文書画像を対象とした未知単語の抽出法

1M－8

仙田修司　美濃導彦　池田克夫
京都大学工学部

1　はじめに

データベースに蓄積された文書を効率的に検索することを目的として，文書の自動分類[1]やキーワードの自動抽出[2]といった研究が行われている．これらの研究では，文書中に出現する単語の分布を手がかりにするので，人手をかけずに大量の文書を処理することができる．同様に，単語の共起関係に基づいて検索質問の作成を支援する研究[3]においても，文書中から抽出した単語を利用している．

筆者らは，与えられたキーワードの文書画像中での位置を高速に捜し出すことのできる文書画像検索システムについて報告した[4]．このシステムは，文書画像の文字認識結果を曖昧な仮説(文字ラティス)の形で蓄積するので，文字認識率が低い場合でも高い検索率を達成している．このシステムで用いた検索手法は同時に複数のキーワードを検索することができるので，あらかじめ用意した辞書単語を与えることによって，文書画像から辞書単語の抽出を行うことができる．これによって，文書画像の自動分類や文書画像からのキーワードの自動抽出が可能となる．しかし，辞書中に存在しない単語(未知単語)は，文書を特徴づける重要なキーワードとなっていることが多いにもかかわらず，上記の手法では抽出することができない．

本稿では，日本語で書かれた文書画像を対象として，日本語の性質と単語の出現頻度を利用した未知単語抽出法を提案する．さらに，実験によって，本手法の有効性を示す．

2　文字ラティス

文字ラティスは，文字切り出し位置の候補をノード，文字の候補とその確信度をアークとする有向グラフで表される(図1)．文字ピッチや文字フォントが既知でない文書画像を対象とした場合，文字切り出し，文字認識のそれぞれの処理の結果を一意に決定してしまうことは非常に難しく，言語的な知識を用いた後処理を行うことが望ましい．文字ラティスは，そのような，文字認識処理の結果を後処理に渡すためのデータ構造として非常に有用である．

文字ラティスの有用性を示すために，サンプル文書画像(文字数1515)について，候補を一つに絞った場合と文字ラティスによって平均2.6個の候補を持った場合とを比較した．候補を一意に決定する手法として

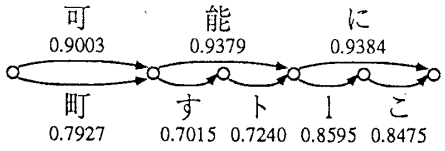

図1: 文字ラティスの例

は，文字候補の確信度の高いものから順に決定していくこととした．その結果，候補を一つに絞った場合，正解率が92.6%，切り出し誤りが0.89%，文字認識誤りが6.5%であった．それに対して，文字候補数(文字ラティスのアーク数)が3917の文字ラティスの場合には，正解を含む割合(再現率)は99.9%であった．このことは，文字切り出しおよび文字認識の処理結果を文字ラティスで表現することにより，後処理次第で，認識率を92.6%から99.9%に向上させることが可能であることを示している．

3　字種による未知単語抽出法

本稿でいう未知単語は，辞書中に存在しない単語である．よって，未知単語抽出法とは，辞書を用いない単語抽出法のことであり，辞書を用いる辞書単語抽出法とは区別する．本節では，文字ラティスからの未知単語抽出法の基礎となる“字種による未知単語抽出法”を示し，その有効性を通常のテキストに対して調べる．

まず，日本語の文字をその字種によって，カタカナ，漢字，英数字，その他の4種類に分類する．そして，カタカナ(もしくは漢字，英数字)のみからなる文字列のうち，長さが2文字以上で極大のものを単語として抽出する．ただし，極大な文字列とは，他の文字列の一部になっていない文字列のことをいう．この手法は，同一字種の連接からなる文字列は単語を形成し易いという日本語の性質に基づいたものであるため，ひらがなを含む単語や1文字の単語は抽出できない．しかし，後の実験でも示す通り，こうした単語の数はそれほど多くないことから，4節の文字ラティスからの未知単語抽出法では抽出の対象外とする．

この“字種による未知単語抽出法”によって，どの程度の単語が正しく抽出されるかを調査した．サンプル文書(文字数1515)に上記の手法を適用した結果，256個の単語が抽出された．そのうち，明らかに誤りであるものは，動詞(“区切る”)の語幹が1個，名詞(“二通り”，“入力誤り”)の一部が2個であった．よって，

Extraction Method of Unknown Words from Document Images
Shuji SENDA, Michihiko MINOH, Katsuo IKEDA
Faculty of Engineering, Kyoto University

切り出し候補数(文字ラティスのノード数－1)は2061であった．

抽出された単語のうち正しいものの割合 (適合率) は，98.8% である．ただし，形容詞 (“容易に” など) の語幹 15 個と，複合名詞 (“指向錯誤”，“誤入力” など)36 個は正解であるとした．逆に，サンプル文書に含まれる単語のうち，9 個の 1 文字名詞 (“木” など) が抽出されなかった．よって，抽出すべき 262 個の単語のうち実際に抽出できたものの割合 (再現率) は，96.6% であった．これらの結果から，字種による単語抽出法は，単純な手法であるが十分に有効であることが確認できた．

4 文字ラティスからの未知単語抽出法

文字ラティスからの未知単語抽出法は，以下の三つの Step で構成される．

Step1 字種による単語候補の抽出

文字ラティス中から，同一字種の連接となる文字列全てを単語候補として抽出する．文字切り出しや文字認識の誤りに対処するため，他の文字列の一部となっている場合でも，異なる字種の文字に挟まれている文字列は全て単語候補とする．また，生成された単語候補 $c_1c_2\ldots c_n$ の確信度を $\frac{1}{n}\sum P_{c_i}+\alpha n$ と定める．ただし，P_{c_i} は文字 c_i の確信度である．この定義によって，文字の確信度が高く，長さが長いものほど単語として抽出されやすくなる．

Step2 出現頻度による単語確信度の補正

類似文字の存在などを考えれば，文字の確信度のみに頼って抽出すべき単語を決定することは得策ではない．そこで，同一文書内に何度も出てくる単語候補は重要な単語である可能性が高いという点に着目して，出現頻度の高い単語候補の単語確信度が高くなるように補正する．具体的には，文字列としての表記が同一の単語候補に対して，それらの単語確信度の総和を新たな単語確信度とする．この補正によって，出現頻度の高い単語候補から順に抽出されることになる．

Step3 単語確信度の高い単語から確定

最後に，単語確信度の高い単語候補から順に単語として抽出する．その際に，既に抽出された単語と位置的に重なるような単語候補は抽出しない．また，単語確信度が一定値β以下のものも抽出しない．

後の実験で示すように，Step1-3 の未知単語抽出法のみで単語抽出を行うよりも，辞書単語抽出法を併用した方がよい結果が得られる．辞書単語照合法では，最初に，文字ラティスから辞書単語の候補を抽出し，Step1 で定義した単語確信度に従って Step3 の確定処理を行うことにより，確からしい単語だけを辞書単語として抽出する．次に，抽出された単語を構成するアーク，および，それらと位置的に重なるアーク全てを文字ラティスから削除する．残った文字ラティスに対して Step1-3 の未知単語抽出法を適用すれば，辞書中に存在しない未知単語が抽出される．複合名詞の分離が可能となる点も辞書を用いるメリットの一つである．

	Step1-3 のみ	辞書のみ	辞書+Step1-3
再現率	98.1%	90.1%	100%
適合率	98.1%	100%	99.2%

表 1: 文字ラティスからの単語抽出結果

提案した未知単語抽出法の有効性を示すために，2 節で述べたサンプル文書画像の文字ラティスに対して，Step1-3 のみを適用した場合，辞書単語のみを抽出した場合，辞書単語を抽出した後で Step1-3 を適用した場合について，それぞれ再現率と適合率を求めた (表 1)．ただし，$\alpha = 0.02$，$\beta = 1.0$ とし，漢字単語は 2 文字以上のもの，カタカナ単語と英数字単語は 3 文字以上のものだけを抽出した．再現率が高いことは，抽出したい単語を漏らしていないことを意味し，適合率が高いことは，無駄なゴミを抽出していないことを意味する．表 1から，Step1-3 の手法を未知単語の抽出に使用すると，わずかなゴミが抽出されるが，抽出すべき単語は全て抽出できていることが分かる．

5 おわりに

本稿では，文字ラティスから単語を抽出する手法について述べた．辞書単語を抽出したのち，字種による単語候補の生成，出現頻度に応じた単語確信度の補正によって，未知単語が抽出できることを示した．

本手法によって，文字候補数が比較的少ない文字ラティスに対しては非常によい結果を得ることができるが，文字候補数の非常に多い文字ラティスを対象とした場合，極端に結果が悪くなる恐れがある．これは，抽出すべき文字列に関する制約が弱いためと考えられるので，こうした場合には，形態素解析などのより一般的な自然言語処理が必要となるだろう．今後は，文字ラティスに対する形態素解析について考察したい．

参考文献

[1] 湯浅, 上田, 外川: “大量の文書データから自動抽出した名詞間共起関係による文書の自動分類,” 情処技法, NL98-11, 1993.

[2] 鈴木, 増山, 内藤: “語の意味分類の出現傾向を考慮したキーワード抽出の試み,” 情処技法, NL98-10, 1993.

[3] 津田, 仙田, 美濃, 池田: “自動作成された単語間リンクによる検索質問作成支援,” 第 48 回情処全大, 1993.

[4] Senda, S., Minoh, M. and Ikeda, K., “Document Image Retrieval System Using Charcater Candidates Generated by Charcater Recognition Process,” *Proc. of 2nd ICDAR*, pp. 541–546, 1993.