日本語校正支援システム (Joyner) の研究について (3)
－正解語探索－

2J－6

中村 直人　徐 国偉　伊吹 潤　松井 くにお
富士通研究所

1　はじめに

従来，片仮名表記の揺れ誤りや仮名漢字変換誤りなど単語の綴り誤りに対して，それぞれ校正処理の枠組が提案されている．Joyner ではこれらの誤りを単一の枠組で扱うことを試みた．Joyner の処理手順は，文に含まれる各種単語の綴り誤りの検出と候補の推定を行ない，推定した候補を原文の別綴り可能性 (綴り曖昧さ) とする．そして，文の綴り曖昧さの中から最尤解を探索し (正解語探索)，最尤解の綴りが原文と異なる時に，原文に綴り誤りがあると推定する [1].

綴り曖昧さのある文の最尤解釈の探索は，文字認識の分野で文字切り出し曖昧さの処理として検討され，2 端子グラフ (ラティス) の最短経路を求める動的計画法で処理できることが知られている [2]．これを Joyner に適用し，動的計画法の一種である CYK 法で実現した．本稿では，校正支援のためのラティスを CYK 法で処理する方法について Joyner での実現を中心に報告する．

2　曖昧さを付加した綴りのラティス

Joyner は，入力文の単語や文字に対して，考えられる正しい綴りの候補を曖昧さとして原文に付加し (展開)，これをラティスの構造で表す (図 1)．文字認識のラティスでは辺が文字とその信頼度を持つのに対し，Joyner のラティスでは文字列とコストを持つ．

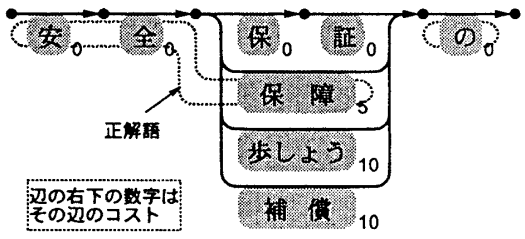

図 1: ラティスの例 (入力:「安全保証の」)

2.1　辺に対応付けた文字列

形態素解析の際にはラティス中の辺の並び (パス) に従って解析する文字列を組み立てる．この時，辺が持つ文字列は組み立ての単位となる．これは，単に組み立てるという意味だけでなく，辞書引きの単位としての意味を持たせた．つまり，辺一つまたは複数の辺を連結した表記の形態素を認めるが，一つの辺の文字列を分割するような形態素は認めない．この解釈を設けることで，原文を展開してラティスを発生する時に意識した曖昧さのパターンをより忠実に表現でき，その結果，形態素解析での無用な探索を排除できた．

2.2　辺に対応付けたコスト

辺のコストは，その辺を用いて合成した綴りの尤度の低さを表す．辺のコストを設定する一般的な方針は，原文に含まれる綴りのままである辺に 0，展開で生成した辺に正数を与える．特に，並行したパスの間では，推定した誤りの不自然さに差があれば大小関係を持たせる．このコストは，形態素解析での最尤評価 (コスト最小解の探索) において，他のコスト要素 (形態素自体の特徴に対するコスト，隣接する形態素の組合せによるコスト) と合算される．このようにして，最尤解の選択に辺が持つコストが加味される．

3　CYK 法による解析

Joyner では CYK 法を用いた形態素解析を正解語探索に用いた．その理由は，CYK 法を用いた形態素解析システムが手元に既にあり開発上有利であったこともあるが，節 2.1で述べた辞書引きをする表記の組み立てに関する制約が直感的に表現できる (後述) からである．

以下では，辞書引きをする表記の組み立てに関する制約を CYK 表に表す方法と，その結果を用いて辞書引きを行う方法 (CYK 表の初期化) について述べる．CYK 法による形態素解析の説明は省略する．

3.1　ラティスの CYK 表への割り付け

文字列の形態素解析の場合，文字列と CYK 表との対応付けは文字列を先頭から一字ずつ割り付ける．あ

Japanese spelling corrector(Joyner)(3)
Naohito NAKAMURA, Guowei XU, Jun IBUKI, Kunio MATSUI
Fujitsu Laboratories Ltd.
1015 Kamikodanaka, Nakahara-ku, Kawasaki-shi 211 Japan

とはCYK表に沿って辞書引きを行うことで，文字単位(割り付け単位)で辞書引きが行われる．文字列の形態素解析の場合は特に意識するまでもないが，入力とCYK表との割り付けがこのように後の辞書引きの単位を決定している．既に述べたように，Joynerではラティスの辺を辞書引きをする表記の構成単位にする必要があるので，各辺をCYK表に割り付けることにした．

CYK表上での形態素解析は，先頭文字(あるいは，文頭形態素)の列を辞書引きをして得た単語で末尾に向かって直線的に成長させることで完了する．同じような手順でラティスを手戻りなく検索するには，ラティスの先頭の辺から末尾に向かう順序で辺をCYK表へ割り付けなければならない．この制約を，次のように捉えた．ここで，「位置」はCYK表の文頭から数えた列の番号を表す．また，辞書引きの際に必要な『仮想長』を導入した．仮想長の解釈は後で述べる．

- 二辺にラティス上で順序関係が成立すれば，それらの位置の大小関係はこれを保存する
- 始点が同一な辺はすべて同じ位置を持ち，かつ，同じ位置を持つ辺はすべて同一の点を始点とする
- ラティス上で連続している2つの辺(分岐点をはさんで連続する場合も含む)の間の連続性を保存する(辺の位置にその仮想長を加えた値が後続の辺の位置となる)

図1に示したラティスでこの割り付けを行った様子を図2に示す．ここでCYK表上の同一の位置に割り付けられた複数の辺において，それぞれ異なった仮想長を持ち得ることに注意されたい．

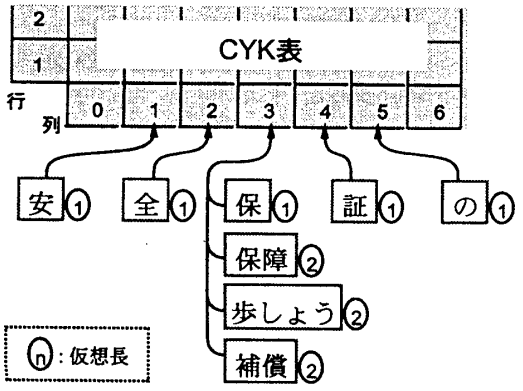

図 2: CYK表への辺の割り付け

3.2 辞書引き

前節で確認したように，文字列の形態素解析の場合，CYK表の隣あった列(文字に対応)を組み合わせた文字列を表記として辞書引きを行っている．この隣あった列を順次結合して表記を生成してよいという性質を，『各列の仮想長が1である』と捉えることにした．つまり，ある列iに後続させて表記生成に使って良い列jは，列iから列iの仮想長だけ進めた列であると考えた．

辺を割り付けたCYK表の辞書引きは，前節の制約に従って定めた仮想長を用いて，この解釈を行えばよい．例えば図2で列2に割り付けた「全」から始まる辞書引きは表1のようになる．ここで，辞書引きの結果を登録する行が，結合した辺の仮想長の合計であることに注意されたい．

表 1: 辞書引きをする表記

表記: 辺の表記 (位置, 仮想長),···	結果を登録する行
全 (2,1)	1
全 (2,1), 保 (3,1)	2
全 (2,1), 保 (3,1), 証 (4,1)	3
全 (2,1), 保 (3,1), 証 (4,1), の (5,1)	4
全 (2,1), 保障 (3,2)	3
全 (2,1), 保障 (3,2), の (5,1)	4
全 (2,1), 歩しょう (3,2)	3
⋮	

我々はこの辞書引きを，ダブル配列を用いたトライ辞書機能を拡張することで実現した．なお，表1を見ると表記「全保」や「全保証…」の表記系列，「全保障…」の表記系列があり，これらを「全保」まで併合する辞書引きの手間が減らせると考えられる．しかし，併合処理(検索を伴う)よりトライ辞書の引き直しの処理コストが小さいと考えて併合処理は行っていない．

4 まとめ

綴り曖昧さをもつ文を表すラティスからCYK法で最尤解を求める校正支援システムを作成し，そこで採用したデータの解釈と処理方法について報告した．

最後に，正解語探索のベースとなったシステムの提供や技術検討をしてくれた颯々野研究員を始め，西野研究員および小川研究員に感謝する．

参考文献

[1] 松井他: “日本語校正支援システム (Joyner) の研究について (1) −綴り誤り自動訂正について−”, 情処学会第52回大会, 2J-4(1996)

[2] 村瀬他: “言語情報を利用した手書き文字列からの文字切り出しと認識”, 信学論文誌, Vol.J69-D, No.9 (1986)