北京大学学报(自然科学版) 第 52 卷 第 1 期 2016 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 52, No. 1 (Jan. 2016)
doi: 10.13209/j.0479-8023.2016.001

基于字形与语音的音译单元对齐方法

刘博佳 徐金安† 陈钰枫 张玉洁

北京交通大学计算与信息技术学院, 北京 100044; † 通信作者, E-mail: jaxu@bjtu.edu.cn

摘要 为了解决仅采用基于语音或基于字形的音译方法造成的误差过大问题, 以汉英音译为主要研究对象, 运用统计与规则的理论思想, 提出融合基于语音和字形的音译单元对齐方法, 设计了 4 个实验, 与传统方法进行对比。实验结果显示, 该方法能够很好地提高机器音译的准确性。

关键词 机器音译; 对齐; N-gram 模型; 基于语音的音译方法; 基于字形的音译方法

中图分类号 TP391

Integrating of Grapheme-Based and Phoneme-Based Transliteration Unit Alignment Method

LIU Bojia, XU Jin'an†, CHEN Yufeng, ZHANG Yujie

School of Computer and Information, Beijing Jiaotong University, Beijing 100044; † Corresponding author, E-mail: jaxu@bjtu.edu.cn

Abstract In order to solve the errors caused by only using the pheneme-based method or the grapheme-based method, applying the theory of statistics and rules, this paper proposes a new method for transliteration unit alignment which integrates the two main transliteration methods. Four experiments are designed to compare with the traditional methods. Experimental results show that proposed method outperforms other methods in terms of performance in machine transliteration.

Key words machine transliteration; alignment; N-gram model; grapheme-based method; phoneme-based method

在自然语言处理应用中, 机器音译常被用于解决未登录词(out-of-vocabulary, OOV)的问题, 音译结果的准确度直接影响到实际应用[1]。对于采用不同字母表和发音系统的不同语系之间(如英语与汉语, 英语与日语, 英语与阿拉伯语等), 机器音译的难度往往很大。根据音译的方向, 可以分为正向音译 (forward-transliteration) 和反向音译 (backward-transliteration), 也可分为基于规则的方法和基于统计的方法。经过历年的发展, 音译的主流方法经历了从基于规则到基于统计的发展过程[1]。根据音译要素分类, 主要分为基于语音(phoneme-based)的音译框架[2]和基于字形(grapheme-based)的音译框架[3]。

基于规则的方法需要人工针对特定的语言对和音译方向建立音译规则[4]。Wan 等[4]提出从英文到中文的基于规则的正向音译方法, 该方法的思想被大量应用在规则音译系统中。蒋龙等[5]指出, 规则的音译框架采用跨语言的语音对应表, 这种方法的典型不足就是不能为表中的每一种对应提供一个概率值, 以便排序选择最优翻译。同时, 由于完备的规则系统需要完全通过手工撰写语言规则, 需要很大的人力投入, 且获取的规则不容易泛化。因此, 随着 NLP 领域的发展, 机器音译的方法逐渐向统计方法靠拢。

在基于统计方法的音译中, 经常使用对齐模型 IBM model 1-3 和 HMM[3]。GIZA++①是一个融合了 IBM model 1-5 和 HMM 模型的开源对齐工具。

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

国家自然科学基金(61370130, 61473294)、中央高校基本科研业务费专项资金(2014RC040)和国家国际科技合作专项(2014DFA11350)资助
收稿日期: 2015–06–18; 修回日期: 2015–08–16; 网络出版日期: 2015–09–29

很多音译方法将一个音译人名对看做 SMT 中的一个句子对[6]，将每个音译单元看做句子中的单词，并直接使用 GIZA++进行对齐，取得较好的翻译效果。

理论上，基于语音的音译框架能够更好地提高准确率。Gao 等[7]在 2004 年提出一种不同于噪声信道模型的基于音素的音译模型，直接使用源语言到目标语言的生成概率计算音译结果。但是，由于一个音译单元可能存在多种发音形式，并且由于不同语系之间拼写规则的不同，从源语言的语音转化成目标语言语音的步骤之间存在很大误差。基于字形的音译框架能够避免从字形转换到语音，从语音再还原成字形的音译单元的误差，摆脱对发音规则的依赖。李海舟研究小组[6,8–9]在英到中的音译中使用直接对齐，采用基于噪声信道模型进行音译，取得较好的效果，但是由于跳过了语音环节，会不可避免地产生信息丢失。

综合考虑以上方法的优缺点，本文在构建基于统计机器音译框架后，引入音译方法中的规则，在使用基于字形的音译框架的同时，融合语音要素的音译方法，提出音译单元的融合对齐方法。

1 流程描述

按本文方法构建的音译系统的流程如图 1 所示，主要包括数据前处理、训练音译模型、解码实验及后处理 4 个部分。

首先，在前处理阶段，数据来源分为训练语料与测试语料。将双语平行训练语料分别按照基于字母的音节划分规则和基于字形与字音并结合汉语与英语音节细划分规则，进行音译单元的粗划分与细划分。将测试语料也依据给出的音节划分规则进行相应的音译单元的划分操作。第 2 步，将已划分好音译单元的训练语料用提出的方法进行双语音译单元对的对齐。第 3 步，用已对齐的平行语料训练音译模型。第 4 步，对已划分好音译单元的源语言测试语料进行解码实验。第 5 步，将解码实验之后输出的目标语言音译结果进行还原操作，主要是进行音译单元的还原与格式还原。同时，倘若出现数据稀疏问题所造成的未登录词，则引入维基百科的数据，用于解决未登录词的翻译问题，有效地缓解数据稀疏问题。

本文主要论述音译系统中前处理、训练模型与解码实验的部分，后处理部分只做简单叙述。

2 数据前处理

前处理部分的重点在于对源语言语料与目标语言语料进行音译单元的划分。我们采取基于音节的音译单元划分规则，将音译单元的划分过程分为粗划分和细划分两个阶段。

2.1 音译单元粗划分阶段

英文名的音节划分规则是按照文献[5]给出的规则方法，首先将英文 26 个字母进行分类，分类情况如表 1 所示。完成对英文字母的分类后，按照表 2 所示的音节划分规则进行粗划分。

2.2 音译单元细划分阶段

根据以上粗划分的结果，我们发现划分后的语料中存在一些不合理现象，如音译对“埃 利 欧/E LIOU”、“罗 密 欧/ROM MEO”、“阿 布 拉 霍 尔/A B RA HA L L”等，通过日常的发音习惯可以清楚地分辨出，此处出现的“欧”或单独的“L”和“R”

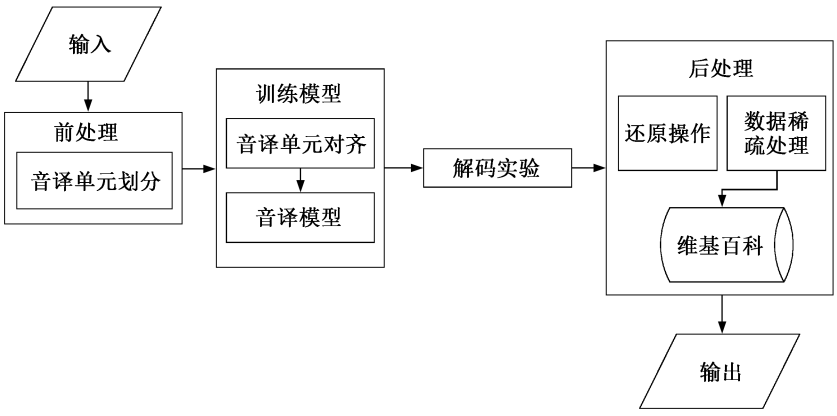

图 1 音译系统流程

Fig. 1 Frame diagram of the transliteration system

表 1　英文字母分类情况

Table 1　English letter classification

字母	详情	特殊情况
元音	{a, e, i, o, u}	字母 y 后面紧跟辅音字母时，视其为元音
辅音	鼻音: {m, n}	当“鼻音”被元音包围时，双写该鼻音字母
		当“鼻音”字母后面紧跟一个元音字母时，这两个字母重新组合，形成一个新的“元音”替代原来的两个音
	其他	字母 y 后面紧跟元音字母时，视其为辅音

表 2　音译单元粗划分规则

Table 2　Coarse segmentation rules for the transliteration unit

中文	英文
按字与字之间进行划分，将一个汉字视为一个音译单元	将连续的“辅音”字母划分开
	将连续的“元音”字母视为一个新的“元音”
	一个“辅音”和其后面紧跟的一个“元音”形成一个音节
	剩下的单独的“元音”或“辅音”则可以被视为一个单独的音节

表 3　音译单元细划分规则

Table 3　Further segmentation rules of transliteration unit

中文	英文
中文音译中常出现的辅助发音的汉字，如“尔”, “儿”, “欧”等，当出现该类词语时，将其与前一个音译单元合并为一个新的音译单元	当辅音 C, S, Z 后紧跟辅音 H 时，在细划分阶段，将以辅音 H 开头的后一音节与前一音节合并为一个音节
	若辅音 R 前后均为辅音时，实际英文发音过程中，这个辅音 R 总是辅助前一个音节发音，在细划分阶段，将此种情况的 R 与前一个音节合并为一个音节

等均是用于辅助前一音节发音的作用，此时将它们与前一音节合并为一个音节更符合发音规律。经统计，此种情况不在少数。

因此，我们依照数据统计结果改良发音规则，对粗划分的划分结果进行细化，如表 3 所示。例如，对于给定英文名 CHURTON，它的音译单元划分过程如图 2 所示。

在以往的研究中，对于音节的划分方法常常局限在一个步骤上，缺少相应的细化过程，会对后面步骤的效果产生影响。本文采用两个阶段的划分过程，经实验 2 和 3 (见 5.2 节)论证，能够更好地提升音译效果。

英文名:	CHURTON
经粗划分规则划分:	C/HU/R/TON
经细化分规则划分:	CHUR/TON
对应中文名:	丘/顿

图 2　“CHURTON”的划分过程

Fig. 2　“CHURTON” segmentation process

3　音译模型

3.1　规则与统计相结合的自动对齐方法

音译单元等级自动对齐的主要目的在于使汉英双语名字各自的音译单元相互对齐。例如上述例子“丘顿/CHUNTON”，自动对齐的结果就是“丘/CHUN”和“顿/TON”。在机器音译中，双语音译单元的对齐效果直接影响音译结果的好坏，同时由于在音译过程中不存在音译单元的调序问题，通常情况下，源语言音译单元的对齐结果就是目标语言相同序号的音译单元。

由于在上一步分词过程中常存在源语言与目标语言划分的音译单元个数不同的情况，一般的自动对齐常存在一对多与一对空的问题，这样的对齐结果往往不具有代表性，对提升音译效果起阻碍作用。因此，自动对齐的难点在于选择正确的音译单元对，尽量消除上述问题。我们采用基于规则的自动对齐算法，具体步骤如下。

1) 对于分词后汉语与英语名字音译单元个数相同的情况，采取直接对齐的规则，即将相同序号的

音译单元对齐, 形成音译单元对, 例如: “欧 文/ERWIN”。

2) 对于分词后汉语与英语名字音译单元个数不相同的情况。

① 首先将汉语名字分词结果转化成拼音的表示形式, 例如, “埃格德/AAGAARD”表示为“AI4 (1) GE2(2) DE2(3)/AA(1) GAA(2) R(3) D(4)”。

② 根据音节首字母匹配规则, 以汉语的音译单元首字母为准, 分别对应英语的音节首字母, 即用 A, G, D 这 3 个字母, 将英文名字“AAGAARD”重新划分成“AA”、“GAAR”和“D”三部分。同时根据汉英字母发音的规律, 按照文献[4]中的权重分配规则, 将划分方式进一步细化。

③ 经过上述步骤, 将得到一个英语名字的一种或几种的划分方式$<c, e>_i$, (i= 1,2, …, n)。

④ 将它们都列入候选集合$\mathfrak{R}$中, 同时经过步骤④和步骤⑤, 得到概率最高的划分方式。

⑤ 计算第 i 种划分方式中, 单个音译单元对$<c_k, e_k>$的概率:

$$P(<c_k,e_k>)_i=\frac{|<c_k,e_k>|}{|c_k|}, \tag{1}$$

其中, $|<c_k, e_k>|$与$|<c_k>|$表示该音译单元对在所有对齐方式中的统计与在所有名字中对应音译单元的统计。

⑥ 计算第 i 种划分方式的概率:

$$P(<c,e>)_i=\prod_{k=1}^{n}P(<c_k,e_k>)_i, \tag{2}$$

比较 n 种划分方式的概率大小, 取概率值最大的划分方式作为最终划分方式。

3.2 N-gram 音译模型

对于汉英方向机器音译, 假设中文名与英文名可以以字符序列的方式表示, 其中, 中文名表示为 $\alpha=x_1x_2x_3\ldots x_m$ (m 表示中文名汉字数), 英文名表示为 $\beta=y_1y_2y_3\ldots y_n$ (n 表示英文名字母数), 经过前处理与对齐的步骤后, 中、英人名对被分别表示为音译单元的序列。

中文名字: $\alpha=c_1c_2c_3\ldots c_k$; 英文名字: $\beta=e_1e_2e_3\ldots e_k$。$c_i$ 和 e_j (i = 1, 2, 3, … k, j= 1, 2, 3, … k)分别表示第 i 或 j 个中文或英文音译单元, 即中英文音译单元的数目相同。

由此, 中文音译单元 c_i 与英文音译单元 e_i 就形成对齐关系。α 与 β 的对齐关系 γ 表示如下:

$$\gamma=<\alpha,\beta>=\{<c_1,e_1>,<c_2,e_2>,<c_3,e_3>,\ldots<c_k,e_k>\},$$

其中, 一个中文音译单元中可能包含一个至多个汉字, 一个英文音译单元中可能包含一个至多个英文字母。

根据上述 α, β, γ 的定义, 汉语到英语的音译过程可以用下式推导:

$$\begin{aligned}\bar{\alpha}&=\arg\max_{\alpha}P(\alpha,\beta)\\&=\arg\max_{\alpha}\sum_{\gamma}P(\alpha,\beta,\gamma)\\&\approx\arg\max_{\alpha}(\arg\max_{\alpha}P(\alpha,\beta,\gamma))\\&=\arg\max_{\alpha,\gamma}P(\alpha,\beta,\gamma),\end{aligned} \tag{3}$$

其中, $P(\alpha, \beta, \gamma)$表示 α, β, γ 的联合概率。

经过实验对比, 我们采取 N-gram 的音译模型, 其中 n = 3, 式(3)重写为

$$P(\alpha,\beta,\gamma)\approx\prod_{i=1}^{K}P(<c_i\,|\,e_i>|<c_{i-2}\,|\,e_{i-2}>,<c_{i-1}\,|\,e_{i-1}>)。 \tag{4}$$

4 数据后处理

4.1 还原操作

经过解码实验, 输出的最优结果是以音译单元形式表示的目标语言人名(本文研究的音译方向为汉到英, 因此输出的目标语言为英语)的形式, 这并不是我们真正需要的音译结果, 因此, 需要对该数据进行还原处理, 我们主要进行了两个步骤的还原操作。

1) 音译单元还原操作。在音译单元的划分阶段, 特别是在细划分阶段, 存在将鼻音{m, n}双写的情况, 所以在解码实验输出结果的音译单元中也存在这种情况。因此, 当出现“mm”或“nn”时, 若其前后是被元音包围的情况, 将其改为“m”或“n”。

2) 格式还原操作。在实际音译单元划分过程中, 音译单元与音译单元之间是以空格区分的。因此, 此处的格式还原操作为去除音译单元之间的分隔符, 将其还原为一个单词的形式。

4.2 数据稀疏处理

在音译过程中不可避免地会产生数据稀疏问题, 本研究使用维基百科的数据来缓解这一问题。主要方法是, 将出现数据稀疏问题的源语言人名再次进行前处理操作, 同时从维基百科中抽取汉英人名对作为参考语料, 对其进行与之前的训练语料相同的处理操作后, 利用式(1)和(2), 选取与问题人名

中音译单元对应的概率最大的目标语言音译模型，并将其作为新的解码实验的输出结果，再进行还原操作。

5　实验分析

实验使用的双语语料来自 I2R 2009 的音译数据[6,8–9]。该数据包含 31961 条惟一的英文词条及其对应的官方音译结果，各部分数据的使用量如表 4 所示。

表 4　实验数据

Table 4　The dataset

训练集	测试集	维基百科
31961	2896	37151

5.1　实验评价

对于本次实验结果的评价方法，采用的是 PRF 系统评测模型，其中 P (Precision)为准确率，R (Recall)为召回率，F 值用于均衡准确率与召回率的误差。本次实验中对准确的定义是音译结果与参考集中给定的参考结果完全一致。

$$P=\frac{\text{正确的音译数}}{\text{总音译结果数}}, \tag{5}$$

$$R=\frac{\text{正确的音译数}}{\text{待音译总数}}, \tag{6}$$

$$F=\frac{2PR}{P+R}。 \tag{7}$$

5.2　实验结果

为从整体上比较本文方法与只使用基于字形的音译方法，我们设计了以下 4 个实验。

1) 基线实验。本文基线系统采用文献[10]提出的方法，以评价提出方法的性能。仅采用基于字形的音译单元对齐方法，对英文语料进行简单的按音节的音译单元划分方法，对中文语料采取按空格音译划分方法，并用 GIZA++工具进行音译单元的对齐，训练音译模型并输出最好的一个结果，将其实验结果作为对比参照。

2) 粗划分实验。将训练语料只进行音译单元的粗划分，并用 GIZA++工具进行简单的汉英音译单元的对齐，训练我们的音译模型，并输出最好的一个结果。

3) 双重划分实验。将训练语料进行音译单元的粗划分与细划分，并使用 GIZA++工具进行简单的汉英音译单元对齐，训练我们的音译模型，并输出最好的一个结果。

4) 对齐改进实验。将训练语料进行音译单元的粗划分与细划分，并使用我们提出的对齐改进方法处理对齐结果，用该数据训练我们的音译模型，并将 Top1 作为输出结果。

与基线系统相比，我们的系统得到较好的性能表现(表 5)，分析如下。

1) 单纯的基于字形的音译方法，音译效果不理想，例如“斯滕尼/STENY”，用该方法的输出结果是“STENNY”，而在其他两个实验中均能获得正确结果。这种鼻音的单写双写问题在现实应用中并不少见，因此该方法不能直接用于机器音译中。

2) 引入新的划分步骤之后，音译单元的划分更加准确，例如“斯托克迈/STOCKMAYER”，在粗划分时会被划分成“斯 托 克 迈/S TO C K MA YE R”，英文音译单元明显划分不够准确，在经过细划分后，成功地变为以“S TO CK MA YER”表示的更准确的形式。音译系统的准确率、召回率与 F 值均有提高，足以证明该方法的可行性。

3) 运用我们提出的对齐方法后，P, R 和 F 值都有明显提升，进一步验证了字形与语音融合的音译单元对齐方法既降低了语音转换步骤中的误差，又减轻了仅采用基于字形的方法造成的信息对视问题。由此可以得出，基于字形和语音的音译单元对齐方法能够提高音译的效果。

6　总结及未来工作

本文提出一种新的融合的方法用于音译单元的划分与对齐过程。经过实验验证得知，我们提出的方法能够很好地提高音译的准确率，同时在解决音译单元对齐的一对多与一对空问题方面表现较好。本研究有如下创新。

1) 提出融合字形与语音的音译单元对齐方法。在以往的研究成果中，大部分的工作将关注点投放

表 5　实验结果

Table 5　Results of experiments

测试集	P	R	F
基线实验结果	0.163152	0.162983	0.163067
粗划分结果	0.172544	0.172306	0.172426
双重划分结果	0.194000	0.193674	0.193836
对齐改进结果	0.222545	0.222307	0.222425

在字形或者语音音素一个纵向的方面。在本次研究中，我们致力于将字形与语音的研究成果结合起来，吸收两者的优点，弥补其中一方的缺点，更好地提升音译效果。

2)结合规则与统计音译方法各自的优点，提出规则与统计相结合的音译单元划分与自动对齐的方法，将其运用在相应过程中，并通过实验验证了该方法的可行性。

但是，对于来源不同的英、汉人名，存在不同的音译习惯，在我们的音译过程中并没有很好地解决这个问题。下一步的工作将引入更多的音译单元划分规则与对齐规则，同时更好地利用维基百科的数据，对来源不同的人名进行不同处理，希望能够进一步提高音译的效果。

参考文献

[1] 李婷婷. 基于非参数贝叶斯学习的多语言人名音译研究[D]. 哈尔滨: 哈尔滨工业大学, 2013

[2] Lin Weihao, Chen Hsin-Hsi. Backward machine transliteration by learning phonetic similarity // Proceedings of the 6th Conference on Natural Language Learning. Taipei, 2002: 1–7

[3] Zaidan O. Z-MERT: a fully configurable open source tool for minimum error rate training of machine translation systems. Prague Bulletin of Mathematical Linguistics, 2009, 91: 79–88

[4] Wan S, Verspoor C M. Automatic English-Chinese name transliteration for development of multilingual resources // Processing of the 17th ICCL. 1998: 1352–1356

[5] 蒋龙，周明，简立峰. 利用音译和网络挖掘翻译命名实体. 中文信息学报, 2007, 21(1): 23–29

[6] Li Haizhou, Kumaran A, Zhang Min, et al. Whitepaper of NEWS 2009 machine transliteration shared task // Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration. Singapore: Association for Computational Linguistics, 2009: 19–26

[7] Gao Wei, Wong Kam-Fai, Lam Wai. Phoneme-based transliteration of foreign names for OOV problem // Proceedings of the 1st International Joint Conference on Natural Language Proceedings, Lecture Notes in Computer Science. Hainan, 2004: 110–119

[8] Li H, Zhang M, Su J. A joint source-channel model for machine transliteration // Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Barcelona, 2004: 1190–1194

[9] Zhang Min, Li Haizhou, Su Jian. Direct orthographical mapping for machine transliteration // Proceedings of the 20th International Conference on Computational Linguistics (COLING'04). Sydney, 2004: 716–722

[10] Wang Dandan, Yang Xiaohui, Xu Jin'an, et al. A hybrid transliteration model for Chinese/English named entities — BJTU-NLP Report for the 5th Named Entities Workshop. Beijin, 2015