北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2016.006

基于双语对齐的汉语–新蒙古语命名实体翻译

杨萍[1,2]　侯宏旭[1,†]　蒋玉鹏[1]　申志鹏[1]　杜健[1]

1. 内蒙古大学计算机学院，呼和浩特　010021; 2. 临汾职业技术学院计算机系，临汾　041000;
† 通信作者, E-mail: cshhx@imu.edu.cn

摘要　汉语—新蒙古语命名实体翻译在跨汉语–新蒙古语信息处理中具有重要意义，而直接使用机器翻译的方法不能达到满意的结果。针对上述问题，提出一种从汉语–新蒙古语平行语料中自动的抽取汉语–新蒙古语命名实体翻译对的方法。该方法只需对汉语端进行命名实体标注；然后基于双语 HMM 词对齐结果利用滑动窗口的方法抽取所有候选命名实体翻译对；最后基于融合 5 种特征的最大熵模型对所有候选翻译单位进行过滤，选取与汉语端命名实体相对应的置信度最高的新蒙古语命名实体翻译单位。实验结果表明，该方法优于基于 HMM 的方法，在对齐模型只是部分准确的情况下，也获得了较高准确率的汉语–新蒙古语命名实体翻译对。

关键词　命名实体；识别；翻译；双语对齐
中图分类号　TP391

Chinese-Slavic Mongolian Named Entity Translation Based on Word Alignment

YANG Ping[1,2], HOU Hongxu[1,†], JIANG Yupeng[1], SHEN Zhipeng[1], DU Jian[1]

1. College of Computer Science, Inner Mongolia University, Hohhot 010021; 2. Department of Computing, Linfen Vocational and Technical College, Linfen 041000; †Corresponding author, E-mail: cshhx@imu.edu.cn

Abstract　Chinese to Slavic Mongolian Named Entity Translation in cross Chinese and Slavic Mongolian information processing has a very important significance. However, using the machine translation method directly cannot achieve satisfactory result. In order to solve the above problem, a novel approach was proposed to extract Chinese-Slavic Mongolian Named Entity pairs automatically. Only the Chinese named entities need to be identified, then extracting all of the candidate Named Entity pairs using sliding window method based on HMM word alignment result. Finally filtering all of the candidate Named Entity translation unit based on Max Entropy Model integrated with four features, and choose the most probable aligned Slavic Mongolian NE of the Chinese NE is. Experimental results show that this approach outperforms HMM model, achieves high quality of Chinese-Slavic Mongolian Named Entity pairs with relatively high precision, even though sometimes the word alignment result is partially correct.

Key words　named entity; recognition; translation; bilingual word alignment

命名实体在人类语言中传递着非常重要的信息[1]。命名实体可以指出文档里“何人(何组织)……何时……何地……”等主要内容，因此找出它们是准确理解文档的基础。命名实体的识别在网络信息抽取、网络内容管理和知识工程等领域都具有非常重要的地位[2]。命名实体翻译对机器翻译、跨语言信息检索等多语言信息处理领域而言意义重大。因此，有很多学者致力于命名实体识别和翻译的研究。最早的命名实体翻译研究开始于英语和阿拉伯语之间，Al-Onaizan 等[3]使用音译模型以及词典查

国家自然科学基金(61362028)资助
收稿日期: 2015–06–07; 修回日期: 2015–08–18; 网络出版时间: 2015-09-30 12:38:18

找的方法进行英语和阿拉伯语之间的命名实体翻译。随后越来越多的命名实体翻译研究在不同的语言之间开展。Knight 等[4]和 KEITA Tsuji[5]进行了日语和英语命名实体翻译的研究。韩语和英语的命名实体翻译主要有 Lee 等[6]的工作。近些年来, 汉语和英语命名实体之间的翻译也受到越来越多的关注。Huang 等[7]提出基于多特征代价最小的自动抽取汉语–英语命名实体翻译对的方法。Wan 等[8]和 Feng 等[9]也分别提出不同的汉语–英语命名实体翻译方法。

近年来, 我国与蒙古国的经济、政治、文化交流日益深入, 对新蒙古语信息处理技术的发展起到极大的促进作用, 同时也提出更高的要求。在传统蒙古语的命名实体识别方面, 那顺乌日图等[10]采用基于规则的方法进行人名的自动识别, 召回率达到 89%、准确率为 86%。通拉嘎[11]采用最大熵的数学模型实现蒙古语人名自动识别系统, 封闭测试的 F 值为 89.61%。这些研究只针对传统蒙古语的人名识别, 未涉及传统蒙古语地名及机构名的识别。在新蒙古语的命名实体识别和翻译方面, 尚无相关论述。

采用音译或意译命名实体直接翻译的方法进行汉语–新蒙古语命名实体的翻译缺乏对命名实体自身组成结构以及上下文信息的考率, 必然会影响翻译结果。如果使用命名实体对齐的方法, 则需要命名实体的识别和命名实体间的对齐都被很好的处理。目前, 需要懂得新蒙古语言的工作者在语料上进行命名实体的标注, 工作量大、周期长; 新蒙古语语料规模相对于英语、汉语等其他语言规模尚小, 必然会影响新蒙古语命名实体识别的效果。在命名实体识别中的部分识别、识别错误等问题在对齐过程中不能被很好地纠正。

针对上述问题, 本文提出一种从只在汉语端标注了命名实体的汉语–新蒙古语平行语料中抽取汉语–新蒙古语命名实体翻译对的方法。我们先用 HMM 词对齐模型对双语语料进行对齐, 然后基于对齐模型, 利用相关短语抽取技术[12], 抽取出与汉语端相对应的新蒙古语端的候选命名实体翻译单位。用融合 5 种特征的最大熵模型对所有候选命名实体翻译单位进行过滤, 得到与汉语端命名实体最为匹配的新蒙古语端命名实体翻译单位。实验结果表明, 我们的实验结果优于 HMM 模型, 在语料库上得到的命名实体翻译对的正确率为 86.51%, 召回率为 87.32%, F 值为 86.91%。

1 词对齐模型

IBM 信源信道翻译模型[13]包括语言模型和翻译模型。其中, 翻译模型可建模为:

$$p(s|t)=\sum_{a} p\,(s,a\,|\,t)\,, \tag{1}$$

a 是一个表示源语言和目标语言句子中词和词对齐情况的隐含变量, $a=a_1\ a_2\ \ldots a_I$, 其中 a_I 表示源语言句子里第 I 个词对应的目标语言句子中词的位置。在一对句子的所有对齐方式中, 其训练对齐模型中的最大可能的对齐方式通常被称为最大近似对齐。

在 IBM 对齐模型中,

$$p(s,a|t)=\sum_{I=1}^{i} p(a_i|i,I)\times p\,(s_i|t_{a_i})\,, \tag{2}$$

在 HMM 对齐模型下, 用 Viterbe 算法实现最大近似对齐, 即对齐 a_I 满足

$$\hat{a}_I=\arg\max p_{a_I}\left(s,a_I|t\right), \tag{3}$$

$$p(s,a|t)=\sum_{i=1}^{I} p(a_i\,|\,a_{i-1},I)\times p(s_i|t_{a_i})\,, \tag{4}$$

这里, $p(a_i|a_{i-1}, I)$表示源语言句子当前词对齐位置 a_i 对前一个词对齐位置 a_{i-1} 的依赖关系, I 表示源语言的句长 $p(s_i|t_i)$表示词的翻译概率。

相比较于 IBM 词对齐模型, HMM 对齐模型考虑了当前词对齐位置 a_i 对前一个词对齐位置 a_{i-1} 的依赖关系, HMM 模型比 IBM 模型更有利于对平行语料库中的局部化现象进行有效地建模。因此, 我们在 HMM 词对齐结果上来抽取候选汉语–新蒙古语命名实体翻译对。

2 基于对齐模型的候选汉语–新蒙古语命名实体翻译对的抽取

本文命名实体翻译对的抽取经过 3 个步骤: 1) 汉语端命名实体的识别; 2) 基于词对齐模型, 生成与汉语端命名实体对应的新蒙古语端候选的翻译单位; 3) 对新蒙古语端的候选翻译单位进行置信度估计, 从中选出置信度最高的汉语–新蒙古语命名实体翻译对。

本文使用 CRF 模型进行汉语端命名实体识别。因为汉语命名实体识别不属于本文重点要讨论的内容, 本文不再赘述。下面重点介绍汉语–新蒙古语候选命名实体翻译等价对的生成和候选翻译等价对的置信度估计这两部分工作。

2.1　候选汉语–新蒙古语翻译对的生成

平行句对中，源语言句子 S_i 与目标语言句子 T_j 中词与词之间的对应情况可以用词对齐图表示。在图 1 中，叉线所在的单元表示由最大近似对齐得到的词对齐结果。在一个平行句对中，可以用一个四元组假设 $H_p(c_s, c_e, m_s, m_e)$来表示一个翻译等价对。其中，c_s 和 c_e 分别表示汉语命名实体的起始位置和结束位置；m_s 和 m_e 分别表示与汉语端对应的新蒙古语端候选翻译单位的起始位置和结束位置。例如，在图 1 中(2,3,2,4)就可以表示一个翻译等价对，即汉语端由词(s_2, s_3)组成的命名实体与新蒙古语端由词(t_2, t_3, t_4)组成的候选翻译单位对应。本文的翻译等价对抽取任务就是找出合适的汉语与新蒙古语之间的翻译对。

采用滑动窗口的方法从对齐图中找出与汉语端对应的新蒙古语端的所有候选命名实体翻译单位。如图 1 所示，如果(s_2, s_3)是汉语端的一个命名实体，那么图中粗线框选的所有对齐点所对应的新蒙古语端的词就构成了一个候选翻译单位。即 t_2, (t_2, t_3)和(t_2, t_3, t_4)就是与(s_2, s_3)对应的所有候选翻译单位。利用这样的方法可以产生较大数量的候选翻译等价单位，即使在对齐模型只是部分准确的情况下，依然可能抽取到正确的命名实体翻译对。

2.2　候选汉语–新蒙古语命名实体翻译对的置信度估计

考虑到最大熵模型可以很好地融合不同的特征，在此框架下对所有候选翻译对进行置信度估计。对于汉语端命名实体 $\mathrm{ne_c}$ 和与之对应的所有候选新蒙古语端命名实体 $\mathrm{ne_m}$，假设有 M 个特征方程 $H_m(\mathrm{ne_c}, \mathrm{ne_m})$, m=1, 2, … M, 对于每个特征函数，都有一个对应的模型参数 λ_m, m=1, 2, … M。汉语端与新蒙古语端命名实体对齐的概率可以被定义为式(5)[14]。

$$P(\mathrm{ne_m}|\mathrm{ne_c}) = P_{\lambda_1^M}(\mathrm{ne_m}|\mathrm{ne_c}) = \frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\mathrm{ne}_m, \mathrm{ne}_c)\right]}{\sum_{\mathrm{ne}_{m'}}\exp\left[\sum_{m=1}^{M}\lambda_m h_m[\mathrm{ne}_{m'}, \mathrm{ne}_c]\right]}。\quad (5)$$

选择出与汉语端命名实体对应的最有可能的新蒙古语端命名实体翻译单位，如式(6)[14]所示。

$$\widehat{\mathrm{ne_m}} = \arg\max_{\mathrm{ne_m}}\left\{P(\mathrm{ne_m}|\mathrm{ne_c})\right\} = \arg\max_{\mathrm{ne_m}}\left\{\sum_{m=1}^{M}\lambda_m h_m(\mathrm{ne_m}, \mathrm{ne_c})\right\}。\quad (6)$$

结合命名实体翻译的特点，我们采用 4 个特征：对齐一致性得分、翻译得分、共现得分、边界得分。下面对这 4 个特征分别详细介绍。

2.2.1　对齐一致性得分

任意一个汉语端的命名实体与它所对应的新蒙古语端的任何一个候选翻译单位，都在词对齐图中划分了一个范围。我们以这个划分是否与最大近似对齐中的对齐点一致来对候选翻译对进行对齐一致性置信度估计。对齐点 $A_p(x, y)$与 $H_p(c_s, c_e, m_s, m_e)$定义的划分一致是指这个对齐点所对应的源语言端词的位置与目标语言端词的位置均在 H_p 所划分的范围内。对齐点 $A_p(x, y)$与 $H_p(c_s, c_e, m_s, m_e)$定义的划分被认为不一致，当且仅当满足：

$$c_s \leqslant x \leqslant c_e \wedge (y < m_s \vee y > m_e),\quad (7)$$

$$m_s \leqslant y \leqslant m_e \wedge (x < c_s \vee x > c_e)。\quad (8)$$

对于每个 $H_p(c_s, c_e, m_s, m_e)$都包括一个与该划分一致的对齐点的集合和不一致的对齐点的集合。例如,在图 1 中，H_p(2,3,2,4)就包括与其一致的对齐点{(s_2, t_2), (s_3, t_3), (s_3, t_4)}和与其不一致的对齐点集合{(s_1, t_4), (s_4, t_4), (s_2, t_6)}。用式(9)计算任意一个 $H_p(c_s, c_e, m_s, m_e)$的对齐一致性得分。

$$\mathrm{Score_{ma}}(H_p(c_s, c_e, m_s, m_e)) = \frac{\mathrm{num(cons)}}{\mathrm{num(cons)} + \mathrm{num(incons)}}。\quad (9)$$

式(9)中，num(cons)和 num(incons)分别表示与四元假设 $H_p(c_s, c_e, m_s, m_e)$划分范围一致的对齐点的个数和不一致的对齐点的个数。在汉语—新蒙古语命名实体候选翻译对的四元假设的划分中，如果一致的对齐点越多，不一致的对齐点越少，则该翻译对的对其一致性得分就越高。

2.2.2　翻译得分

组成汉语命名实体中的词与组成新蒙古语命名实体的词之间的翻译概率对于考察汉语端命名实体

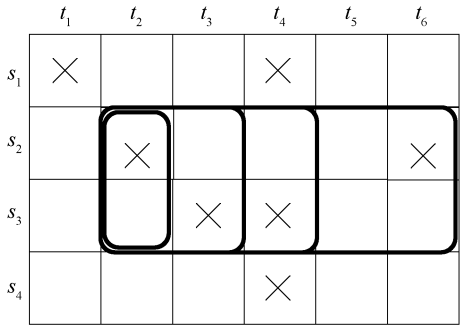

图 1　源端句子 S_i 与目标端句子 T_j 词对齐图

Fig. 1　Sample source S_i and target T_j alignment map

与新蒙古语端命名实体的相近程度具有非常重要的作用。假设汉语端命名实体由 s 个词组成 $ne_c=\{c_1, c_2, \dots c_s\}$，新蒙古语端候选命名实体翻译单位由 t 个新蒙古语词组成 $ne_m=\{m_1, m_2, \dots m_t\}$，则这个候选双语命名实体对的翻译得分可以由 c_i 与 m_j 之间的翻译概率计算得到:

$$\text{Score}_{\text{lex}}(H_p(c_s,c_e,m_s,m_e))=\sum_{j=1}^{t}\sum_{i=1}^{s}P(m_j\mid c_i)。\quad(10)$$

式(10)给出的是候选双语命名实体对中的词互译的概率。可以看出，该特征倾向于给含有词数更多的命名实体翻译单位以更高的分数。

2.2.3 语言模型得分

为了使与汉语端命名实体对应的新蒙古语端的翻译单位最大程度地符合新蒙古语的语法，在新蒙古语语料库上进行了语言模型的训练 LM(mn),对候选新蒙古语端命名实体翻译单位进行语言模型打分。如式(11)所示:

$$\text{Score}_{\text{lm}}(H_p(c_s,c_e,m_s,m_e))=p(m_s m_{s+1}\dots m_e)\approx p(m_s)p(m_{s+1}\mid m_s)\dots p(m_e\mid m_{e-1},m_{e-2})。\quad(11)$$

对应于汉语端同一个命名实体，在新蒙古语端包含词数较多的命名实体翻译单位倾向于获得更高的翻译得分，这样容易在新蒙古语命名实体翻译单位中引入一些多余的词。加入对语言模型得分的估计后，候选命名实体翻译单位中多余词的存在会使该翻译单位获得很低的语言模型得分，避免了翻译得分带来的偏差。例如，在未加入语言模型得分之前，我们获得“孔子学院–Күнзийн Институт улсын”的对应关系，包含多余的词“улсын”。但加入语言模型得分后，我们得到准确的命名实体翻译对“孔子学院–Күнзийн Институт”。

2.2.4 共现得分

假设汉语端命名实体与候选新蒙古语端的命名实体翻译单位在双语语料库中常常是同时出现的，那么它们为翻译等价对的可能性就非常大。从整个语料库中得到的知识可以作为对句对间局部对齐信息特征的一个有效补充。用式(12)计算源汉语端命名实体与候选新蒙古语端命名实体的共现得分:

$$\text{Score}_{\text{co}}(H_p(c_s,c_e,m_s,m_e))=\frac{\text{num}(ne_c, ne_m)}{\sum\text{num}(*, ne_c)}。\quad(12)$$

其中，$\text{num}(ne_c, ne_m)$是 ne_c 和 ne_m 共同出现的次数，$\text{num}(*, ne_c)$是 ne_c 出现的次数。

2.2.5 边界得分

新蒙古语命名实体词的开头字母是大写字母，这是新蒙古语命名实体的一个重要特征。这一特征对于新蒙古语命名实体边界的确定具有重要的作用。但在实际语料库中存在着部分不规范的现象，部分首字母应大写的命名实体词并未大写。为了尽量减少上述错误对计算边界得分的影响，我们不直接考察组成命名实体的首词或尾词是否为首字母大写。边界得分的计算是在该翻译单位中，首字母大写的词的个数占所有词的个数的比例，即:

$$\text{Score}_{\text{bd}}(H_p(c_s,c_e,m_s,m_e))=\frac{\text{num (CapWords)}}{\text{num (words)}},\quad(13)$$

其中，num(CapWords)指在新蒙古语命名实体翻译单位中，首字母是大写的词的个数，num(words)代表在该翻译单位中包括的所有词的个数。

2.2.6 基于最大熵模型的汉–新蒙命名实体候选翻译对的过滤

上面定义了 5 个特征函数。对于在汉语端标注出的每个命名实体，要计算与之对应的每个候选新蒙古语端命名实体翻译单位的特征分数，从而得到与汉语端命名实体对应的最佳的新蒙古语端翻译单位。根据式(5)，使用 MEM 建模工具 YASMET[①]进行最大熵模型的训练。由于没有汉语–新蒙古语命名实体翻译对的标准训练集，采用 bootstrapping[15]方法指导训练过程。首先在包括所有的候选汉–新蒙命名实体翻译对的训练集上对模型进行训练，根据训练得到的对各个候选翻译对的概率估计对初始训练集进行精简，得到剪裁后的训练集，并且对候选翻译对进行排序。反复进行上述步骤直至模型收敛或得到的实体翻译对变化不明显为止。

3 实验结果及分析

3.1 实验设置

为了验证本文提出的汉语–新蒙古语命名实体翻译方法的有效性，我们使用实验室整理得到的 12400 句对的汉语–新蒙古语平行语料，从中选取出 300 个汉–新蒙古语平行句对作为标准测试集(每个句对中至少包括一个命名实体翻译对)；并用人工标注出这 300 个句对中所有的汉语和新蒙古语命名实体，作为命名实体翻译对的标准答案。训练集和测试集中包括各类命名实体的数量如表 1 所示。

① http://www.fjoch.com/YASMET.html

表 1　训练集和测试集实体数目

Table 1　Number of named entities in training set and test set

实验数据集	句对数	实体总数	人名	地名	机构名
训练集	12100	18620	5680	6840	6100
测试集	300	628	177	330	121

使用基于 CRF 模型的汉语命名实体识别方法，在剩余的 12100 平行句对的汉语端进行汉语命名实体识别，并进行汉语–新蒙古语命名实体翻译对抽取的训练。各个实体类别的数目见表 1。

3.2　评价标准

假设 S^*是汉语端标注出的所有的命名实体的集合，S 是用本文的方法在 S^*基础上抽取得到的汉语–新蒙古语命名实体翻译对的集合，T 是双语语料中基于 S^*的所有的正确的命名实体翻译对。我们用准确率(P)、召回率(R)、F 值作为评价标准。

$$P=\frac{|S\bigcap T|}{|S|}, \tag{14}$$

$$R=\frac{|S\bigcap T|}{|T|}, \tag{15}$$

$$F=\frac{2PR}{P+R}。 \tag{16}$$

3.3　实验方法与结果

首先用实验室完成的基于 CRF 模型的汉语命名实体识别方法，对双语语料的汉语端进行命名实体的标注。采用 GIZA++工具包[16]训练得到从汉语–新蒙古语、新蒙古语–汉语单向最大近似对齐结果，并使用 GROW-DIAG-FINAL 算法[17]对两个方向的对齐文件进行合并，得到汉语与新蒙古语双向最大近似词对齐结果。然后用 SRILM①训练一个新蒙文端的 3-gram 语言模型。为了考察词切分对基本对齐以及命名实体翻译对抽取的影响，我们进行了两组实验：第一组对汉语端进行分词，训练汉语–新蒙古语双向词对齐，在此基础上，用本文提出的方法进行双语命名实体翻译对的抽取；第二组实验不对汉语端分词，只分成一个一个的字。实验得到的汉语–新蒙古语命名实体翻译对如表 2 所示，实验结果如表 3 所示。

表 2　汉语–新蒙古语命名实体翻译对示例

Table 2　Examples of Chinese-Slavic Mongolian named entity translation pairs

汉语命名实体	实体类型	新蒙古语命名实体
澳大利亚	LOC	Австрали
亚	LOC	Ази
欧	LOC	Европын
武文斌	PER	У Вэньбинь
东莞	LOC	Дүнгуаньд
北京语言大学汉语水平考试中心	ORG	Бээжингийн Хэлний Их Сургуулийн Хятад хэлний төвшний шалгалт авах төвөөс
包头	LOC	Бугат
布林贝赫	PER	Бүрэнбэх
江苏东大通信公司	ORG	Жянсү мужийн Дүнда холбооны компанийн
巴雅尔赛汉	PER	Баярсайхан
丁俊晖	PER	Дин Жюньхүйн
维信羊绒股份有限公司	ORG	Вэйшинь ямааны ноолуурын хувьцаат компанийн
中国国务院	ORG	БНХАУ-ын Төрийн Зөвлөл
内蒙古蒙牛乳业股份有限公司	ORG	Өвөр монголын Мэнню сүүний аж ахуйн ХК

① http://www.speech.sri.com/projects/srilm/

表 3　实验结果

Table 3　Experiment result

实验模型	准确率	召回率	*F* 值
HMM (未分词)	0.6516	0.6952	0.6727
HMM+MEM (未分词)	0.8651	0.8732	0.8691
HMM (分词)	0.6054	0.6134	0.6298
HMM+MEM (分词)	0.7837	0.8325	0.8111

在表 2 中, HMM 是直接在 HMM 对齐模型上抽取得到的汉语–新蒙古语命名实体翻译对的实验结果, 作为基线系统。HMM+MEM 指在 HMM 对齐模型上抽取汉语–新蒙古语候选命名实体翻译对, 再对候选翻译对进行融合 5 种特征的最大熵模型进行置信度估计, 选取置信度最高的命名实体翻译对。从实验结果可以看到, 无论是 HMM 还是本文方法, 不对汉语端进行分词, 抽取出的命名实体翻译对的 *F* 值都高于分词后的结果。最主要的原因是减少了分词错误对句对间词对齐以及命名实体翻译对抽取的错误传递。

实验表明, 本文选择用来刻画汉语–新蒙古语命名实体翻译对的特征, 对于命名实体翻译对的抽取是非常有帮助的。对齐一致性得分为命名实体翻译对的抽取提供了句对间的上下文信息。翻译得分指明了汉语端命名实体与候选新蒙古语端翻译单位的相近程度。语言模型得分使抽取到的新蒙古语端命名实体单位尽量符合新蒙古语语法。共现得分为命名实体翻译对的抽取提供了整个训练语料库中汉语词与新蒙古语词之间的共现知识。而边界得分充分考虑了新蒙古语命名实体词首字母大写的特性。

4　结束语

命名实体翻译中, 对称对齐的方法需要在源语言端与目标语言端都进行命名实体识别, 且在一端识别错误, 即使是另一端识别正确的情况下, 该错误也无法在对齐过程中纠正。目前, 可用于新蒙古语命名实体识别的标注语料规模尚小, 直接影响到新蒙古语命名实体的识别效果。针对上述问题, 本文给出一种只需在汉语端进行命名实体标注, 从汉–新蒙古语平行语料中抽取汉–新蒙古语命名实体翻译对的方法, 在 HMM 词对齐模型上抽取候选汉–新蒙古语翻译单位, 然后用基于最大熵模型对候选翻译对进行过滤, 最终得到质量较高的实体翻译对。实验表明, 本文方法与基于 HMM 的方法相比, 实验结果有了大的提高。本文抽取出的一些实体翻译对还有不正确的地方, 在下一步工作中, 可以考虑新蒙古语命名实体自身的语言特征, 并可以加入一些规则, 使得实验效果更好。

参考文献

[1] Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder // Proceedings of the Fifth Conference on Applied Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 1997: 194–201

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

[3] Al-Onaizan Y, Knight K. Translating named entities using monolingual and bilingual resources // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 400–408

[4] Knight K, Graehl J. Machine transliteration. Computational Linguistics, 1998, 24(4): 599–612

[5] Tsuji K. Automatic extraction of translational Japanese-KATAKANA and English word pairs from bilingual corpora. International Journal of Computer Processing of Oriental Languages, 2002, 15(3): 261–279

[6] Lee J S, Choi K S. A statistical method to generate various foreign word transliterations in multilingual information retrieval system // Proceedings of the 2nd International Workshop on Information Retrieval with Asian Languages (IRAL'97). New York, 1997: 123–128

[7] Huang F, Vogel S, Waibel A. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization // Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition—Volume 15. Stroudsburg, PA: Association for Computational Linguistics, 2003: 9–16

[8] Wan S, Verspoor C M. Automatic English-Chinese name transliteration for development of multilingual resources // Proceedings of the 17th International Conference on Computational Linguistics—Volume 2. Stroudsburg, PA: Association for Computational Linguistics, 1998: 1352–1356

[9] Feng D, Lü Y, Zhou M. A new approach for English-

Chinese named entity alignment // Proceedings of the Conference on Empirical Methods in Natural Language Processing Emnlp. Stroudsburg, PA, 2004: 372–379

[10] 那顺乌日图, 雪艳, 淑琴, 等. 蒙古文人名自动识别研究//语言计算与基于内容的文本处理: 全国第七届计算语言学联合学术会议论文集. 北京: 清华大学出版社, 2003

[11] 通拉嘎. 基于蒙古文语料库的人名自动识别[D]. 北京: 中央民族大学, 2013

[12] Venugopal A, Vogel S, Waibel A. Effective phrase translation extraction from alignment models // Proceedings of the 41st Annual Meeting on Association for Computational Linguistics—Volume 1. Stroudsburg, PA: Association for Computational Linguistics, 2003: 319–326

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

[14] Och F J, Ney H. Discriminative training and maximum entropy models for statistical machine translation // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 295–302

[15] Abney S. Bootstrapping // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 360–367

[16] Och F J, Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003, 29(1): 19–51

[17] Koehn P, Hoang H, Birch A, et al. Moses: open source toolkit for statistical machine translation // Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Stroudsburg, PA: Association for Computational Linguistics, 2007: 177–180