基于双向融合的日语外来语翻译技术的研究*

高永磊　张桂平　季铎　国宪宇　蔡东风
沈阳航空工业学院知识工程中心 辽宁 沈阳　110034
E-mail: yonglgao@163.com

摘　要：日英翻译需要处理日语片假名的英译还原问题。本文利用了基于短语的统计翻译方法对片假名进行日到英、英到日的双向翻译，在对翻译结果进行分析的基础上总结了日英互译过程中的各自特点，提出了一种基于双向融合的片假名翻译方法，该方法优化了日英的翻译结果，实验结果表明，基于双向融合的翻译策略翻译准确率达到78.5%，比基于短语的日英片假名翻译提升了4.5个百分点，有效地解决了日语片假名的英译还原问题。

关键词：片假名，双向融合，基于短语的统计翻译

Research on Katakana Translation Based on Bi-Directional Integration

Yonglei Gao, Guiping Zhang, Duo Ji, Xianyu Guo, Dongfeng Cai
Knowledge Engineering Research Center, ShenYang Institute of Aeronautical Engineering, ShenYang, 110034
E-mail: yonglgao@163.com

Abstract: Japanese-English translation must deal with the problem of Katakana reduced to English. This paper exploits the phrase-based statistical machine translation model to achieve Katakana translation from Japanese to English and back-translation from English to Japanese. After analyzed the results and summarized the characteristics of mutual translations, this paper describes a translation method based on bi-directional integration; our method optimized the Japanese -to-English translation results. The experimental results indicate that the translation precision reaches 78.5%, our method outperforms phrase-based Japanese-English Katakana translation model 4.5%, effectively addresses the problem of the Katakana reduced to English.

Keywords: Katakana, Bi-directional Integration, Phrase-based Statistical Machine Translation.

1　引言

片假名是日语中音节文字的一种，主要用来表示外来语及拟声语。日英的翻译中需要处理由片假名表示的外来语（本文指从英语中吸收来的科技术语），虽然外来语可以通过双语词典来解决部分翻译问题，但由于新词的不断涌现，双语词典不可能收录所有的外来语及其翻译[1][2]，因此如何对双语词典中未收录片假名进行翻译是日英翻译需要解决的问题[3]。

片假名是对外来语进行读音的近似转换的表示，因此将片假名还原为英语最常用的方法是基于语音的翻译[4](transliteration，简称音译)。音译就是用转录的方法将源语言转录成目标语言，转录通常是根据发音的近似实现的。现有的音译的方法有：基于字形的方法、基于音素的方法以及混合的音译方法[5][6][7]。基于字形的方法[8]是通过罗马字符与英文字母的编辑距离运算进行拼写校正的翻译模式，基于音素的方法[4]则利用了源语言的音素生成目标语言的字形。基于音素的方法处理过程比基于字形的方法复杂，但是基于字形的方法仍需要用罗马字符表示片假名，然后利用拼写校正的方法翻译。

为进一步简化音译的中间过程，本文提出的方法省略了片假名用罗马字符表示，然后与英

*[基金项目]国家自然科学基金项目，项目号：6084200。
[作者简介]高永磊(1983-)，男，研究生，主要研究方向为自然语言处理，机器翻译。

文单词进行对齐这一过程，将片假名与英文字母分别进行全切分，进行对齐训练、抽取短语表、训练英文字母的语言模型，然后利用基于短语的统计翻译模型[9][10]对片假名进行翻译。为进一步提高翻译结果正确率，本文根据片假名的日英、英日翻译的特点，提出了基于双向融合的方法，实验结果表明，双向融合的方法能够有效的提高翻译质量。

本文的其它部分组织如下：第二部分介绍基于字形的方法和基于短语的统计翻译模型，第三部分介绍本文的双向融合策略，第四部分介绍实验及实验结果分析，最后给出本文工作的总结和下一步工作展望。

2 相关研究

2.1 基于字形的方法

Brill & Moore[2]利用拼写校正模型[8]实现了片假名到英文的翻译。此模型学习 $\alpha \to \beta$（α 为片假名的罗马字符的组合，β 为对应英文字母的组合，也可以为空）之间的映射参数和概率，然后利用公式 1(其中 Katakana 是罗马字符表示的片假名)：

$$\arg\max_{English} P(English \mid Katakana) \tag{1}$$

找出此片假名的最佳翻译结果。$\alpha \to \beta$ 之间编辑概率的获取方法如下：首先将片假名和英语双语对<source，target>进行单个字母对齐，然后利用改进的编辑距离算法，对片假名和英语对进行训练，得到任意子串的概率，再通过拼写校正模型得出输入的片假名最可能的英文原文结果。例如双语对<コンテキスト,context>，首先对片假名进行罗马字符转化，并对罗马字符和英文字母进行对齐，得到罗马字符与英文字母的对齐表，其结果如图 2.1 所示，假定窗口大小为 4，然后对罗马字符串的所有片段集进行编辑距离概率计算（例如 $c \to k, co \to ko, con \to kon$ 等），最后利用公式 1 进行计算，得出最终结果。

本文的方法不同于基于字形的方法，区别在于本文将片假名和英语单词进行全切分，利用 GIZA++[11]对齐，抽取翻译模型，而省略了将片假名用罗马字符表示，利用罗马字符与英文字母对齐的过程。本文通过比对片假名的罗马发音，发现对齐效果基本符合发音对齐效果，例如图 2.2，サ的罗马发音为 sa，ム的发音为 mu，ソ的发音为 so，ン的发音为 n，与对齐结果基本相同，因此利用字形对齐抽取翻译模型进行片假名翻译是可行的。

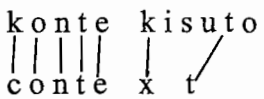

图 2.1 罗马字符与英文单词对齐结果

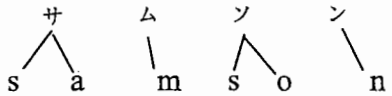

图 2.2 片假名字与英文单词对齐结果

2.2 基于短语的翻译模型

基于短语的方法首先将输入的源语言句子 f 分割成由 I 个短语组成的短语序列 f_1^I，然后将每个源语言短语 f_i 翻译成为目标语言短语 e_i，从而生成目标语言句子[9]。

翻译模型采用的是对数线性模型[12]，其翻译过程可以描述为：

$$p(e \mid f) \approx P_{\lambda_1 \ldots \lambda_M}(e \mid f) = \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(f,e)]}{\sum_{f'} \exp[\sum_{m=1}^{M} \lambda_m h_m(f',e)]} \tag{2}$$

其中 f、e 是机器翻译的源语言和目标语言句子，$h_1(f,e),\ldots,h_M(f,e)$ 分别是 f、e 上的 M 个特征，$\lambda_1,\ldots,\lambda_M$ 是与这些特征分别对应的 M 个参数（或权值）。

其整体翻译概率是各个特征加权后的乘积。对于给定的S，相应的最优译文T为：

$$E_{best} = \arg\max_{e}\{p(e \mid f)\} = \arg\max_{e}\{\sum_{m=1}^{M} \lambda_m h_m(f,e)\} \tag{3}$$

特征函数选取语言模型和翻译模型作为基本特征，则：

$$E_{best} = \arg\max_{e}\{\lambda_T p_T(f \mid e) + \lambda_{LM} p_{LM}(e)\} \tag{4}$$

其中 $p_T(f \mid e)$ 代表短语的翻译概率，每个短语的翻译概率为 $p(f_i \mid e_i)$，则 $p_T(f \mid e)$ 的计算公式为：

$$p_T(f_1^I \mid e_1^I) = \prod_{i=1}^{I} p(f_i \mid e_i) \tag{5}$$

$p_{LM}(e)$ 代表目标语言的语言模型概率。

3 基于双向融合的片假名翻译

本文利用基于短语的统计机器翻译方法对片假名进行翻译，并对实验结果进行了分析，得知正确结果在前几个候选单词内出现的频率较高，例如“キット”的候选有“kitt，kit，kite，cotte，ket”，与其正确结果“kit”比较，可以看出字母多译现象严重，如“キ”可翻译为“ki，co，ke”等，针对多译现象，本文将“kitt，kit，kite，cotte，ket”译回片假名，可以从这些英文单词译出的片假名候选找到原片假名“キット”，因此本文提出将日英、英日翻译结果进行整合，并通过打分排序筛选出能够译回原片假名的英文结果作为最佳翻译结果。

3.1 融合方法

本文首先利用日英翻译系统（记为：JE）对片假名进行翻译，并取翻译的前N-best结果，利用英日翻译系统（记为：EJ）反译成片假名，然后将两次翻译的结果进行融合，取出能够反译成原片假名的英语单词作为最佳翻译结果。翻译实例如图3.1。

本文利用日英翻译系统得到片假名“オロゲン”的N-best翻译结果，去重之后得到M（此例为5）个候选单词，然后利用英日翻译系统对此M个单词进行翻译，得到多个候选片假名，再通过译出的片假名与原片假名“オロゲン”校对，得到反译回片假名“オロゲン”的K（此例为3，黑体表示）个英文单词，最后通过如下方式输出1-best结果：对此K个英文单词利用公式6进行融合打分，并将结果按照从大到小的顺序进行排序：

$$score_{match} = score_{JE} + \beta \times score_{EJ} \tag{6}$$

其中 $score_{JE}$ 为日英翻译结果的打分值（同理 $score_{EJ}$ 为英日翻译结果的打分值），β 为调节系数，最终得到1-best结果并输出。

3.2 融合算法

本文提出的利用日英翻译系统和英日翻译系统的融合算法如下：

1. 输入片假名，利用日英翻译系统翻译出N-best结果，并提取各自的打分值；
2. 对于step1的N-best结果去重过滤之后取出M-best结果及其对应的打分值；
3. 对于step2得到M-best的每一英文单词，利用英日翻译系统译出其K-best结果，并提取打分值；
4. 将step3的英文单词与其译出的K-best结果利用公式6进行融合，其中对于能够译回原输入片假名的结果，其 $score_{EJ}$ 值采用其对应的打分值，不能译回原输入片假名的结果，本文赋给其一个极小值；

5. 重复 step3、step4 直至 M-best 的所有单词均融合完毕；
6. 将最终的融合结果进行排序，输出 1-best 结果。

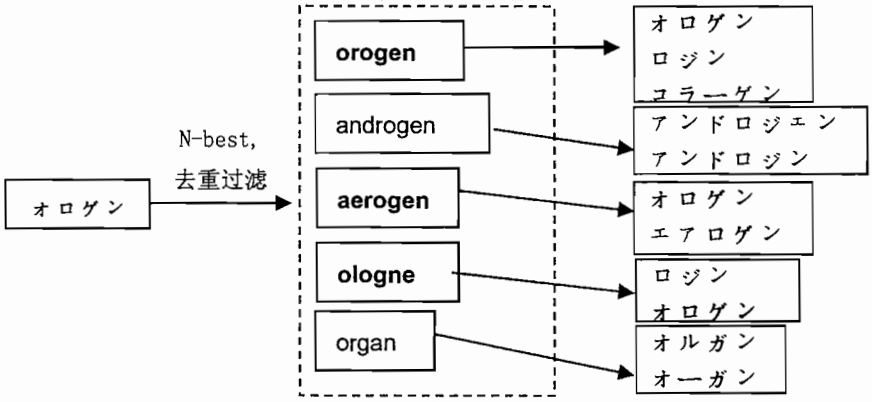

图 3.1 片假名双向翻译

4 实验及结果分析

本文采用 13820 个日英词对作为语料集，其中训练语料取 11981 个词对，测试语料取 1839 对，英语和片假名语言模型分别采用了更大的英文单词和片假名语料集进行训练，分别得到了从 6 元到 12 元不同 gram 的语言模型。对于翻译模型的训练，本文利用 GIZA++从日英、英日两个方向进行训练获得字母对齐，利用 GROW-DIAG-FINAL 方法进行优化对齐，然后抽取短语得到日英和英日两个短语翻译概率表。本实验以基于短语的日英统计翻译系统为基线系统，将双向融合的系统与基线系统进行比较。实验结果表明，双向融合的方法要优于基于短语的方法。本文还测试分析了语言模型元数对翻译效果的影响。

本文的实验以系统给出的最优结果的正确个数来评价系统的性能（公式 7），主要考察了系统输出的前 15 个结果。输入一个片假名，得到系统输出的前 15 个结果后，观察这 15 个结果是否含有正确答案（翻译出的结果单词与标准答案完全匹配则为正确答案），以及正确答案在这 15 个结果中首先出现的位置。本文统计了结果集在 1、2-5、6-10、11-15 这四个区间段内的翻译准确率。

$$准确率(Precision)=\frac{正确翻译的个数}{总的翻译个数}\times 100\% \qquad (7)$$

4.1 日英翻译结果及分析

由于片假名翻译不需要调序，本实验没有使用调序模型。表 4.1 是片假名在不同 gram 英语语言模型上的开放性测试结果。通过表 4.1 可以看出基于短语的翻译系统对于片假名的翻译效果基本令人满意，第一候选的正确率比较高，在 2-15 个候选中，2-5 的候选占的比重比较大，这说明通过一定的选择策略，可以进一步提高第一候选的正确率，同时可以看出翻译性能在 11gram 以后并没有提升，因此语言模型在 11 元以后已经不能提高翻译质量。

通过对错误翻译结果的分析，可以看出在翻译过程中多译漏译现象以及部分译错现象比较多，多译现象例如“トリプル”的正确结果为“triple”， 而统计翻译的结果却为“tripple”；“ウエーブ”的正确结果为“wave”， 而统计翻译的结果却为“weave” ；“エーテル”的正确结果为“ether”， 而统计翻译的结果却为“eather”。漏译现象如“コサイン”的正确结果为“cosine”，而统计翻译的结果却为“cosin”；“ジョーク”的正确结果为“joke”，而统计翻译的结果却为“jok”。译错现象例如“ミオジン”的正确结果为“myosin”， 而统计翻译的结果却为“myogen”；“ケ

ール”的正确结果为“kale”， 而统计翻译的结果却为“cale”。

表 4.1 基于短语的日英翻译在不同 gram 语言模型的结果

	1	2-5	6-10	11-15	1-15
6gram	65.80%	17.24%	2.88%	1.03%	86.95%
7 gram	72.05%	12.94%	1.79%	1.14%	87.92%
8 gram	73.25%	12.18%	1.63%	0.98%	88.04%
9 gram	73.46%	11.96%	1.69%	0.92%	88.03%
10 gram	73.84%	11.96%	1.74%	0.82%	88.36%
11 gram	73.95%	11.80%	1.79%	0.82%	88.36%
12 gram	73.95%	11.80%	1.79%	0.82%	88.36%

4.2 日英与英日融合翻译结果及分析

由 4.1 日英实验结果分析可知在 2-5 候选中正确的结果占的比重比较大，通过一定的选词策略可以提高 1best 的正确率。本文只针对单一片假名进行翻译，缺乏可以用来指导选词的信息，为了提高 1best 的正确率，本文对日英和英日系统的翻译结果进行融合实验。表 4.2 是在不同 gram 的片假名语言模型上，利用 1839 对测试语料对英日翻译系统测试的实验结果。

表 4.2 英日翻译在不同 gram 语言模型的结果

	1	2-5	6-10	11-15	1-15
6gram	51.28%	24.36%	4.84%	2.34%	82.82%
7 gram	51.44%	23.93%	4.89%	2.56%	82.82%
8 gram	51.39%	24.09%	4.79%	2.56%	82.83%
9 gram	51.44%	24.03%	4.79%	2.56%	82.82%
10 gram	51.06%	24.52%	4.73%	2.50%	82.81%
11 gram	51.01%	24.58%	4.73%	2.50%	82.82%
12 gram	50.95%	24.63%	4.73%	2.50%	82.81%

通过表 4.2 可以看出在同一训练和测试语料的基础上，英日翻译的效果明显低于日英翻译的效果，第一候选的正确率仅仅过半，而 2-5 的候选占的比重非常大。在 6-12 元语言模型中，采用 9 元语言模型翻译的第一候选的正确率最高，6-9 元呈上升趋势，9-12 元呈下降趋势，因此语言模型取 9 元最优。

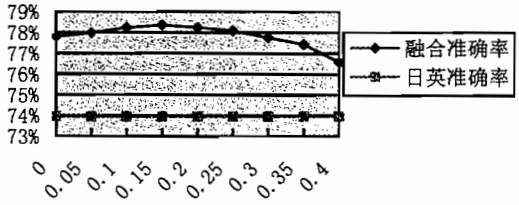

图 4.1 不同 β 值的实验结果比较

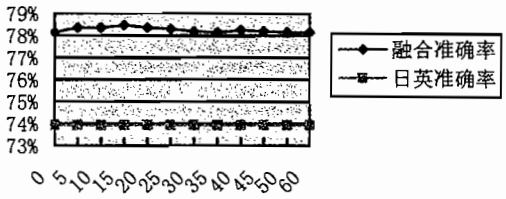

图 4.2 英日翻译系统不同候选个数的实验比较结果

由表 4.1 实验结果可以看出，对于前 5-best 候选，正确率可以达到 85%，因此本文融合实验取日英翻译系统翻译结果的个数为 5。本实验先假定英日翻译系统取得候选值为 10，然后调节 β 的值，使融合系统的翻译性能达到最佳，并以此确定 β 的值。由图 4.1 的实验数据可以看出随着 β 值由 0 向 0.15 递增，融合系统的翻译准确率也是逐步上升，而当 β 值从 0.15 向 0.4 递增时，融合系统的翻译准确率呈下降趋势，而且后期下降明显，因此本文由图 4.1 实验结果，确定调节系数 β 的值为 0.15。在 $\beta = 0.15$ 的情况下，本文调整英日翻译系统不同的 N-best 候选翻译结果，

以求双向融合的翻译结果准确率达到最高，得到图 4.2 的实验结果。通过对图 4.2 的分析可以看出，在英日翻译系统取前 15-best 的情况下，准确率最高并达到 78.47%，比基线系统高出 4.5 个百分点，这说明日英与英日的双向融合策略能够提高 1-best 的正确率。

本文通过对双向融合翻译结果的分析，可以看出融合之后的结果能够有效改变候选排序，提高第一候选正确率。例如“カストロ”的前 5 个单词候选为“castor，caster，castle，astro，castro”，而正确的单词“castro”排名很低，通过英日翻译之后的结果融合，仅能得到“castro”翻译成“カストロ”，使得“castro”成为第一候选。

5 总结

本文提出了双向融合的策略进行片假名翻译，该方法的优点在于通过对融合结果的排序有效改变仅基于短语的翻译候选结果的排序。实验表明，基于双向融合的翻译方法比基于短语的日英翻译方法的结果提升了 4.5 个百分点，达到了较为理想的效果。

本文使用的语料规模比较小，统计翻译的结果仅达到 78%，本文下一步将扩充语料规模来验证本文双向融合翻译的性能。另外，本文实验仅考虑了将片假名还原成英文的情况，而片假名分为多种，对于日语而言，只要是外来语，一律用片假名标识，因此片假名可以翻译为很多国家的语言，例如“エネルギー”被翻译为德语“Energie”，“アフロディーテ”被翻译为希腊语“Aphrodite”等，如何识别片假名引入语言的归属，并进行相应语言的还原是下一步需要解决的问题。

参 考 文 献

[1] Jeong, K. S., Myaeng, S. H., Lee, J. S., & Choi, K. S..automatic identification and back-transliteration of foreign words for information retrieval[C]. Information Processing and Management, 1999.523-540.

[2] E. Brill, G. Kacmarcik and C. Brockett. Automatically harvesting katakana-English term pairs from search engine query logs[C]. In Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, 2001.393-399.

[3] Goto, N. Kato, N. Uratani, and T. Ehara. Transliteration considering context information based on the maximum entropy method[C]. In Proc. of IXth MT Summit, 2003.125-132.

[4] Kevin Knight, Jonathan Graehl. Machine Transliteration[C]. In Proceedings of ACL, 1998.599-612.

[5] S.Bilac and H.Tanaka.Improving Back-Transliteration by Combining Information Sources[C]. In Proc. Of the First International Joint Conference on Natural Language Processing, 2004.542-547.

[6] Jong-Hoon Oh, Key-Sun Choi and Hitoshi Isahara. A Comparison of Different Machine Transliteration Models[C]. In Journal of Artificial Intelligence Research, 2006.119-151.

[7] Andrew Finch, Eiichiro Sumita. Phrase-based Machine Transliteration[C]. In Proc. of IJCNLP 2008, Workshop on Technologies and Corpora for Asia-Pacific Speech Translation,2008.13-18.

[8] E. Brill and R. C. Moore. An improved error model for noisy channel spelling correction[C]. In Proceedings of the 38th Annual Meeting of the Association for Computarional Linguistics, 2000.286–293.

[9] Philipp Koehn, Franz J.Och, and Daniel Marcu. Statistical Phrase-Based Translation[C]. In Proceedings of HLT-NAACL, 2003.127-133.

[10] R. Zens, F.J. Och, H. Ney. Phrase-Based Statistical Machine Translation[C]. In KI-2002: Advances in artificial intelligence.25.Annual German Conference on AI, KI2002, Vol.LNAI2479:18-32.

[11] F. J. Och and H. Ney. Improved statistical alignment models[C]. In Proceedings of the 38th Annual Meeting of the Association for Computarional Linguistics, 2000.440-447.

[12] F. J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation[C]. In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, 2002.295–302.