拼読：Pinyomi
漢字対応に基づく日中辞書検索

靳 志輝　　Lars Yencken*　　田中久美子
東京大学情報理工学系研究科

概 要

拼読は、日本人が中国語を日本語で辞書引きするためのインターフェースである。ユーザが中国語の日本語読み、あるいは、対応する日本語漢字を入力すると、拼読は、日中の漢字対応 (grapheme transliteration) を利用して、確率的に母国語の候補を列挙する。本稿では、本システムの設計を示すと共に、簡単な評価実験について報告する。(http://www.ish.ci.i.u-tokyo.ac.jp/pinyomi/)

1　はじめに

日本人が中国語の単語の意味を調べたいときには、pinyin の読みに基づく方法が最も一般的である。その際、中国語の読みがわからない場合には、各漢字の pinyin を調べ、その上で辞書を引くことになり、かなり時間がかかる。

本稿では、中国語を日本語の読みあるいは日本語の漢字で引くことのできる拼読システムについて報告する。たとえば、中国語「香蕉」を調べたいとき、日本語の読み「か・しょう」を入力する。すると、システムはその読みに対応する中国語の候補を挙げ、その中には「香蕉」が含まれる。これを選択すると、その意味は「バナナ」、読みは “xiang jiao” であることがわかる。同様に、中国人が pinyin を入力することにより、日本語の漢字を引くこともできる。

この辞書引きは日中の漢字対応を基本とする。母語の漢字、あるいは母語の漢字に対応する読みを入力すると、目的言語での単語を得る仕組みである。

拼読のアイデアは、漢字という共通基盤にのっとって、日本人は中国語を日本語読みで読み、中国人は中国語読みで読んでいることに基づく。この現象は、より広くは language transfer [5][6] —誰しも外国語を学ぶときには母語に即して学ぶ— という普遍的な現象の一つとして捉えることができる。実際、西欧諸語の間でもフランス人が英語をフランス語読みで発音することなど、よくあることである。

読みの上での language transfer は自然言語では、transliteration system として研究されてきており、昨今では特に情報検索技術と関連して、枚挙にいとまがない。しかし、これまでの transliteration は、表音文字対応を基本とし、また、固有名詞や技術用語を対象としていることが多い。たとえば、アラビア語英語対応などは当然表音文字上の対応関係を論じるものであり [1]、中国語英語対応であっても、pinyin との transliteration が主な論となる [8]。英日間の研究はこれらの口火を切ってきたが [7][3]、そこで問題とされたのも、英語と日本語のカタカナ語表記の対応であった。総じて、外来語の情報検索が焦点となることが多いため、固有名詞や技術用語が対象とされてきた。一方で、拼読は、日中の共通言語としての漢字上の結びつきにのっとっているため、grapheme transliteration がテーマである。したがって、日中の大部分の単語が transliteration の対象となる点も大きな特徴である。

別の関連研究の観点からは、本システムは外国人のための日本語検索システム FOKS[2] の日中版であると捉えることもできる。FOKS は、必ずしも正しくない日本語の読みが入力されても国語辞典を引くことのできるシステムで、日本語を学ぶ外国人用に開発された。たとえば「発表」を「はつひょ」などで引くことができる。拼読は同様に中国語を「当たらずとも遠くはない」日本語で引くという意味で、FOKS 的な側面を持っているシステムである。

実は日本語の漢字の読みで中国語漢字を検索するシステムは、既存の電子辞書などに実装されている[1]。しかし、現状では日本語の読みで中国語の漢字一字が引

*正式所属はメルボルン大学情報科学科。2006 年 10 月〜2007 年 3 月末まで東大情報理工に協力研究員として滞在。

[1] http://www.sony.jp/products/Consumer/DD/

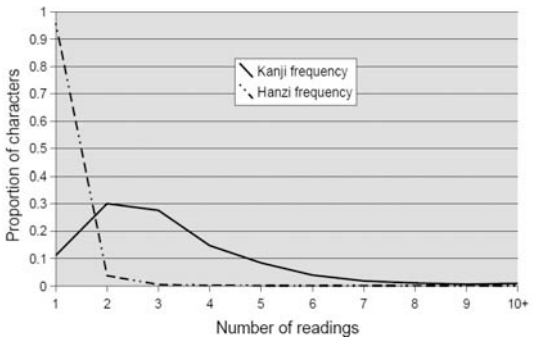

図 1: 日中の漢字の読みの数の分布

けるだけで、自明なものにとどまる。拼読は、この考え方を一般化し、単語を検索することができるように拡張したシステムとも位置付けることができる。

2　日中の漢字対応

一般に、二国語間の表記の対応には二つの異なる軸がある：意味と綴りである。拼読は漢字の綴りの対応を基本とするので、綴りの軸で漢字対応を考察する。

日中の漢字は、それぞれ別の歴史をたどり、簡略化されてきたので、同起源であっても綴りが異なるものも多い。日本語、あるいは中国語のある漢字に関して、相手言語での対応する漢字について考えると、３つの場合がありうる。

- まったく同じ綴りの漢字がある。たとえば、「主」。
- 同意味、同起源などの背景から、綴りは異なっても対応するものがある。これには、「紅」と「红」や「問」と「问」と推測可能なものから、より推測が難しい「関」「关」や「書」「书」などといったものもある。また、「発」「發」と「发」などのように対応が曖昧であるものもある。
- 対応する漢字がない。日本語については「峠」「込」など。中国語については「你」など。

すなわち、日本人、中国人どちらにおいても、自国語の漢字にない字、あるいは対応がわからない漢字があるということになる。拼読は、この現状を鑑みて設計する必要がある。

また、拼読に関する別の差異として、日本語での漢字の読みの方が、中国語の漢字の読みよりも、はるかに種類が多いことがあげられる。図 1 は読みの数に対する漢字割合の分布を示している。ここからわかるように、日本語の漢字はほとんどの場合に複数の読みがあるのに対し、中国語においては、ほとんどの場合一つの読みしかない。したがって、拼読 は日中でまったく対称というわけではなく、日本人のための中国語検索システムの方が候補が多く挙がり、実現が難しくなることを示唆する。

3　システム

3.1　ユーザインターフェース

拼読のテーマは、漢字から成る単語を調べることである。したがって、母語で対応する漢字を入力することができさえすれば、直接漢字対応に基づいて調べることができ、あえて読みを入力して曖昧性を増やして事態を複雑にする必要はない。しかし、つぎの三つの理由をふまえ、拼読では読みからも引けるようにした。

第一に、対応する母語の漢字がわからない場合に、漢字部分の読みなどから検索が可能な場合もあるからである。第二に、外国語の単語を対応する母語の漢字列を入力する場合は、単漢字入力を繰り返すこととなってしまい、困難となる場合が多いからである。第三に、母語の漢字を入力してから探すのでは、読みから母語への漢字列へ一度変換し、その上でさらに外国語の漢字列を調べる二段階の処理になってしまう。一方、読みから引く場合には、一段階の処理で済むし、またどの言語にも一般的な「読みから辞書を引く」感覚で外国語の単語を引くことができるという利点がある。

以上から、拼読では、ユーザは外国語の単語が与えられたとき、つぎのいずれかを入力することができる。

- 自国語の読み（日本字の場合はかな、中国人の場合には pinyin)
- 自国語の対応する漢字
- 読み、対応する漢字ともわからない場合には X

読みと漢字は現在のところ同時に用いることができないが、X は読み、漢字のいずれの場合にも混入させることができる。

たとえば「电视」を引きたい場合には、漢字対応に想像が付けば「電視」あるいは「でん・し」と入力することができるし、第一番目の漢字の対応に想像が付かなければ、「X し」などと入力することができる。図 2 に、「でん・し」と入力して実際に引いている例を示す。日本語で「でん・し」と読むことのできる中国語の単語としては複数のものがあるため、図のようになる。第一番目の候補から「电视」はテレビの意味であることがわかる。

図 2: かな入力「でん・し」で中国語を調べた結果

3.2 候補生成

ユーザの入力が母語の漢字である場合の処理は、入力が読みである場合に含まれる。そこで、ここではユーザの入力が母語の読みであるものとして、入力からどのように出力が生成されるかを説明する。

s を漢字から成る外国語の単語、$\phi(s)$ を s を母語漢字に対応させた漢字列の集合、$\phi(s)$ の要素を t、t の読みを r で表すものとする。r が実際の入力で、母語の読みである。たとえば、s=「电视」、$\phi(s)$ は図 2 に挙がった候補の集合で、r=「でん・し」となる。

拼読では候補は $P(s|r)$ の降順に挙げる。$P(s|r)$ は、以下のように変形することができる。

$$\begin{aligned}\Pr(s|r) &\propto \Pr(r|s)\Pr(s)\\ &= \Pr(s)\sum_{t\in\phi(s)}\Pr(r,t|s)\\ &= \Pr(s)\sum_{t\in\phi(s)}\Pr(r|t,s)\Pr(t|s)\\ &= \Pr(s)\sum_{t\in\phi(s)}\Pr(r|t)\Pr(t|s) \qquad (1)\end{aligned}$$

すなわち、コーパスから計測しなければならない項としては、$\Pr(s)$, $\Pr(r|t)$, $\Pr(t|s)$ の 3 つがある。それぞれつぎの簡単なモデルを用いる。

- $\Pr(s)$ については、単語頻度でモデル化した。
- $\Pr(t|s)$ については、[4] を利用して求めた。$t \in \phi(s)$ について母語のコーパス中の相対頻度でモデル化する。
- $\Pr(r|t)$ は、ある漢字列 t が r と読まれる確率であるが、t を成す漢字列を t_i、それに対応する r 中の読み部分を r_i とし、

$$\begin{aligned}\Pr(r|t) &= \Pr(r_1\ldots r_n|t_1\ldots t_n)\\ &\approx \prod_{i=1}^{n}\Pr(r_i|t_i) \qquad (2)\end{aligned}$$

で近似した。$\Pr(r_i|t_i)$ は、t_i に対応する読みの内の r_i の相対頻度とした。

以上の言語モデルを求めるため、日本語コーパスは毎日新聞コーパス (100MB)、中国語は北京大学コーパス (200MB) を用いた。このほか、漢字の読みの辞書として、日本語は Kanjidic 辞書[2]、中国語の場合には GBK の pinyin の辞書を利用した。

X については、X が表す漢字の集合が読み X を持つものとして、上記の処理をそのまま用いた。といって、X を単純なワイルドカードとしてすべての漢字に対応させていては、X が入力された場合の候補の数が非常に多くなってしまう。ユーザが X を入力する場合には、

- **unmapped:** 漢字に対応がない場合
- **unknown:** 母語への漢字への対応がわからない場合
- **unreadable:** 対応はあっても使用頻度の差などから、母語で読めない場合

の三つの場合がある。そこで、日中で対応する漢字について、コーパス中の利用頻度の差がある与えられた閾値より大きい場合、ならびに、フォントの bitmap 上での形状の差異がある閾値より大きいものを X に割り当てるものとした。$\Pr(\mathrm{X}|s_i)$ は一様分布とした。

表 1: 目的の単語の候補中の平均順位

	かなによる中国語の検索		**pinyin** による日本語の検索	
単語長	Best	Random	Best	Random
1	9.41	12.40	6.65	6.66
2	1.63	2.02	2.00	2.00
3	1.01	1.01	1.12	1.12
≥4	1.00	1.00	1.00	1.00

4 評価

まず、各単語につき読みを入力した際に挙がる候補中の順位を統計的に調査した結果を表 1 に示す。単語長が 1,2,3,4 以上の時別に、各単語について、最も高い確率の読みを入力した場合 (Best)、ランダムに読みを選択して入力した場合 (Random) について挙げた。

単語長が 2 以上のときには、平均順位は比較的高く、現実的な値となっている。単語長が 3 以上となると、

[2] http://www.csse.monash.edu.au/~jwb/kanjidic.html

表 2: ユーザ入力に基づく検索性能

	順位	候補数	検索成功率
	かなによる中国語の検索		
読み上位 1	1.85	10.24	0.61
読み上位 3 まで	1.99	10.01	0.64
	Pinyin による日本語の検索		
読み上位 1	1.31	2.54	0.81
読み上位 3 まで	1.30	2.54	0.83

表 3: エラー解析結果

エラーの種	かな	pinyin
読み	14.5%	37.6%
形状	43.5%	22.0%
X	25.8%	10.8%
その他	16.1%	29.6%

順位はほとんど 1.0 に近くなった。一方で、単漢字で本システムを用いるのは、特に日本語において現実的でないことがわかる。

さらに、実際にユーザが外国語の漢字から成る単語を見てどのように読みを入力するかのアンケート調査を行い、その答えに基づき、拼読を調べた際の候補の順位、候補平均総数、検索成功率を調べた。12 人の日本人、28 人の中国人に、ランダムに事前に選んだ 100 の日中の単語の中からさらにランダムに 30 ずつユーザに見せ、各単語につき最大で 3 つづつ読みを答えてもらった。

ユーザ入力をもとに実際に元の単語が候補に挙がった場合に成功とみなして、検索成功率を求める。というのも、ユーザの入力には、入力ミスや読み間違いなどが含まれるために、必ずしも検索が成功するとは限らないのである。

結果は表 2 となった。日本語は中国語に比べ、漢字に複数の読みがある分、順位が低く候補数が多くなり、検索成功率が低くなっている。

検索不成功であった部分については、その理由をすべて解析し、結果を表 3 に示した。「読み」とは、かなや pinyin 入力上の問題を示す。たとえば、日本語には連濁や音便などがあり、これは現在のシステムでは適切に処理されていない。また、漢字の読みを送り仮名混じりで入力した場合も、日本語の漢字とその読みの辞書に登録されていないために、候補は挙がらない。「読み」の問題は、特に中国人の場合には目立つ。これは pinyin の入力誤りが原因であった。「形状」については、読めない漢字について、部首や共通の部首を持つ別の漢字の読みを入力して、それが誤りであった場合である。日本人が中国語を見る場合に特にこれが多かった理由は、中国語の漢字が簡体字であることが挙げられる。この点は、漢字対応を部首や類似漢字へと広げることで解決できよう。「X」とは、X を入力に含めたために、候補が多くなりすぎてしまった場合をいう。これを見ると、X の使い方を制限するなどの工夫が今後は必要となる。

より楽しい日中の言語交流を目指して、今後は拼読システムをよりよくしていきたい。

参考文献

[1] N. AbdulJaleel and L. S. Larkey. Statistical transliteration for English-Arabic cross language information retrieval. In *CIKM*, pages 353–360, 2003.

[2] S. Bilac, T. Baldwin, and H. Tanaka. Bringing the dictionary to the user: the FOKS system. In *Proc. 19th International Conference on Computational Linguistics*, pages 85–91, 2002.

[3] E. Brill, G. Kacmarcik, and C.Brockett. Automatically harvesting katakana-English term pairs from search engine query logs. In *NLPRS*, pages 393–399, 2001.

[4] C. Goh, M. Asahara, and Y. Matsumoto. Building a Japanese-Chinese dictionary using kanji/hanzi conversion. *Proceedings of IJCNLP 2005*, pages 670–681, 2005.

[5] S. D. Krashen. *Second Language Acquisition and Second language Learning*. Oxford:Pergamon, 1981.

[6] T. Odlin. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge University Press, 1989.

[7] Y. Qu, G. Grefenstette, and D. A. Evans. Automatic transliteration for japanese-to-english transliteration. In *SIGIR*, pages 353–360, 2003.

[8] P. Virga and S. Khudanpur. Transliteration of proper names in cross-language applications. In *SIGIR*, pages 365–366, 2003.