其定义如下：

□ $$Z_o = \sum_{S} \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t)) \quad (2)$$

□ 公式(1)中的 f 通常是一个二值表征函数，用于表达上下文可能的语言特征，其定义为：

□ $$f_k(s_{t-1}, s_t, o, t) = \begin{cases} 1, \text{如果满足条件} \\ 0, \text{否则} \end{cases} \quad (3)$$

□ CRFs 模型通过特征函数能够整合任何特征，包括可观察序列O在时刻t时由当前字及其上下文组成的字串序列特征，以及上下文中隐含词位的转移特征$s_{t-1} \to s_t$，λ_k。λ_k是一个训练过程中需从训练语料中学习的参数，表示的是相应的特征函数 $f_k(s_{t-1}, s_t, o, t)$ 的权重，其取值范围可以是$-\infty$到$+\infty$。对于一个由公式(1)给定的条件随机场模型，对任意的一个输入字串，其最可能的标记序列可以由以下公式求出：

□ $$S^* = \arg\max_{S} P_{\Delta}(S \mid O) \quad (4)$$

可以使用 Viterbi 算法对公式(4)进行解码，从而求出使得 $P\Delta(S \mid O)$ 最大的标记序列。

本文所采用的命名实体识别工具是实验室独自研发的基于CRFs的单语命名实体识别工具，该系统选取较为复杂的特征模板进行识别，性能较高，其从中日双语的维基百科数据库的单语语料库中抽取汉语人名和日语假名人名。

3. 2 归纳学习

日语假名属于外来词，大多是通过音译得到的，而其对应的汉语人名，同样属于外来词，也是通过音译而来的。因此，从发音规律上来讲，彼此之间存在一定的对应关系[15]。为了探索和发现其规律，可以将假名人名对应的汉字和日语假名分别转换为对应的中文拼音和罗马字，如“路易斯恩里克”和“ルイスエンリケ”分别转换为“lu|yi|si|en|li|ke”和“ru|i|su|e|n|ri|ke”，

在这里我们将中文拼音以每个汉字拼音进行分词，而日语假名对应的罗马字以其发音的音节进行分词，通过分析，我们可以简单地得到这样的规则对，“lu~ru”，“yi~i”，“si~su”，“en~e|n”，“li~ri”，“ke~ke”，如表 3 中所示，值为“1”的即是汉字拼音和假名罗马字的规则对。我们希望在更多的汉日假名人名对中得到更多这样的规则对，以用来识别我们未知的汉日假名人名翻译等价对。

表 3 汉字拼音和假名罗马字的对应图

	lu	yi	si	en	li	ke
ru	1					
i		1				
su			1			
e				1		
n				1		
li					1	
ke						1

我们将抽取得到的日汉假名人名分别转换成罗马字音节序列和汉语拼音序列。对于音译而来的假名人名，它的汉语发音和日语发音都是顺序的，不会出现逆序的情况。因此，给出如下处理步骤：

1. 假设实体等价对之间的首尾发音音节具有对应关系。如“lu|yi|si|en|li|ke”和“ru|i|su|e|n|ri|ke”，它们的首尾有“lu~ru”，“ke~ke”的对应关系。

2.为了确定一个汉字拼音对应几个罗马字音节，给定窗口设置，在一定范围内获取候选汉字拼音和罗马字音节之间的对应关系规则对，例如设窗口为 2，我们可以从上例中抽取到“lu~ru”、“lu~ru|i”、“ke~ke”和“ke~ri|ke”这些规则对。

为了提高归纳学习法的效率，本文采用一定规模的既有假名人名翻译等价对作为学习数据，得到候选的汉字拼音和罗马字音节的规则对的初始集合，然后使用根据获取规则的权重，设定阈值过滤部分低置信度的规则获取高置信度的汉字拼音和罗马字音节规则表，之后，根据相似度计算，获取实体等价对，再进行校正处理和反馈学习，通过迭代生成新的音译规则，

并更新规则的权重。表 4 给出了一个规则库实例。

表 4 规则库实例

拼音序列	罗马字音节序列	规则库
lu|yi|si|en|li|ke	ru|i|su|e|n|ri|ke	lu~ru，ke~ke
yi|si|en|li	i|su|e|n|ri	lu~ru，ke~ke，yi~i，ri~li
si|en	su|e|n	lu~ru，ke~ke，yi~i，ri~li，si~su，en~e|n

本方法在利用既有假名人名翻译等价对作为学习数据构建初始汉字拼音与罗马字音节的规则库之后，我们从日汉双语的单语语料库中使用基于 CRFs 的单语命名实体识别工具分别进行假名人名的识别，得到两个单语的假名人名集合，通过预处理得到分好“词”的假名人名拼音列表和罗马字音节列表，使用上述初始的汉字拼音和罗马字音节规则库，计算候选的假名人名翻译等价对的相似度，相似度计算如公式(5)所示。

$$Score_{de}(na_c, na_j) = \frac{\sum_{i=1}^{k}(c_i + j_i)}{m+n} \qquad (5)$$

其中，nac 为假名人名拼音序列，naj 为假名罗马字音节序列，k 为候选假名人名对在规则库中找到的规则对的个数，ci 为规则对中拼音个数，ji 为规则对中罗马音节个数，m 为假名人名拼音个数，n 为假名罗马字音节个数。

如对“lu|yi|si|en|li|ke”和“ru|i|su|e|n|ri|ke”，假设规则库中存在“lu~ru”，“ke~ke”，“yi~i”，“ri~li”这四个规则，则其相似度 $Score_{de} = \frac{4+4}{6+7} = \frac{8}{13} = 0.61$。

然后，选取相似度大的候选假名人名翻译等价对进行归纳学习、人工校正和反馈处理，通过迭代处理得到新的候选规则，对规则给定阈值，获取可信度大的候选规则更新规则库，再利用更新的规则库迭代计算相似度，直到收敛为止。对于一对多或多对一的规则情况，本文采用了计算其之间的编辑距离来进行过滤。针对所产生的新规则，根据语言学知识进行判定和校正处理，以提高规则的正确性。

4 实验

4.1 基线系统

本文采用之前在命名实体翻译等价对中比较常见的统计机器翻译系统作为基线系统。如文[2]中所述的方法，采用基于短语的统计机器翻译实现日语假名人名到中文的翻译。具体使用 Moses[16]训练翻译模型来实现假名人名的翻译，基线系统实验数据共包括 13032 对日汉假名人名对，实验中将数据分为训练集、开发集、测试集三部分，其中测试数据与下文实验中数据一致。基线实验所用的实验数据如表 5 所示。

表 5 基线系统实验数据

	训练集	开发集	测试集
人名数量	11032	1000	1000

如文[2]中所述，实验结果评价指标不用 BLEU 值来估计，直接用翻译准确率如公式(6)来测试，表 6 是测试结果，这也与文[2]中的实验结果相近。

$$\text{正确率} = \frac{\text{翻译正确的人名数}}{\text{总的人名数}} \qquad (6)$$

表 6 实验一结果

	测试人名数	正确翻译数	正确率
Moses	1000	433	43.3%

4.2 实验设置

4.2.1 实验语料

实验中所用的单语语料库来源于中日双语的维基百科数据库，本实验从日语单语篇章和汉语单语篇章中使用本实验室基于 CRFs 的命名实体工具进行识别，共识别汉语人名 88203 个，日语人名 73322 个，并从中抽取 13032 个假名人名日语条目，并手工进行词对齐的校正工作，作为实验的数据。

4.2.2 实验工具

实验中用到的工具包括，基于 CRFs 的单语命名实体识别工具，由本实验室研究小组自

主开发，其他还有 GIZA++工具[17]，汉字转拼音工具[18]，假名转罗马字工具[19]等。

4.2.3 参数设定

实验中的参数设定，主要是指对规则抽取的阈值的设定，在迭代过程中，该阈值应逐渐放宽，否则，随着迭代次数的增加，难以获取新规则。然而在初始时，该阈值却不能选的较低，否则将导致规则库过冗余。另外，对于相似度的阈值设定，实验中我们取初值为 0.3，随着迭代的进行，我们进行动态的调整。

4.2.4 评价方法

实验结果指标采用准确率(P),召回率(R)和 F 值来作为评分标准，其中 P，R 和 F 的计算方式如公式 7,8 和 9 所示。

$$P = \frac{Num_{correct}}{Num_{\min ed}} \times 100\% \qquad (7)$$

$$R = \frac{Num_{correct}}{Num_{total}} \times 100\% \qquad (8)$$

$$F = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 P + R} \times 100\% \qquad (9)$$

其中，Numcorrect 为抽取到的正确的条目，Nummined 为抽取到的所有条目，Numttotal 为语料中存在的所有正确条目，当 β =1 时，表示准确率(P)和召回率(R)权重相同，就是一般所说的 F1 值。本文我们认为准确率和召回率同等重要取 β =1。

4. 3 实验结果

本实验从中日双语的维基百科数据库中的日语单语篇章和汉语单语篇章中，使用本实验室基于 CRFs 的命名实体工具进行识别，共识别汉语人名 88203 个，日语人名 73322 个，实验中基于 CRFs 的单语命名实体识别工具的识别效果分别如表 7 和表 8 所示。并从中抽取 13032 个假名人名日语条目，并手工对齐，作为实验的数据。其中训练数据 12032 个，测试数据 1000 个。

表 7 中文人名识别效果

P(%)	R(%)	F(%)
95.56	92.54	94.03

表 8 日语假名人名识别效果

P(%)	R(%)	F(%)
95.36	67.3	78.9

在本文中，我们提出了一种基于单语语料的面向日语假名的日汉人名翻译对自动抽取方法。采用基于实例的归纳学习法自动获取人名实体的日汉音译规则库，通过反馈学习来迭代重构音译规则库。使用音译规则库计算日汉人名实体之间的相似度判定人名实体翻译等价对。实验结果见表 9.

表 9 实验二结果

迭代数	P(%)	R(%)	F(%)
1	82.29	80.06	81.15
2	83.25	81.28	82.25
3	85.32	82.05	83.65
4	86.71	82.24	84.41
5	87.60	82.36	85.05
6	87.64	82.66	85.08

表 10 实验三结果

	P(%)	R(%)	F(%)
校正前	87.64	82.66	85.08
校正后	88.73	83.72	86.15

在表 9 中，在每次迭代对所取的相似度阈值不同，通过多次对比实验，采用贪心算法进行参数优化。对于每次迭代，对不同相似度阈值下的结果进行比较，取最优结果。实验中，

第一次迭代和第二次迭代的相似度阈值取 0.3.第三次迭代取 0.4，第四、五、六次迭代取 0.5。

由实验二的结果，我们可以看出，使用本文所提出的方法，从单语语料中抽取假名人名对的效果较之使用统计机器翻译系统的方法准确率提高了很多。例如我们使用机器翻译系统翻译“伊云尼斯域”并不能得到正确的结果“イワニセビッチ”，而用本文的方法容易抽取到这样的命名实体翻译对。随着迭代次数的增加，经迭代重构的规则库越完备，取得实验效果越好。实验证明所提方法简单高效。当然，由于语料的局限性，某些命名实体对也可能抽取不出来，例如“宽”和“クアン”，这种情况下我们可以通过对规则库进行人工校正来解决。由于抽取规则的不确定性，实验中，我们对规则库做了少量的人工校正，对于明显不符的规则如“dang~mu”，直接剔除，对于有稍许偏失的规则予以修正，对未能提取到的规则直接加入规则库，由表 10 中可以看到校正后，实验的效果会有所提高。

5 总结和未来工作

本文提出了一种基于单语语料的面向日语假名的日汉人名翻译对自动抽取方法。首先，该方法使用条件随机场模型，分别从日语和汉语语料库中抽取日语和汉语人名；然后，采用基于实例的归纳学习法自动获取人名实体的日汉音译规则库，通过反馈学习来迭代重构音译规则库。使用音译规则库计算日汉人名实体之间的相似度判定人名实体翻译等价对。实验结果表明，提出方法简单高效，在实现系统高精度的同时，克服了传统方法对双语资源的依赖性。我们下一步的工作，将考虑利用更多的特征，如词长度信息，编辑距离，所属文本的文体等特征来对单语语料中的命名实体翻译等价对进行自动抽取，同时，我们尝试采用规则获取和其他统计方法相结合的方法来解决此类问题。同时，使用本方法，我们还将对地名、组织结构名称等其他纯假名命名实体对的自动获取进行扩展。

参考文献：

[1] D. Bikel, S. Miller, R. Schwartz, etc. A high-performance learning name-finder.[C] // Proceedings of Applied Natural Language Processing,Washington DC:1997.

[2] 李婷婷，赵铁军，张春越. 基于统计的日本人名的识别和翻译[J]. 智能计算机与应用, 2012, 2(1) :4-7.

[3] 赵军. 命名实体识别、排歧和跨语言关联[J]. 中文信息学报，2009，23(2):3-17

[4] 邹波，赵军. 英汉人名音译方法研究[A]. 第四届全国学生计算语言学研讨会会议论文集[C],2008:24-30.

[5] Jenq-Haur Wang, Jei-Wen Teng, Pu-Jen Cheng,etc. Translating unknown cross-lingual queries in digital libraries using a web-based approach[C]//Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries.ACM,2004:108-116

[6] Jiang L,Zhou M,Chien L F,et al.Named entity translation with web mining and Tansliteration[C] //IJCAI.2007,7:1629-1634

[7] 蒋龙，周明，简立峰. 利用音译和网络挖掘翻译命名实体[J].中文信息学报，2007,21(1):23-28.

[8] Huang, F., Vogel, S., Waibel, A.: Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-Feature Cost Minimization[C] / / Proceeding of Association of Computational Linguistics, Sapporo,Japan:2003.

[9] 茹旷. 日汉双语命名实体对获取方法及其应用研究[D]. 北京交通大学，2014.

[10] Ru K,Xu J,Zhang Y,et al.A Method to Construct Chinese-Japanese Named Entity Translation Equivalents Using Monolingual Corpora [A] // Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg,2013:164-175

[11] 荒木健治，高橋祐治，桃内佳雄，栃内香次.帰納的学習を用いたかな漢字変換[C],電子情報通信学会論文誌 D-II、Vol.J79-D-II, No.3,1996:391-402.

[12] 罗晓莹. 日语假名罗马字标记法的历史及发展[J]. 郑州航空工业管理学院学报(社会科学版). 2014(06)

[13] 孙镇，王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术，2010,(6):42-47.

[14] John Lafferty, Andrew McCallum, and Fernando C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", . June 2001.

[15] 何功星. 日语中日汉人名的声调规则[J]. 科技信息，2011(17)

[16] http://www.statmt.org/moses/

[17] http://code.google.com/p/giza-pp/downloads/detail?name=giza-pp-v1.0.7.tar.gz

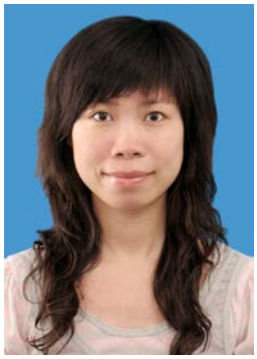
[18] http://www.aies.cn/pinyin.htm

[19] http://o-oo.net.cn/katakana-Roman.asp

作者简介：王东明（1985——），男，硕士研究生，主要研究领域为自然语言处理、统计机器翻译。Email:13120428@bjtu.edu.cn；徐金安（1970——），男，副教授，主要研究领域为自然语言处理和机器翻译。 Email: jaxu@bjtu.edu.cn；陈钰枫（1981——），女，副教授，主要研究领域为自然语言处理和机器翻译。Email: chenyf@bjtu.edu.cn；张玉洁（1961——），女，教授，主要研究领域为自然语言处理和机器翻译。 Email: yjzhang@bjtu.edu.cn。

 ：王东明

 ：徐金安

 ：陈钰枫

 ：张玉洁