知的な辞書FOKS

誤読でも検索可能な辞書システム

東京工業大学大学院情報理工学研究科後期課程 Slaven Bilac (博士課程 2 年、計算工学)
Stanford University, CSLI 研究員 Timothy Baldwin (2001 博士課程卒、計算工学)
東京工業大学大学院情報理工学研究科教授 田中穂積 (1966 修, 制御工学)

1 はじめに

８０年代から電子辞書が普及してきた。電子辞書の特長は検索の早さ、装置の軽さの他に使いやすさをあげることができる。一つの装置に複数の辞書、たとえば、国語辞書、類似語辞書、和英辞書などを入れておけば、辞書から辞書へと簡単に移ることもできる。検索速度は紙辞書の検索に比べて圧倒的に早い。

熟語の場合には読みがわかればその熟語の検索が容易である。しかし熟語が含まれている個々の漢字の読みを知っていても熟語全体の読みが分からないことが多い。電子辞書では、(読みではなく) 漢字の熟語をそのままの形で入力して、その熟語の辞書項目を検索することができる。このことは、熟語の読みが分からない日本語学習中途の留学生などには、都合のよい辞書引き機能のように思えるかも知れない。カナ漢字変換システムを用いて、熟語を漢字の形に変換することができそうに思えるからである。ところが熟語の読みが分からなければ、カナ漢字変換システムは使えない。ある程度の漢字の読解ができる日本語学習者なら、一文字毎の読みを与えたカナ漢字変換を繰り返すことも考えられるが、それは効率が悪いだけでなく、たとえば「旅」という漢字をいつも「りょう」と誤読して覚えた日本語学習者の場合には、カナ漢字変換システムは役に立たない。

そこで個々の漢字の読みある程度を知っている日本語学習者が、熟語を誤読しても辞書引きできるシステムがあれば便利である。しかも個々の漢字の読みを誤読して憶えていても辞書検索できるシステムがあれば、さらに便利である。たとえば、「春」を「はる」、「雨」を「あめ」と読み、「春雨」を「はるあめ」と誤読しても、「はるさめ」という正しい読みとともに、その辞書項目を検索し表示するシステムの開発である。実際、我々はそのようなシステム FOKS を開発し公開している。FOKS とは、寛大なオンライン漢字検索システム (Forgiving Online Kanji Search) という意味が込められている。FOKS は、誤読を正しい読みに変換してから辞書を検索する方式をとっていない。誤読に対して、誤読の可能性のある辞書項目の候補をスコア順に表示し、そのなかからユーザに適切なものを選択させ、それを正しい読みとともに表示する。

そのために、FOKS は誤読の原因を分析し、確率的なモデルを用いて、辞書引きの対象となる熟語の考えられる読み (正確あるいは不正確) を全て生成しておき、それらにあらかじめスコアを付けておく。このようなデータをあらかじめ作成することにより、ユーザが推測した読み (正確あるいは不正確であるかは無関係) に対して、そのような読みに対応する辞書項目の候補をすべて抽出することができる。

たとえば、今読んでいる本の中に「山車」という熟語が現れ、それに対して「さんしゃ」と誤読した入力を FOKS に与えても、FOKS は直接「山車 (だし)」、「三者 (さんしゃ)」などを表示することができる。これは、「山車」と「三者」の正しい読みと誤読のリストのなかに、「さんしゃ」が含まれているからである。表示する順番は、あらかじめ計算しておいたスコアの順に並べる。そこで、ユーザが表示さ

れた候補のなかの「山車」を選択すると（これは本の中の活字の「山車」と比べる）、FOKS は、正しい読みと共に説明文を出力する。

FOKS は WWW の上に実装し、一般に公開されている。`http://www.foks.info/` でアクセスできるので使ってみて欲しい。”Using is Believing” かもしれない。FOKS のキーポイントは、各熟語に対して正しい読みと考えられる誤読のリストと読みのスコアを予め作成しておくことであった。これについては、次節以降で簡単に説明する。

2　誤読の原因

2.1　日本語の表記について

日本語の文字は三種類（平仮名、片仮名、漢字）ある[1]。

平仮名と片仮名はそれぞれ 100 文字程度であるので覚えるのが容易であるだけでなく、音声表記（読み）とほぼ一対一に対応している。

一方、漢字は数が多いだけでなく、それらの読みが大きな問題となる。1981 年に日本政府が常用漢字の数を 1945 字としたが、実際には新聞や本には約 3000 字程度が使われている。漢字は表意文字であり、 意味があると同時に読みもある。その読みに「音」と「訓」があることも、日本語習得者にとって大きな負担になる。

たとえば、「頭」の訓読みと音読みは、それぞれ「あたま、かしら」と「とう、ず」で、「上」の訓読みと音読みは、それぞれ「うえ、かみ、あがる、のぼる」と「じょう、しょう」である。留学生などは熟語の「頭上」を辞書で調べる場合に、「ずじょう」という正しい読みが推測できない可能性がある。「あたまうえ」、「とうじょう」など誤読してしまう可能性がある。普通の電子辞書では誤読すると辞書引きに失敗する。

また、音読みと訓読みの組合せだけでは、熟語の正しい読みが作り出せない場合も多い。たとえば、「山」の訓読みと音読みはそれぞれ「やま」と「さん」であり、「車」の訓読みと音読みは、それぞれ「くるま」と「しゃ」であり、これらを組み合わせた読みからは、「山車」の正しい読み「だし」を作り出すことができない。これは、当て字と呼ばれる現象で、音訓読みの組合せからは正しい読みが推測できないので、個々に覚える必要がある。覚えない限りは通常の辞書検索が失敗する。FOKS は、このような場合にも正しい読みとともに、その辞書引きができるシステムである。

[1]最近はアルファベットも用いることが普及しているが、ここではアルファベットを考慮しないことにする。

2.2　原因の分類基準

それではどのような誤読が実際にあったかを調べてみよう。

1. **音訓読みの組合せを誤る場合 :** たとえば、「頭上（ずじょう）」を「とうじょう」もしくは「あたまうえ」読んでしまう。

2. **当て字の場合 :** たとえば、「山車（だし）」の他に「海星（ひとで）」を「うみぼし」や「かいせい」と読んでしまう。

3. **音韻変化の場合**。たとえば、「国境（こっきょう）」を「こくきょう 」と読んでしまう。

4. **長音と短音を区別できない場合**。 たとえば、、「旅行（りょこう）」を「りょうこう」と読んでしまう。これは留学生などによくみられる読みの誤りである。

5. **よく共起する字の読みを交換してしまう場合**。たとえば、「挨拶」に対して正しい読みができても、「拶」という字が単独に現れた場合に「あい」か「さつ」か迷ってしまう。

6. **意味類似の場合**。たとえば、「右」と「左」は意味がとても近いので「右側」を「ひだりがわ」で読んでしまう。これも留学生などによくみられる読みの誤りである。

7. **形態類似の場合**。たとええば、「基」と「墓」はよく似ているので「基地」を「ぼち」で読んでしまう。逆に「墓地」を「きち」と読んでしまう。

8. **その他**。たとえば、読みの打ち間違いなど。

上記した種々の誤読をしていても辞書引きができるシステムが FOKS である。その概要を以下に説明する。

3　システムの概要

まずはじめに、システムの概要を簡単に説明する。我々が漢字の読みを学習する時には、よく現れる読み（頻度が高い読み）から頻度の低い読みの順に覚えるのが自然である。そして読みがいづれかが分からない場合には、より高い頻度の読みを優先させる。このような人間に近い読みの学習法をまねるために、多数の例文に現れる多数の漢字を抽出し、それらに対して全ての読みと読みの頻度を与える (3.1、3.2 参照)。次に、熟語に対して、熟語を構成する各漢字の読みを組み合わせて、熟語の読み（誤読も含む）の候補を生成する。このようにして生成した、熟語のあらゆる読みの候補には、各漢字の読みの頻度を使って、さらに確率的なスコアをつけ、このようなデータを全ての熟語について集めてひとつのリストを作る（3.3 参照)。これらをベースにして誤読であっても正しい辞書引きが可能になる FOKS を構築した。

3.1　読みの抽出

漢和辞書には各漢字に対して読みが書かれているが、その読みには頻度も書かれてないし、音韻変化したものなども書かれていない。そのため漢和辞書だけでは漢字の読みの漏れがある。従って、実際の文に現れる漢字の読みを集める必要がある。集めた漢字の読みのなかには音韻変化を伴うものなども、ほぼ網羅的に含まれているからである。

まずはじめに、辞書項目（見出し語）として現れる漢字とその読みを抽出する。そのため辞書の見出し語とその読みから、見出し語に含まれる各漢字の読みを自動的に対応させる。たとえば、

＜発表ーはっぴょう＞ ⇒

＜発ーはっ＞、＜表ーぴょう＞

＜風邪薬ーかぜぐすり＞ ⇒

＜風邪ーかぜ＞、＜薬ーぐすり＞

とする。この時、＜薬ーぐすり＞のように漢字一文字だけなく、＜風邪ーかぜ＞のように複数個の漢字列に対して読みを対応させることもある。両者とも以下ではユニットとよぶことにする。対応方法についてはここでは詳しく述べない。関心がある読者は (Bilac et al. 2002)[2]文献を参照してほしい。

3.2　音韻変化

抽出した漢字の読みには音韻変化した読みも含まれている。しかし、音韻変化したものと音韻変化を含まない読みとを区別する必要がある。区別することにより音韻変化の確率が抽出できるからである。たとえば、「薬」の音韻変化が含まれない読みが「くすり」であることが分かれば、「くすり」の「く」が「ぐ」に変化する確率が頻度情報から計算できる。それにより、「風邪薬」を「かぜくすり」と誤読する確率が計算できる （図 1 参照)。

ある読みが音韻変化を含むか、含まない（基本的な）読みであるかを自動的に判別するために次のようにする。たとえば、「表」に対して「ぴょう」という読みは単独では現れない。一方、「ひょう」は単独でも現れるので、「ぴょう」は「ひょう」という基本読みの半連濁されたものと判断する。このようにして、先の例を変換して得た結果を以下に示す。

＜発ーはっ＞|＜表ーぴょう＞ ⇒

＜発ーはつ ＋音便化＞|＜表ーひょう ＋半濁音化＞

＜風邪ーかぜ＞、＜薬ーぐすり＞ ⇒

＜風邪ーかぜ＞、＜薬ーくすり＋濁音化＞

このような変換後の各ユニット（＜ … ＞の部分）の可能な読みに対する頻度から、それぞれの確率を

[2]S. Bilac, T. Baldwin, and H. Tanaka. Bringing the dictionary to the user: the FOKS system. In *Proc. of COLING 2002*, Taipei, Taiwan, 2002.

計算する。たとえば「表」という文字に対して、「ひょう」読みが 3 回現れて「おもて」読みが 2 回だけ現れたとすると、それぞれの読み確率は 60%(3/5) と 40 %(2/5) となる。さらに、「ひょう」が音韻変化して「ぴょう」になる場合には、音韻変化する頻度から「ひょ」が「ぴょ」になる確率も計算しておく。このようにして各熟語の可能な読みとそれに対する確率的スコアを計算する準備をしておく。

3.3 新たな読み生成とスコア付け

前節で述べた各ユニットの可能な読みから、それらを組み合わせた熟語の可能な読みを生成する。たとえば、「発表」は二つのユニット「発」と「表」からなり、それらの読みを組み合わせて「はつひょう」、「はつおもて」、「はっひょう」、「はっおもて」などを生成し、それぞれのユニット確率の積を熟語読み確率とする。さらに、「はっひょう」、「はっおもて」には音韻変化を含むので、それらの熟語の読みの確率には前節で述べた音韻変化の確率をさらに掛け、それを最終的な熟語の読みの確率とする。

さらに、熟語の読みの確率には熟語の使われる頻度情報をも反映させる。そのために大量の文例データを収集しておく。このようにして、頻度の高い熟語ほど優先させた検索結果をえることができる。たとえば、「さんしゃ」という読みが、「山車」と「三者」の読みのリストに存在するとしよう。そして、大量の文例データから「山車」が「三者」より使用頻度が高いことが判明した時、それぞれの使用頻度を熟語の読み確率に掛け、それを最終的な熟語読みのスコアとする。このようにして「山車」の方を優先させて、「さんしゃ」の辞書検索結果とすることができる。

たとえば、「山車」に対して生成した読みリストとそのスコアを表 1 に示す。生成した読みリストのなかに正しい読みが含まれていない場合には、辞書の見出し語の正しい読みであっても最低の熟語読み確率（一定値）に使用頻度を掛けたものをスコアとして与えて読みのリストに加える。たとえば、表 1[3] の場合は「だし」という読みが生成されなかったため、低いスコアが付けられている。

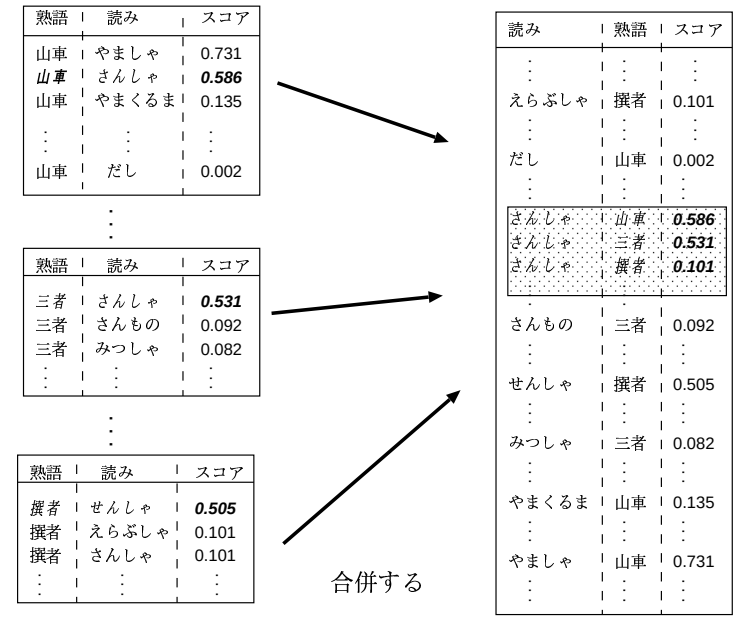
熟語	読み	スコア
山車	やましゃ	0.731902
山車	さんしゃ	0.586403
山車	やまくるま	0.13528
山車	さんくるま	0.108387
山車	ざんしゃ	0.101858
山車	やまじゃ	0.0861448
山車	さんじゃ	0.0690196
山車	やましや	0.0365951
山車	やまぐるま	0.0322644
山車	さんしや	0.0293202
⋮	⋮	⋮
山車	やまじや	0.00430724
山車	さんじや	0.00345098
山車	だし	0.0002

表 1: 「山車」に対して生成した読みとスコア

最後に FOKS が用いる辞書検索用読みリストを次のようにして生成する。これまで述べた方法で、各熟語の読みとそのスコアのリストを生成する。それらを合併してひとつの表を作り、次に読みを見出し語にして並べたリストを作る（図 1(a)&(b) 参照）。これが FOKS が用いる誤読にも対応した読みの最終リストになる。このリストを用いることにより FOKS は、ユーザが入力した読みに対する全ての熟語を簡単に抽出することができる。たとえば、「さんしゃ」を入力すると、図 1 の網がけの部分を抽出することができる。システムはさらに「山車、三者、撰者」などによる辞書引きにより、熟語に正しい読みを対応させるとともに、表 2 に示す候補をユーザに表示する。最後に、ユーザは表示された熟語の中から適切なものを選び、それをクリックすると、その熟語の説明文または訳語が表示される。

[3] 読みのリストには音韻変化や母音の長短音化させた読みなども含める。

[3] ただし、確率的なスコアは誤読のスコアのままで表示する。



a) 熟語の読みとそのスコアのリスト

b) 合併語のリストを読みを見出し語にしたリスト

図 1: FOKS の辞書検索用読みリストの生成

4 システム実装

前節で説明したように、FOKS は、各熟語に対して可能な読みのリストを自動的作り出すが、その時使用した辞書は、一般公開されている EDICT[4] という和英辞書を用いた。EDICT は 10 万以上辞書項目を持ち、その内約 9 万の熟語が含まれている。熟語の使用頻度は約 20 万文の例文集[5] から計算した。FOKS では、（自動的に生成した）図 1(b) に示す読みを見出しにした表を WWW サーバにおき利用している。

動作例を図 2 に示す。FOKS(`http://www.foks.info/`) を動作させて、「春雨」という熟語に対して、「はるあめ」と入力すると、図 2 に示す画面が表示される（正しい」読みも表示されていることに注意)。ここで、Translation キーをクリックすれば、ただちに EDICT のもつ「春雨」の訳語が画面に現れる。

FOKS の主な特長を二つ挙げる。第一に、すべての可能な読みを前処理し、自動的生成し、図 1(b) に示すリストを予めに作っておくので、誤読に対する高速な辞書検索が可能である。第二に、表示される候補は、辞書と大量文例データから抽出された情報から計算された（確率に基づく）スコアの順に並べられている。

見出し語	正確な読み	スコア
山車	だし	0.5864[6]
三者	さんしゃ	0.5313
撰者	せんじゃ	0.1011
三叉	さんさ	0.0744
産物	さんぶつ	0.0341
操車	そうしゃ	0.0078

表 2: 「さんしゃ」に対する見出し語の候補

[4] `ftp://ftp.cc.monash.edu.au/pub/nihongo/`

[5] 日本電子化辞書研究所. *EDR* **電子化辞書** *1.6* **版仕様説明書**, 1995.

図 2: FOKS の動作例

5 おわりに

FOKS システムは 2002 年 10 月に本格的に稼働し公開されている。現在までにおよそ 11 万件の検索要求があった。

簡単な評価をしておく。一つの見出し語に対して生成された読みの数は、平均 48 個であった。一方、ユーザが入力する読みに対して FOKS が出力する見出し語の候補は平均して 1.21 と極めて少数である。30 以上の候補が表示されることもある。このような場合は総数 4,549,152 の読みの内約 0.02%個の読みを入力した場合に過ぎない。従って、FOKS は、一般に読みの入力に対しては、表示される候補の数が少ないので、正しい熟語の選択が容易である。

FOKS は、漢字の読みを習得し始めた留学生などには、熟語の読みの学習や、その意味などを知るためのシステムとして、実際に使われている。このような日本語学習者だけでなく、日本人にとっても有用なシステムである。たとえば日経産業新聞で取り上げられた FOKS の記事によると、「海星」の読みが分からずに「うみほし」と FOKS に入力して、正しい読みが「ひとで」であることを知ったという報告などがなされている。