修士学位論文

論文題目　言語情報を活用する
知的文字認識に関する研究

提出者　東北大学大学院工学研究科
電気及通信工学　専攻
学籍番号　5m123
氏名　秋山秀三

指導教官	阿曽 弘具 教授
審査委員 （○印は主査）	○ 阿曽弘具 教授 1 丸岡 章 教授　2 佐藤雅彦 教授 3 ＿＿＿＿ 教授　4 ＿＿＿＿ 教授

提出者略歴

氏名（ふりがな）	あきやま ひでみ 秋山 秀三　昭和 45 年 1 月 6 日生
本籍	茨城 都道府県 水戸市泉町 1丁目102
履歴事項	
昭和 平成 元年 4月 1日	東北 大学 工 学部 学科入学
昭和 平成 5年 3月 25日	同 卒業 情報工学科
昭和 平成 5年 4月 1日	東北大学大学院工学研究科 電気及通信工学 専攻 前期2年の課程入学
昭和 平成 7年 3月 24日	同 修了
昭和 平成 年 月 日	
昭和 平成 年 月 日	
昭和 平成 年 月 日	
昭和 平成 年 月 日	

備考(1) 履歴事項は、大学入学から年次にしたがって記入すること。

(2) 修士課程の修了年月日は、学位記授与式年月日を記入すること。

修士学位論文

言語情報を活用する
知的文字認識に関する研究

東北大学大学院工学研究科
電気及通信工学専攻

秋山秀三

目 次

図目次

表目次

第 1 章

序論

1.1 研究の背景および目的

高度情報化社会の発展において、我々を取り巻く文字・音声・画像など様々な形態の情報は膨大な量になってきている。これらの情報を効率よく利用するために、雑誌・書籍・書類などの紙の上に文字で表された情報を計算機に取り込み、それらを計算機上で管理しようという社会的要請が高まっている。増大し続ける膨大な量の文字情報を計算機へ入力するためには、その入力手段の省力化が必要である。その要請に応えるものとして、文字情報を計算機に可読な形へ自動的に変換する**文字認識技術**があげられる。文字認識は文書の計算機への入力の完全自動化を目指すものであり、その高速化・高精度化への研究・開発が進められ、いくつかの OCR (Optical Character Reader : 光学式文字読み取り装置) が実用化される段階にきている[1][2]。さらに、文書構造解析技術と結びついた文字・図表・写真が混在した文書の読み取り[3] や、オンライン手書き文字認識[4]を利用したペンコンピュータなど様々な応用も期待されている。

しかしながら、日本語で書かれた文書を文字認識の対象とした場合、以下の点が問題となる。

1. 字種数が多い。JIS 第一水準に登録されている漢字、平仮名、片仮名、数字だけでも 3,196 種類存在する。実際の文書ではこれらと数種の記号が混在している。

2. 類似文字が多い。「大」-「犬」「ほ」-「ぼ」といったものから、「夕」(漢字)-「タ」(片仮名)、「へ」(平仮名)-「へ」(片仮名) など、文字パターンからだけでは識別が難しいものもある。

3. 1 文字が複数個の文字パターンから構成される分離文字 (「読」→「言」＋「売」など) が存在する。また、全角文字と半角文字が混合していたり、アルファベットや記

号など文字ピッチが一定でない文字が存在する。文字ピッチがすべて一定である文書であれば、ピッチ情報から 1 文字の文字領域を判断することができる。しかし文字ピッチが不定な文書は、ピッチ情報からでは分離したものが正しいのか、それらを統合したものが正しいのか判断しにくい。

一般に計算機で文字を認識する場合には、文書中から一文字ごとの文字パターンを切り出し、切り出した個々の文字に対し認識を行うという方法が主である。しかし、主に 1,2 の問題から高精度な認識が、3 の問題から正確な文字切り出しがそれぞれ困難なものとなっている。実際の文書では、かすれやにじみなどの低品質文字が存在し、さらに正確な認識が難しくなる。

人間が文書を読む場合には多少変形した文字や見えない文字があっても、どのような文字が書かれているのかを判断し、文章として意味をとらえることができる。これは前後の文字や文書に関する知識などを使って文字を補完しながら読んでいるためである。文章は単なる文字や記号の羅列ではない。文字が集まり単語を成し、単語が集まり文節を成し、文節の組合せにより意味を持った文章となる。そこには人が見て意味がわかるような文章となるための、文字の並びにおける規則性が存在していると考えられる。そこで計算機による文字認識においても、個々の文字の認識結果に対して日本語の文字列に関する知識 — **言語情報** — を参照することにより、個別文字認識での誤りを自動的に修正し認識の精度を向上させようという**文字認識後処理**の研究が行われている。[7] 文字認識後処理は、文字認識を用いるアプリケーションによって処理の方法が異なってくる。帳票の読み取りやペンコンピュータによる住所・氏名の読み取りにおいては記入するフォーマットや記入する項目があらかじめ定まっているため、単語照合など単語レベルの処理で効果をあげることができる。また、読み取る項目に応じた用語辞書を用いたり、住所の場合は階層関係、氏名の場合は姓名とふりがなの照合など、限定された知識を活用することにより認識精度を高めるといった方法[8] も用いられている。

一方、近年ではフォーマットの定まっていない一般の文書を計算機に読み取らせたいという需要が高まっている。この場合、考慮しなければならない点として次のようなことが考えられる。

- 英語などの場合には、文章は単語単位に区切られているため、単語レベルの処理により後処理を行うことができる。しかし日本語の場合は単語単位で区切られていないため、文字列を単語ごとに区切る処理が必要になる。

- 文字列から単語を抽出した後は単語間の接続を考えるなどの処理を行い、最終的に文としてとらえることになる。構文解析を用いて、より詳細な後処理を行う方法も考えられている[9] が、現在は対象とする文の文法を絞って研究している段階である。与える規則を絞ることによる後処理効果の向上は期待できるが、対象文書を制限するこ

とは一般文書の認識においては実用的ではない。なるべく広い範囲の日本語文を処理できるシステムであることが望まれる。

- 認識する文書の内容に制限がないため、ある程度単語辞書の内容を増やしても文書中に辞書には登録されていない単語が現れてくる。このように辞書に登録されていない単語を未登録語と呼ぶ。未登録語の存在により単語抽出に失敗し、さらにその後の処理に影響が出てくる。

- 記入するフォーマットの定まっていない文書の認識では、1 文字の文字領域を決定する文字切り出しの処理が誤ることがあり、この場合には正しい文章と文字数が異るため、切り出された認識結果を見ただけではこのような誤りを後処理で修正することは困難である。分離した文字画像の前後の認識結果から、言語情報を文字切り出し部にフィードバックして切り出しの精度を向上させる処理は効果的であると考えられ、認識精度向上が期待される。

本研究では、活字で印刷された一般の文書の文字認識に、日本語に関する言語情報を活用することにより文字認識の精度を高めることを目的とする。

まず、文字認識結果である候補文字・認識評価値に言語情報を用いて、日本語として確からしい文字列を選択することにより個別文字認識の誤りを修正する文字認識後処理システムを構築する。(図 1.1) これは日本語の言語情報として文節を単位とした処理を行い、簡単な言語的制約および各候補文字につけられる認識評価値をもとに最も確からしい文節列を選択するものである。

次に一般文書の認識に対応するものとして、文字切り出しの段階から言語情報を活用することにより、切り出し、文字認識、言語処理を統合した高精度な認識システムを構築する。

1.2 本論文の構成

第 1章 序論であり、研究の背景及び目的を述べる。

第 2章 文字認識後処理のシステムで使用する個々の文字の認識手法および認識評価値の算出方法について述べる。

第 3章 言語情報を用いた文字認識後処理システムを構築し、その構成・処理方法を述べ、実験による性能評価を示す。

第 4章 第 3章で示した手法を文字の切り出しの段階から適用し、一般文書に対応した高精度文字認識システムを構築する。また、システム性能評価のための実験を行う。

第 5章 本研究の成果と今後の課題について述べる。

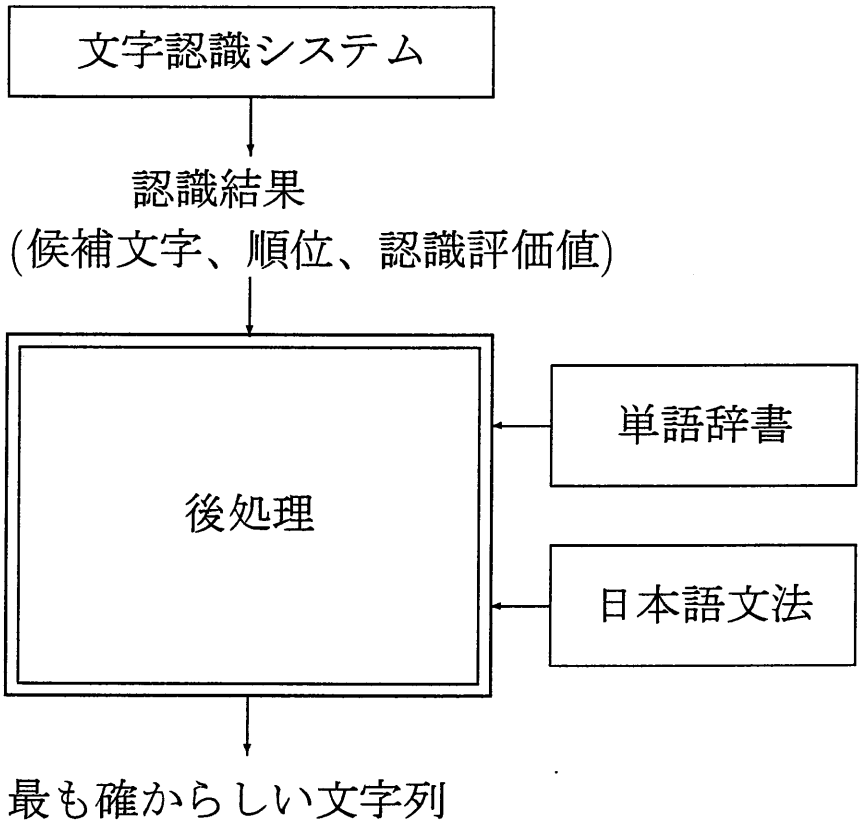


図 1.1: 後処理システムの流れ

第 2 章

個別文字認識

2.1 まえがき

第3章、第4章で構築する文字認識システムでは、切り出された個々の文字イメージの認識を行う過程があり、また認識時に出力される認識評価値をその後の過程で使用する。そこで本章では、個別の文字認識手法について述べる。

2.2 文字認識アルゴリズム

文字認識の方法には大きく分けてパターン整合法、構造解析法の２つがある。パターン整合法はパターン同士の重なり具合から類似度を評価し、それをもとに認識を行うものである。類似文字の識別は難しいが文字の多少の変形やノイズに強く計算機上での実現が容易である。通常は計算量削減のため文字画像そのものではなく文字画像から得られる特徴量を用いて認識を行う。これに対し構造解析法は線分の接続関係や位置関係などの文字構造に着目し、構造の類似性によって認識を行う。これは類似文字の多い漢字の認識や変形の大きい手書き文字の認識などには有効である。しかし、計算機上で実現させる場合には線分の正確な抽出や線分関係の定義が難しく、処理量が膨大であるなどの問題がある。本研究では活字文字を扱うため、その高速性に着目してパターン整合法に基礎をおく方向線素特徴量[10]による認識処理を用いる。

2.2.1 文字入力

文字入力は認識したい文書を二値画像として取り込む処理であり、認識の対象となる文字列のイメージはイメージスキャナを用いて計算機に取り込まれる。入力された文字列のイメージは、文字切り出しの処理により各文字ごとに分割される。

2.2.2 前処理

切り出された文字イメージから特徴量を求めるのに先立ち、次のような前処理が行われる。前処理は、ノイズ除去・スムージング、正規化、細線化、線素化の 4 つの処理によって構成されている。

2.2.2.1 ノイズ除去・スムージング

文字イメージは印刷時の状態やイメージスキャナの性質などによって、ノイズや線分上の凹凸が少なからず発生している。これらは認識精度に悪い影響を及ぼすため、ノイズ除去を行う。ここでは 2×2 ドット以下の孤立点をノイズと見なし除去することにする。スムージング (線分の平滑化) は、3×3 ドットのマスクを用いて周囲 8 近傍を見ることにより凸となっている部分を除去し、凹となっている部分を埋めるという作業をする。ノイズ除去・スムージング後の文字イメージの例を図 2.1(b) に示す。

2.2.2.2 正規化

活字文字は一般的に 4 倍角・縦倍角・全角などさまざまな大きさである。正規化は、切り出した文字の大きさや位置の違いによる影響を吸収するために、もとの文字イメージを一定の大きさに拡大または縮小する処理である。入力図形を縦・横両方向に一定の大きさに線形伸縮することで実現できる (線形正規化) が、このほか入力文字の線密度の大きい部分を拡大、線密度の小さい部分を縮小し、全体を一定の大きさにすることで、線密度を全体に分散させようとする (非線形正規化) もある。本研究では線形正規化を用いた。線形正規化後の文字イメージを図 2.1(c) に示す。

2.2.2.3 細線化

文字線幅の違いを吸収するために、正規化後の画像に対して細線化を行う。細線化は太さを持った図形を幅 1 の図形に変換する処理である。本研究では Hilditch の方法[11] を用いる。この方法は、各画素においてその周りの 3×3 の 8 近傍に着目し、8 連結数をもとに図形を 1 ドットずつ削っていく方法である。図形の変化がなくなるまでこれを繰り返す。すなわちパターンの線分の太さの半分の回数を繰り返すことにより、最終的に線幅 1 の連結図形が得られる。細線化後のイメージを図 2.1(d) に示す。細線化において、つぶれた文字の場合には完全に細線化してしまうと元の形が失なわれてしまう。そこで実際は削る回数の上限を制限し、その後イメージの輪郭線を抽出するものとする。削る回数の上限については文献[10] で検討がなされており、5 回が最適であるとの報告があるため、本研究でも細線化回数は 5 回とした。

2.2.2.4 線素化

線素化は、細線化された図形の各画素について、黒画素であればその周囲 3 × 3 の小領域を参照し最も自然な方向と考えられる 4 方向 (縦「｜」、横「—」、斜め 45°「／」、斜め 135°「＼」) のうちの 1 つの線素に対応させる処理である。線素化後のイメージを図 2.1(e) に示す。

2.2.3 特徴抽出

パターン整合法では、文字パターンの持つ冗長度の削減と処理の高速化、パターン分離の効率化などのために文字パターンを特徴量という数値ベクトルに変換する。この過程を特徴抽出という。

本研究で用いる方向線素特徴量の抽出法は、まず線素化された画像を 8 × 8 の小領域に分割し、図 2.4に示す重なりのある 16 ドット × 16 ドットの領域ごと (計 49 領域) に、その領域に含まれる各線素を重み付きで計数する。これにより得られる 49 × 4 = 196 次元のベクトルを特徴量とする。ここで用いる重みは、図 2.2に示すような、中心部ほど大きい値を持つガウスフィルタ的なものである。図中の数値が重みを表し、ベクトルの値は、

(重みの値 × その重みの領域に含まれる線素数) の和

となる。単なる線素数を特徴量にする場合は、図 2.2に示す重みをすべて 1 にしたものとみなせる。図 2.3に「千」をサンプルにしたときの第 4 領域の方向線素特徴量の抽出例を示す。

候補選出は、あらかじめ作成しておいた認識対象である各々の文字の代表ベクトル (辞書ベクトル、標準パターン) とこのベクトルとの間の認識評価値 (距離) を求め、その値の小さい順に候補文字のコードと認識評価値を出力することで行う。認識評価値としては、ユークリッド 2 乗距離を用いる。

未知入力パターンから求めた特徴ベクトルを

$$\boldsymbol{v} = (v_1, v_2, ..., v_N)$$

とすると、これらの間のユークリッド 2 乗距離 Ee_kは

$$Ee_k = (\boldsymbol{v} - \boldsymbol{m_k})(\boldsymbol{v} - \boldsymbol{m_k})^t = \sum_{i=1}^{N}(v_i - m_{ki})^2$$

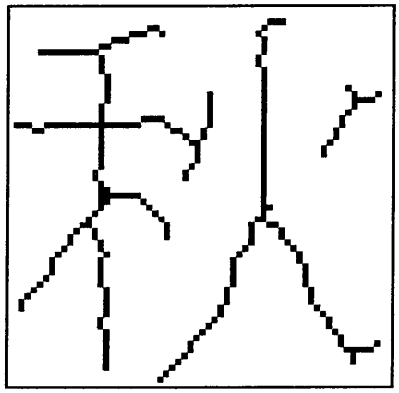
となる。

Ee_k、を最小にする字種 kが、未知入力パターンの 1 位候補文字となり、以下 Ee_k の小さい順に第 2 位候補、第 3 位候補、...、第 N位候補となる。(図 2.5)

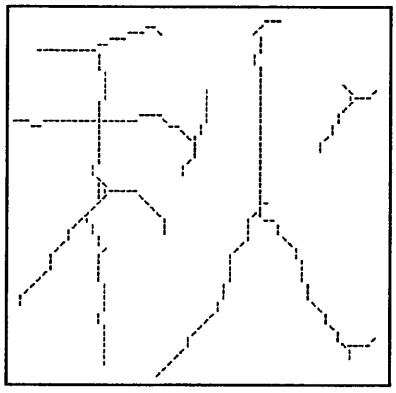
(a) 入力画像

(b) ノイズ除去・スムージング後

(c) 正規化後

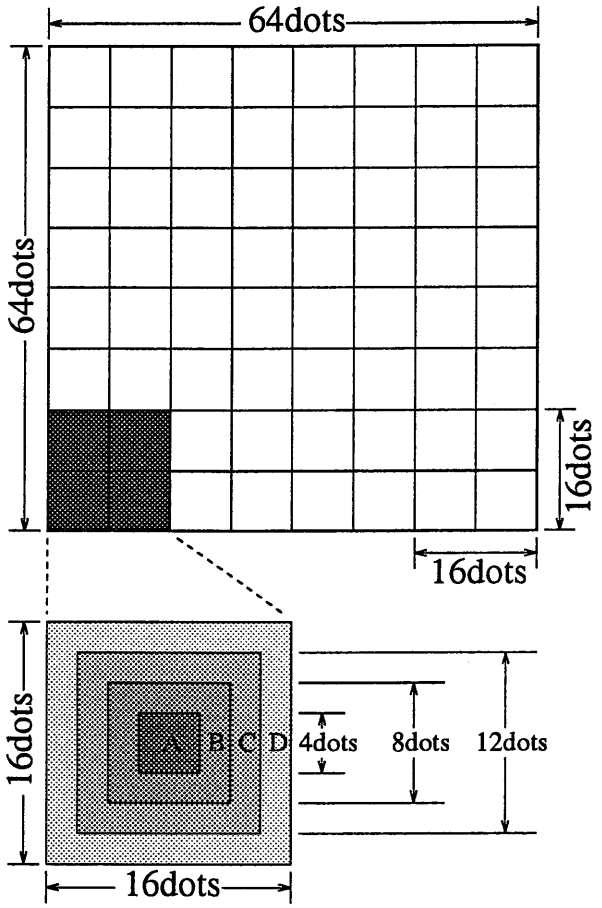


(d) 細線化後



(e) 線素化後

図 2.1: 各処理後のイメージ



方向線素：“|”, “—”, “／”, “＼”

1 領域：16 × 16 ドット

領域数：7 × 7 = 49 個

次元数：4 × 49 = 196 次元

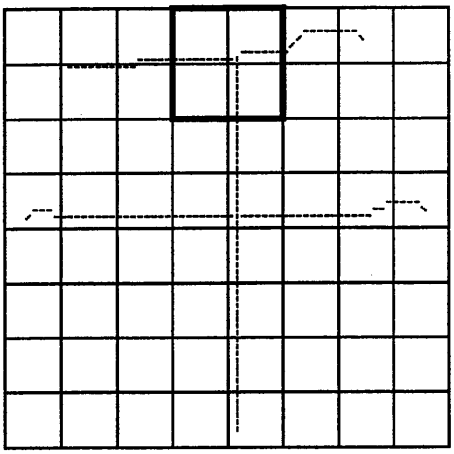
重みのかけ方：　枠 A → 4

枠 B → 3

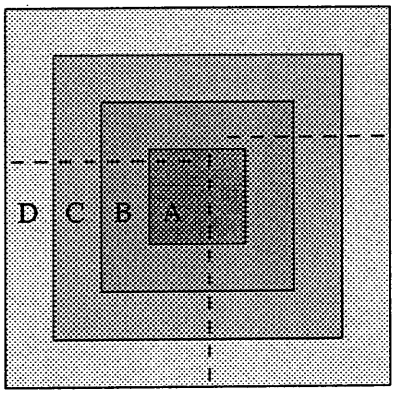
枠 C → 2

枠 D → 1

図 2.2: 方向線素特徴量



「千」



$$
\begin{aligned}
\boldsymbol{F}(P^{千}) &= \left(\boldsymbol{F}(P^{千})_{第\,1\,領域},\ \boldsymbol{F}(P^{千})_{第\,2\,領域},\ldots,\ \boldsymbol{F}(P^{千})_{第\,49\,領域} \right) \\
&= \left(F_{1|},F_{1-},F_{1/},F_{1\backslash},F_{2|},F_{2-},F_{2/},F_{2\backslash},\ldots,F_{49|},F_{49-},F_{49/},F_{49\backslash}\right) \\
&= \left(F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, \ldots, F_{193}, F_{194}, F_{195}, F_{196}\right)
\end{aligned}
$$

第 4 領域の抽出例

$$
\begin{aligned}
F_{13} = F_{4|} &= \overbrace{4 \times \mathbf{4}}^{A} + \overbrace{2 \times \mathbf{3}}^{B} + \overbrace{2 \times \mathbf{2}}^{C} + \overbrace{2 \times \mathbf{1}}^{D} = 28 \\
F_{14} = F_{4-} &= 2 \times \mathbf{4} + 5 \times \mathbf{3} + 4 \times \mathbf{2} + 4 \times \mathbf{1} = 35 \\
F_{15} = F_{4/} &= 0 \times \mathbf{4} + 0 \times \mathbf{3} + 0 \times \mathbf{2} + 0 \times \mathbf{1} = 0 \\
F_{16} = F_{4\backslash} &= 0 \times \mathbf{4} + 0 \times \mathbf{3} + 0 \times \mathbf{2} + 0 \times \mathbf{1} = 0
\end{aligned}
$$

$$
\underline{\boldsymbol{F}_{第\,4\,領域} = (28, 35, 0, 0)}
$$

図 2.3: 方向線素特徴量の抽出例

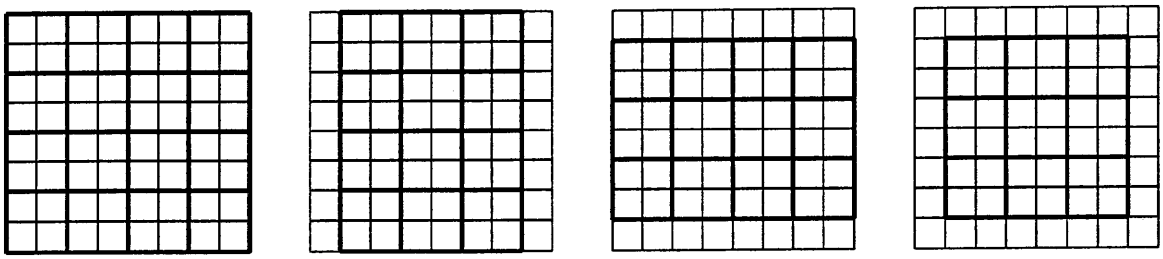
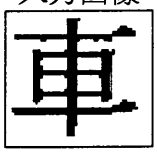


図 2.4: 49 の領域

入力画像



↓

	候補文字	認識評価値
1 位候補	車	7379
2 位候補	革	14562
3 位候補	草	15751
4 位候補	軍	17202
5 位候補	牽	17306
6 位候補	奉	17687
7 位候補	卓	18403
8 位候補	華	19463
9 位候補	東	19778
10 位候補	章	19840
	:	

図 2.5: 個別文字認識結果の出力

2.2.4 認識用辞書

標準パターンの集合を辞書という。本研究で使用する認識用辞書は79種の学習サンプルから求めた特徴量の次元ごとの平均としている。学習サンプルは明朝体・ゴシック体が中心であるが、一部新聞明朝体や教科書体も入っている。そして、記号・数字・アルファベット・仮名と特徴量レベルでの分散が大きい漢字についてはクラスタリングを行って複数の標準パターンを用意し延べ5,451 カテゴリーの辞書となっている。辞書の構成を2.1に示す。

表 2.1: 辞書の構成

種類	字種数	標準パターン数
全角記号	96	3
全角数字・アルファベット	54	3
全角平仮名・片仮名	144	3
JIS 第一水準の漢字	2,965	1〜2
JIS 第二水準の漢字 (一部)	451	1
特殊文字 (「†」)	1	1
2 桁の数字 (「00」〜「99」)	100	1
半角数字	95	1
計	3,906	5,451(延べ)

2.2.5 まとめ

本章では、第3章、第4章で使用する1文字ごとに切り出された文字イメージを認識する手法について述べた。本研究で用いた認識手法はパターン整合法に基礎をおく方向線素特徴量によるものである。

第 3 章

文字認識後処理システム

3.1 まえがき

前章では文字列のイメージから切り出された個々の文字の認識方法、認識評価値について述べた。本章では、この認識系より出力される候補文字・認識評価値を用いて文字認識後処理を行うシステムの内容について述べる。

本章の構成は、まずシステムに使用するための言語情報となる単語辞書・単語接続規則の内容・構造について説明し、その後システムの構成、処理方法を述べ、実際に文章の認識結果を使った後処理実験により評価を行う。

3.2 使用する日本語知識

3.2.1 自立語辞書

本研究で使用する自立語辞書は「九州芸工大自立語辞書ＫＩＤ－Ｊ９４」および検索用辞書 edict の一部 (人名・カタカナ語) から必要な情報を取り出し再構成したものである。一単語についてのフォーマットを図 3.1 に示す。

単語は漢字コード (JIS) 順に並べられており、検索時には単語の先頭文字をキーとしたハッシュテーブルを使用する。ハッシュテーブルには、ある文字が先頭文字として最初に現れる単語の番地とその文字が先頭文字として使われる単語数が登録されている。

また、漢字で登録されている単語が平仮名表記される場合に対応するため、各単語をかな表記した辞書を用意した。これを自立語仮名辞書と呼ぶ。自立語仮名辞書から検索する場合のハッシュテーブルは先頭の２文字をキーとする。

登録されている単語の総数は自立語辞書が約 148,000 語、自立語仮名辞書が約 100,000

語である。

	1	2	3	4	5	6
単語の表記	サ変動詞	活用型	活用行	形容詞	形容動詞	その他の品詞

図 3.1: 自立語辞書のフォーマット

```
応用 1     1
応用式       1
応用的      1
応用力       1
応量器       1
応力       1
押  13
押え  3A  1
押え技       1
押え込  17
押え込み 1     1
押え取  19
```

図 3.2: 自立語辞書の例 (一部抜粋)

自立語辞書に格納されている各文法情報の分類を次に示す。以下の分類は「九州芸工大自立語辞書ＫＩＤ－Ｊ９４」の分類法に準ずるものである。

１．サ変動詞表示部

サ変動詞を、口語と文語*、また、清音と濁音で区別して表わした。

[1]　清音の口語・文語サ変動詞　[愛 (する)]

*本研究では、文語文法については考慮していない。

[2] 濁音の口語・文語サ変動詞 [信(ずる)]
[3] 清音の文語サ変動詞 [解(す)]

2．活用の型表示部

サ変以外の動詞を次のように分類して表した。

[1] 五段(四段)活用
[2] 上一段活用
[3] 下一段活用
[9] カ行変格活用

3．活用行表示部

活用の型に対して、その活用行を次のように表示した。

[2] カ行
[3] サ行
[4] タ行
[5] ナ行
[6] ハ行
[7] マ行
[8] ヤ行
[9] ラ行
[A] ワ(ア)行
[B] ガ行
[C] ザ行
[D] ダ行
[E] バ行

4．形容詞の型表示部

形容詞を次のように活用の型で分類して表した。

[1] 口語形容詞
[2] 文語形容詞ク活用
[3] 文語形容詞シク活用
[4] 口語形容詞と文語形容詞ク活用を兼ねたもの
[B] 口語形容詞と文語形容詞シク活用を兼ねたもの

5．形容動詞の型表示部

形容動詞を次のように活用の型で分類して表した。

[1] 口語形容動詞
[2] 口語形容動詞のうち、「～の」に形のあるもの
[3] 文語形容動詞タリ活用

[4] 文語形容動詞ナリ活用
[B] 口語形容動詞と文語形容動詞ナリ活用を兼ねたもの

6．その他の品詞の表示部

上記以外の品詞 (活用のないもの) を次のように分類した。

[1]	名詞	[5]	感動詞
[2]	連体詞	[6]	助数詞
[3]	副詞	[7]	固有名詞
[4]	接続詞	[8]	記号

3.2.2 付属語辞書

付属語辞書には形式名詞、助動詞、助詞、補助用言、活用語尾の順に、５６０語の付属語が、それぞれの活用型、品詞、活用形がコード化され格納されている。一単語についてのフォーマットを図 3.3に示す。図 3.3の品詞１、品詞２はそれぞれ前に接続する単語による品詞分類、後に接続する単語による品詞分類であり、本研究で構築した後処理システムでは品詞２のみ使用している。これらは後に文節生成に使用するため通常の品詞分類より細分化されておりそれぞれ 130 種に分類されている。付属語辞書の内容は付録 A に記載した。

	1	2	3	4
単語の表記	活用型	品詞 1	品詞 2	活用形

図 3.3: 付属語のフォーマット

3.2.3 単語接続規則

１つの文節は１つの自立語の後に複数の付属語が接続することで構成され、接続可能な付属語は接続される単語の品詞により限定される[5][6]。これは自立語–付属語間の接続、付属語–付属語間の接続のどちらについてもいえることである。(図 3.4ここでは各品詞に接続可能な付属語を 3.2.2 で述べた付属語辞書のアドレスにより指定した。

○：あなた (名詞) → を (格助詞)
×：走る (動詞終止形) → を (格助詞)
○：させ (使役助動詞連用形) → られ (受身助動詞)
×：られ (受身助動詞連用形) → させ (使役助動詞)

図 3.4: 単語間接続可能性

○：書–き (カ行五段活用連用形活用語尾)
×：書–い (カ行五段活用連用形イ音便活用語尾)
○：書–い–て (接続助詞)

図 3.5: 文節成立性

また、ある単語まで接続した場合、その単語列までで文節として成立するかどうかは最後に接続した単語によって決まる。この文節成立性についても各品詞ごとに記述してある。(図 3.5および付録 B 参照)

3.3 文字認識後処理システム

3.3.1 文字認識後処理システムの構成

今回構築した文字認識後処理システムは図 3.6に示すような構成になっている。個別文字認識の結果 (候補文字とその認識評価値) が入力となり、候補文字選択、照合範囲決定、単語照合・文節抽出、最尤文節列選択の各段階を経て、個別文字認識の誤りが修正された高精度な認識結果を出力とする。

以下は各段階について順を追って解説する。

3.3.2 候補文字選択

第 2章で述べた方法により各候補文字およびその順位、認識評価値が求まる。これらが後処理システムへの入力となるが、単語照合の過程では認識候補文字の組合せと単語辞書との照合という作業となるため、認識候補文字数の増加は辞書との照合回数を飛躍的に増加させ、処理速度の低下を招く。また、あまり下位候補まで単語照合の対象とするとその組合せによって誤った単語が生成される場合が生じるため、誤修正の原因となる。

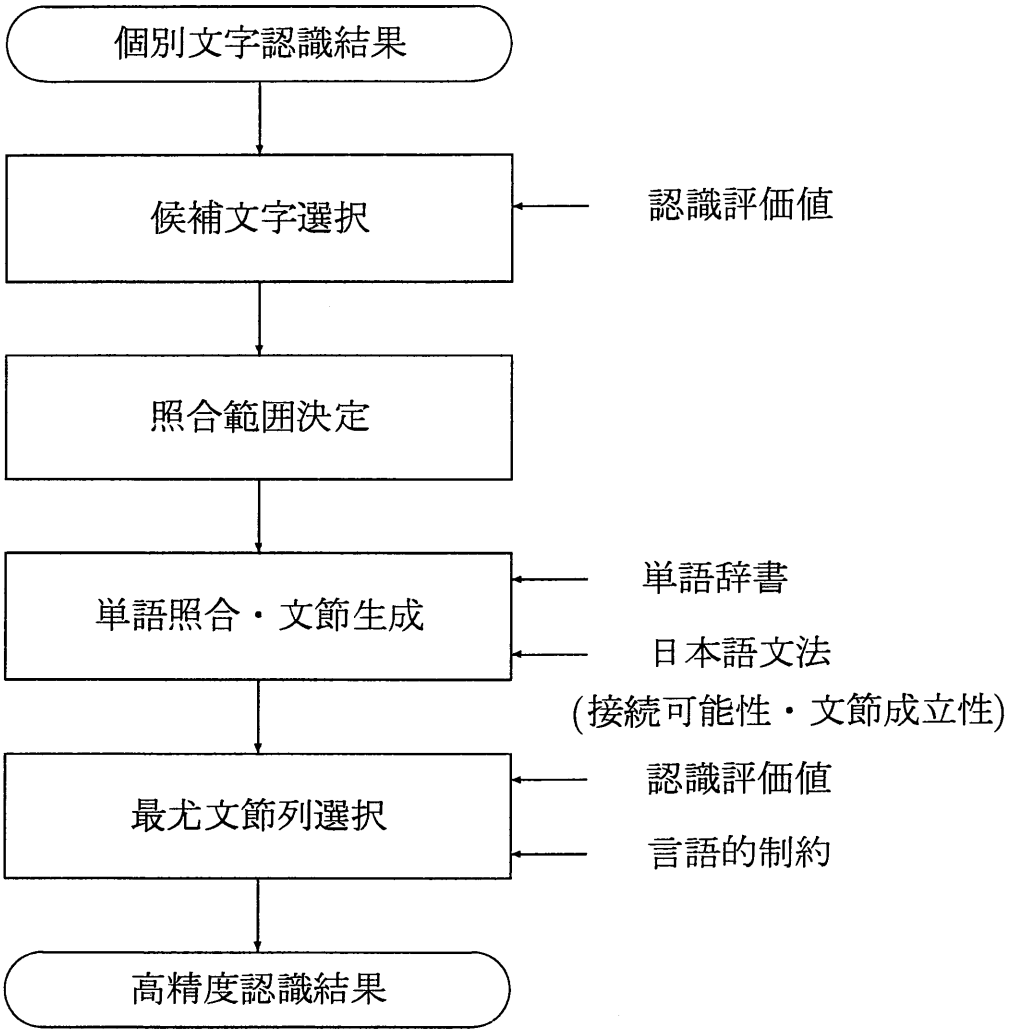


図 3.6: 後処理システムの構成

各候補文字の認識評価値によって1位候補が正解であるかどうかが判定できれば、正解ではないとされた文字についてのみ下位候補も含めて考慮すればよいことになり、単語として成立する組合せもそれだけ限定される。後処理の処理量軽減・高精度化のために、認識における認識評価値を利用して単語照合への入力となる文字数を絞り込むことが有益であると考えられる。(図 3.7)

本後処理システムでは、1位候補が正解であるかどうかの基準として個別文字認識の1位候補と2位候補との認識評価値の比を用いた。一般に認識評価値は1位候補が正解である場合には1位候補と2位候補との認識評価値の比は大きく、そうでない場合には認識評価値の比は小さい値に集中するという傾向がみられる。(図 3.8) この傾向は文字ごとに異なっている。そこで、あらかじめこの傾向を文字ごとに調べておき、1位候補を正解とみなすかどうかの認識評価値の比の閾値を設定しておく。手順としては、数種のサンプルデータを個別文字認識し、その1位候補文字ごとに認識結果を分類して、1位候補と2位候補との認識評価値の比の分布をとる。閾値は1位候補が正解文字であるときの認識評価値の比の分布において、認識評価値の比の大きい値からサンプルデータの N% が含まれている値として決定する。(図 3.9) N が大きいほど、候補文字数に対する絞り込みによって除かれる文字数の割合は増大するが、同時に正解文字が誤って絞り込まれてしまう可能性も高くなる。(つまり正解文字が下位候補に挙がっているにもかかわらず誤っている1位候補を正解とみなしてしまう)

また、1位候補が正解とみなされなかった場合に、下位候補を何位までとるかの基準を決めておく必要がある。これは1位候補文字ごとに分類したデータ中で、1位候補が正解でなかった場合に正解文字の認識評価値の絶対値がどれほどであったかをあらかじめ調べておき、正解の認識評価値がとりえた最大値を閾値とし、それ以上の認識評価値になる候補文字については単語照合の対象とはしないという方法をとった。

実際に実験を行う場合には 100 種類のフォントより認識評価値の比・認識評価値の絶対値の統計をとり、それぞれ閾値を求めた。認識評価値の比を求める際の N の値は 70% とした。以上の方法で単語照合の対象とする候補文字を選択する。

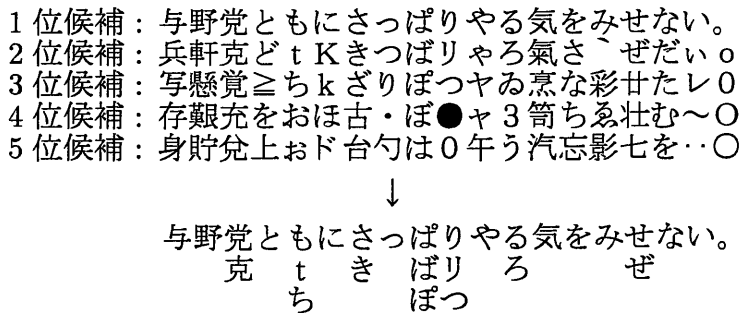


図 3.7: 候補文字の絞り込み

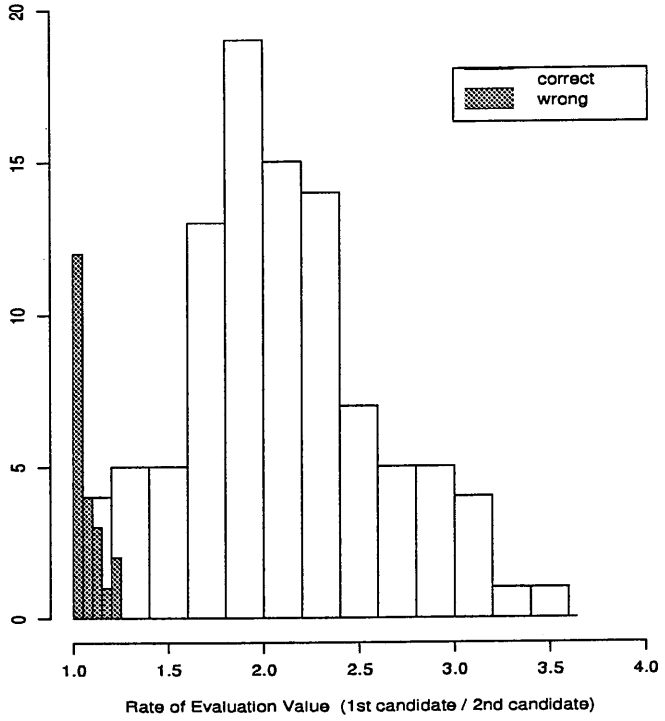


図 3.8: 1位、2位候補の認識評価値の比の分布 (1位候補が「諸」の場合)

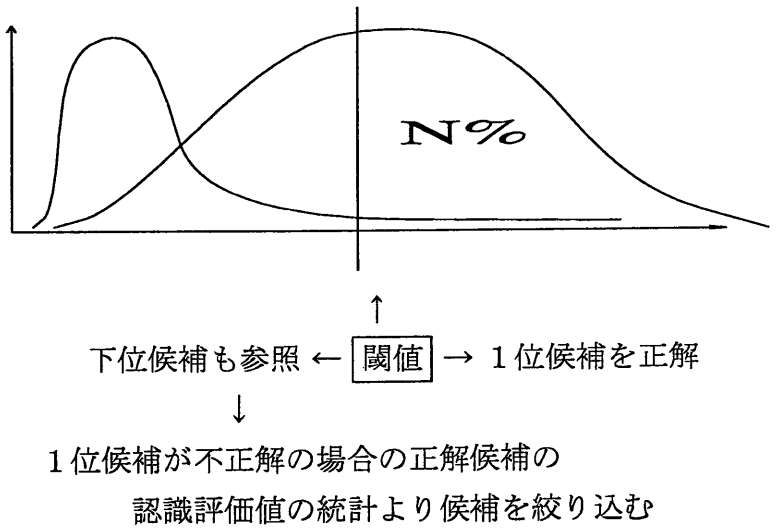


図 3.9: 1位候補を正解とする閾値の決定

3.3.3 照合範囲決定

単語照合では候補文字の組合せと単語辞書との照合を行う。照合を行う範囲は句の境界となる部分を選ぶのが良いと考えられる。ここでは句読点と認識された文字までを照合範囲とする。構文解析などで用いられる手法として、単語照合範囲の境界に字種の変化点 (非漢字 → 漢字) 選ぶことが行われるが、文字認識においては漢字とひらがなの間にも類似文字があり異字種間で誤認識することがある。また「お客さん」「か弱い」など字種にまたがった単語が存在するため字種の変化点を範囲とすることは適当ではない。

句読点を句読点以外の文字に誤認識する場合は、“下位候補に句読点が存在すれば終止形の次には句点がくる” といった知識を使用することにより修正できる。一方、句読点以外の文字を句読点に誤認識した場合は文節の切り出しに誤りが生じるが、このような例は本研究で使用した認識系では起こらなかった。

また文章が句読点を使わないものであったり、フォントの違いによって句読点が他の「０(ゼロ)」「Ｏ(オー)」などの類似文字に間違える、かすれのために「ｃ」に間違える、といったケースが多くなると１位候補が句読点であるものだけを照合範囲の区切りとしていたのでは、範囲が広くなりすぎて処理の効率が悪くなる。このため照合範囲とする文字数に上限を設けておき、その文字数以上になった場合はその時点で強制的に範囲として定める。この場合、照合範囲の境界が文節の区切りではないため、文末付近の文節の抽出が失敗することがある。このため、強制的に定めた照合範囲で一旦最尤文節列 (最も日本語の文章らしいと判断される文節列。3.3.7で述べる。) を決定し、その結果の文末から数えて２文節分戻った位置の文字を次の照合範囲の先頭とする。

3.3.4 単語照合

単語照合方式には以下の二つの方式がある[7]。

- 候補文字の組合せから作られる文字列を基本にして、この文字列が単語辞書に存在するかを検索する候補文字主導型
- 辞書に存在する先頭文字の同じ単語を基本にして、これらの単語の各文字が候補文字集合に含まれているかを調べる辞書主導型

前者の場合、各文字に対する候補文字数を m, 最長単語長を l とすると $\sum_{k=1}^{l} m^k$ の組合せの辞書照合が必要になる。後者の場合、同一の先頭文字をグループ化することにより平均個数が数十個程度の単語について候補文字集合との照合になるため、検索が高速に行える。本研究では後者を用いる。

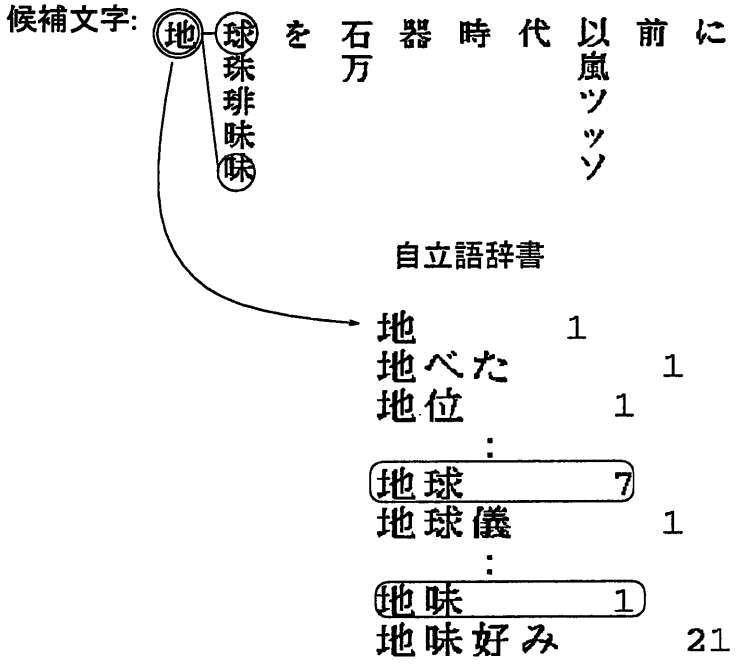


図 3.10: 辞書主導型単語照合

3.3.5 文節生成

日本語の文節は自立語の後ろに複数の付属語が接続して生成されると考えられる。3.3.4で抽出された自立語の後ろに接続可能な付属語が候補文字の組合せにあるかどうかを3.2.2、3.2.3で述べた付属語辞書と単語接続規則により照合する。

1つの自立語に対し複数の付属語が接続する、つまり付属語の後ろにさらに付属語が接続し得るため、接続が可能なかぎり付属語の接続操作を操り返す。その過程で生成される単語列についてその時点で文節として成立するかどうかを判定し、成立するとされた単語列のみ文節候補として扱う。

図3.11の例では「確認されている」という文字列を認識したときの候補文字について、候補文字の組合せから生成される文節を示している。「確」という文字から始まる自立語で候補文字の組合せの中に含まれているものは、サ変動詞の語幹としての「確認」と、名詞としての「確認」の2つである。(図3.11の点線より上の部分) これから付属語の接続を考慮して生成された単語列のなかで、その時点で文節として成立するものが図で四角で囲んだものである。これが文節候補となる。

候補文字の組合せから文節を抽出した例を図3.12に示す。右に書いてあるのは文節が持つ文法データであり、それぞれ文節の最初の単語の品詞分類、文節の最後の単語の品詞分

類、活用形、照合したのは自立語辞書か自立語仮名辞書かあるいは次の 3.3.6 で述べる未登録語なのかの別が示されている。

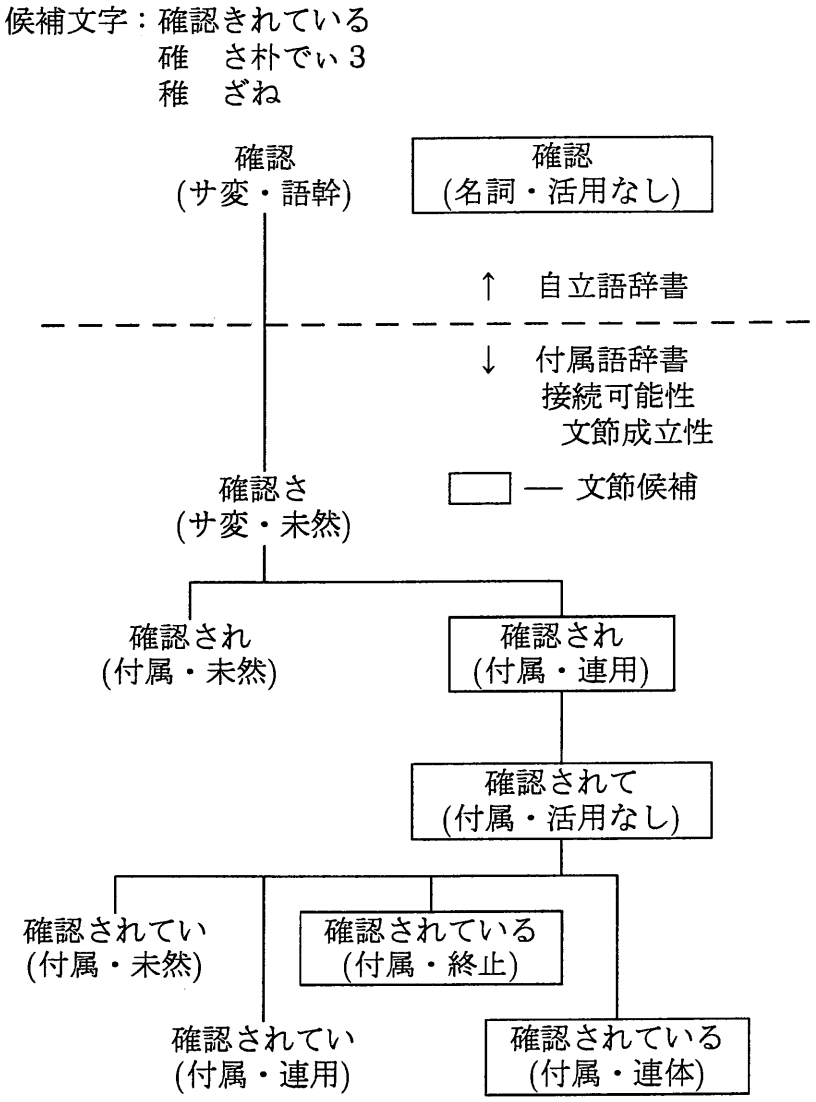


図 3.11: 文節の生成

候補文字

1 位候補：議会内で与党が犬勢を占めるようになったために、
2 位候補：譲　肉て写　か大　　ろょぅ　だつだだぬK
3 位候補：講　向ア兵　ボ太　　3工ラ　　りｔｔ灼ほ
4 位候補：護　的ァ存　潰丈　　うエ）　　・淀危鋤ｋ
5 位候補：諸　　↑身　｀ナ　　ぅェ云　　ぅ彪わ備ぼ

抽出された文節

議会	名詞-(名詞. 活無) 漢字
議会内	名詞-(名詞. 活無) 漢字
議会内で	名詞-(付属. 連用) 漢字
会	名詞-(名詞. 活無) 漢字
会内	名詞-(名詞. 活無) 漢字
会内で	名詞-(付属. 連用) 漢字
内	名詞-(名詞. 活無) 漢字
内で	名詞-(付属. 連用) 漢字
肉	名詞-(名詞. 活無) 漢字
肉で	名詞-(付属. 連用) 漢字
与党	名詞-(名詞. 活無) 漢字
与党が	名詞-(付属. 活無) 漢字
与党か	名詞-(付属. 活無) 漢字
党	名詞-(名詞. 活無) 漢字
党が	名詞-(付属. 活無) 漢字
党か	名詞-(付属. 活無) 漢字
犬	名詞-(名詞. 活無) 漢字
大勢	名詞-(名詞. 活無) 漢字
大勢を	名詞-(付属. 活無) 漢字
占め	下一-(下一. 連用) 漢字
占める	下一-(下一. 終止) 漢字
占めろ	下一-(下一. 命令) 漢字
占めるよ	下一-(付属. 活無) 漢字
占めるよう	下一-(付属. 活無) 漢字
占めるように	下一-(付属. 連用) 漢字
占めろよ	下一-(付属. 活無) 漢字
よう	ワ五-(ワ五. 終止) かな
ように	ワ五-(付属. 活無) かな
うに	名詞-(名詞. 活無) かな
うにだ	名詞-(付属. 終止) かな
うにだった	名詞-(付属. 終止) かな
うにだった	名詞-(付属. 連体) かな
うにだったため	名詞-(付属. 活無) かな
うにだったために	名詞-(付属. 活無) かな
になった	ワ五-(付属. 終止) かな
になったため	ワ五-(付属. 活無) かな
になったために	ワ五-(付属. 活無) かな
になったために、	ワ五-(句点. 活無) かな
にだ	名詞-(名詞. 活無) かな
なり	ラ五-(ラ五. 連用) 漢字
なりた	ラ五-(付属. 終止) 漢字
なった	ラ五-(付属. 終止) 漢字
なったため	ラ五-(付属. 活無) 漢字
なったために	ラ五-(付属. 活無) 漢字
なったために、	ラ五-(句点. 活無) 漢字
なつ	名詞-(名詞. 活無) かな
なつだ	名詞-(付属. 終止) かな
なりたたぬ	タ五-(付属. 終止) かな
つた	名詞-(名詞. 活無) かな
つただ	名詞-(付属. 終止) かな
つたわぬ	ワ五-(付属. 終止) かな
たたぬ	タ五-(付属. 終止) かな
たたぬに	タ五-(付属. 活無) かな
たた	副詞-(副詞. 活無) かな
たため	マ五-(マ五. 命令) かな
ただ	名詞-(名詞. 活無) かな
だだ	名詞-(名詞. 活無) かな
ため	名詞-(名詞. 活無) かな
ために、	名詞-(句点. 活無) かな
だめ	名詞-(名詞. 活無) かな
だために、	名詞-(句点. 活無) かな

図 3.12: 文節候補抽出例 (一部抜粋)

3.3.6 未登録語処理

単語辞書を利用した単語照合検査においては、辞書に登録されていない未登録語を候補として抽出することはできない。一般の文書を対象とする場合には、辞書への登録単語数をいかに増やしても未登録語の出現は避けがたく、何らかの対策が必要となる。通常、文章を単語単位に分割する形態素解析では、入力文が正しいものと仮定できるため単語照合に失敗した部分が未登録語の範囲であると考えられるが、文字認識後処理においては1位候補が正しいとは限らず、未登録語の範囲を決定するのが難しい。

ここでは未登録語が原因で後処理が失敗する例で多いものとして、専門用語や固有名詞に用いられる片仮名表記語、アルファベット表記語、そして数字列に関して未登録語処理を行う。

候補文字の組合せの中で片仮名・アルファベット・数字のそれぞれ同一文字種が複数文字連続であらわれる部分をすべて抽出し、他の自立語と同様に付属語の接続検定を行ない文節を生成する。このときの文字列の品詞は名詞として扱う。数字列に関しては、その後に「個」「種類」などの助数詞が接続することが多いため数字の後に接続可能な助数詞について辞書と照合を行う。このようにして生成された文節を他の文節候補と同様に扱う。

図 3.13の例では「ムバラク」「フセイン」の2語固有名詞が未登録語である。未登録語処理をしない場合、これらの単語は抽出できないだけでなく「パテ」といった誤った単語を抽出している。これに未登録語処理を行うと「ムバラク」「フセイン」ともに抽出できている。

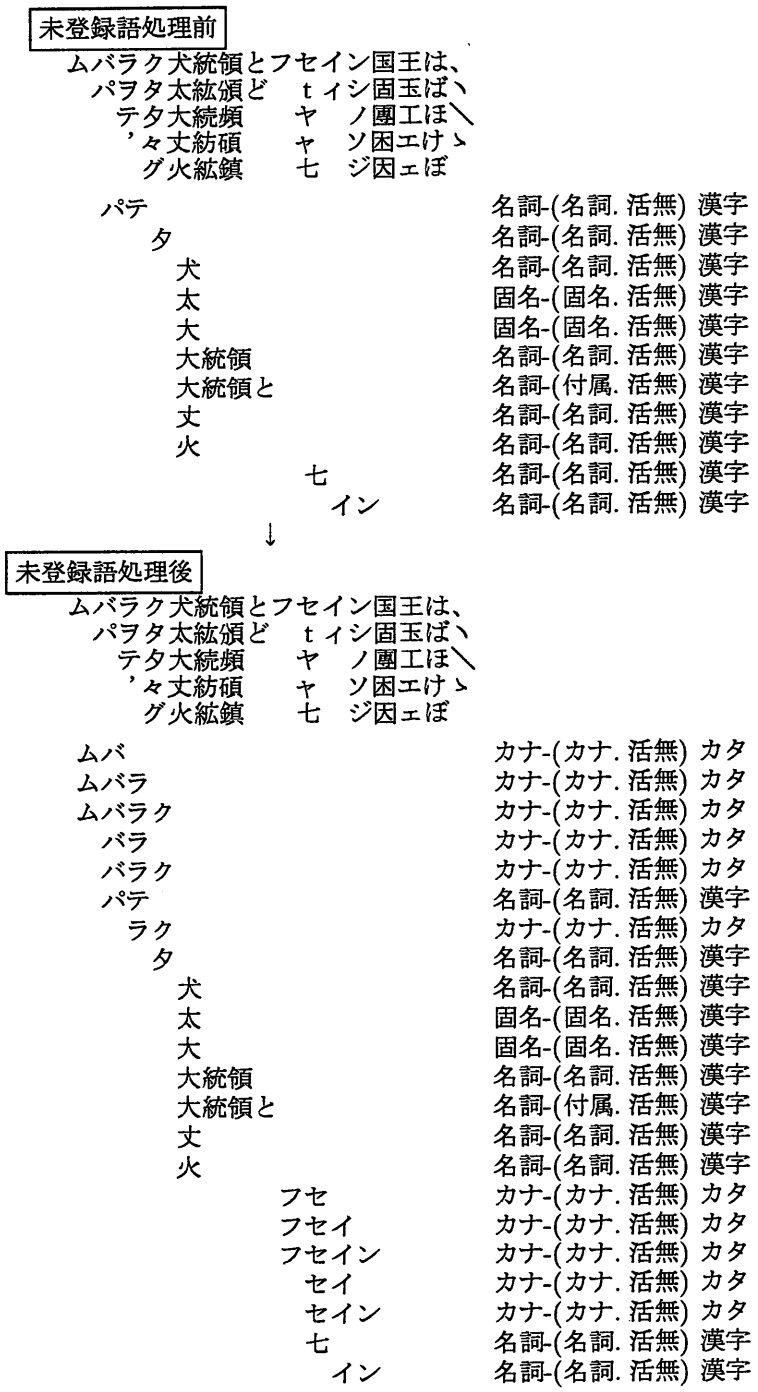


図 3.13: 未登録語処理による片仮名列抽出例

3.3.7 最尤文節列選択

候補文字の組合せから、未登録語も含め図 3.12のように文節候補が抽出される。最尤文節列選択とは、抽出された文節候補を組合せて文節列を生成し、生成した各々の文節列に対して、日本語の文章らしさを評価して、最も日本語の文章らしいと評価される文節列を選択する処理である。ここで、最も日本語の文章らしいと評価された文節列を最尤文節列と呼び、文節列に対して日本語の文章らしさを評価した値をコストと呼ぶ。コストは文節列によって定まる値で、日本語らしい文節列ほど小さい値を示すものである。

各文節列に対するコストは、簡単な言語的制約と個別文字認識より出力される認識評価値を用いて求める。主たる言語的制約として用いているのが以下に述べる形態素解析の一手法である文節数最小法[12]である。

文節数最小法

文節数最小法とは、日本語のワードプロセサの連文節変換などにも用いられている形態素解析の一手法であり、文字列を構成する文節の総数が最小になるように文字列を区分する方法である。文節数を少なくすることで、なるべく長い文節の並びを求めることに相当する。例えば、

にほんのれきしをまなぶ

という文字列があった場合、

／にほんの／れきしを／まなぶ／　　(／日本の／歴史を／学ぶ／)
／にほん／のれ／きしを／まなぶ／　(／日本／乗れ／岸を／学ぶ／)

などの複数の解釈が考えられる。上記の例では文節の総数が最小となる「日本の歴史を学ぶ」という解釈を優先する。この方法によると、ほとんどの場合に文節数が最小になる区切り方と 1 文節多い区切り方までの文節列の中に正しい解釈が存在することが知られている。[12]

この他に使用する言語的制約として用いるのは、3.3.6 で述べた未登録語として抽出された文節と文字数が 1 文字の文節に対してはコストを大きくする、というものである。これは未登録語も 1 文字語も単語・文節としての確信度が他のものより低いと考えられるためである。

これらの言語的制約と認識評価値を使い、コストを次の式から求める。

$$C(N,P,E) = \alpha \cdot N + \beta \cdot P + \gamma \cdot E \tag{3.1}$$

N: 文節数
P: 未登録語、1 文字語の単語数
E: 認識評価値

(文節列を構成する候補文字の 1 位候補との距離比の合計)

α, β, γ : 定数

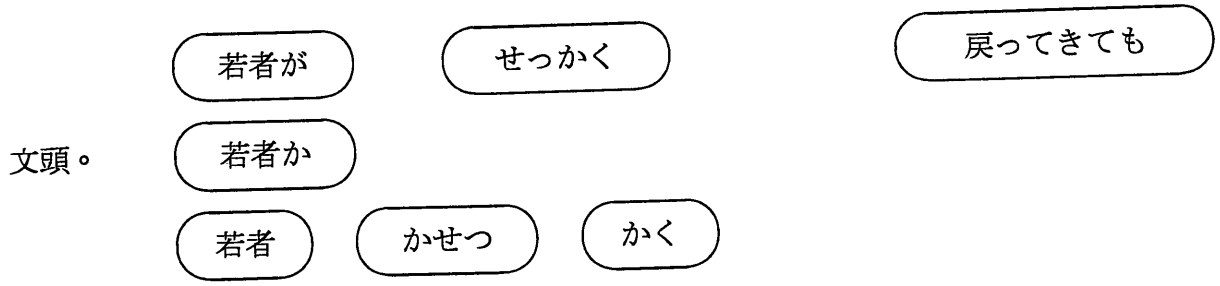
α, β, γ は、どの要素に評価の重みをおくかによって調整する。基本的に文節数が最小のものがコストが最小になり、同じ文節数のものについては認識評価値によってコストの大小が決まるようにパラメータを調節する。文を構成する文節の候補がない文字位置については個別認識の結果 (1 位候補) をそのまま後処理での結果とする。

また同じ文字列、同一の文節数でも異なる文節列の組合せが存在することがあり、この場合コスト関数により求まる値は等しくなることが多い。このとき文節の区切り方は異なったものになるが、文字列は等しいため認識結果として出力するには問題はない。

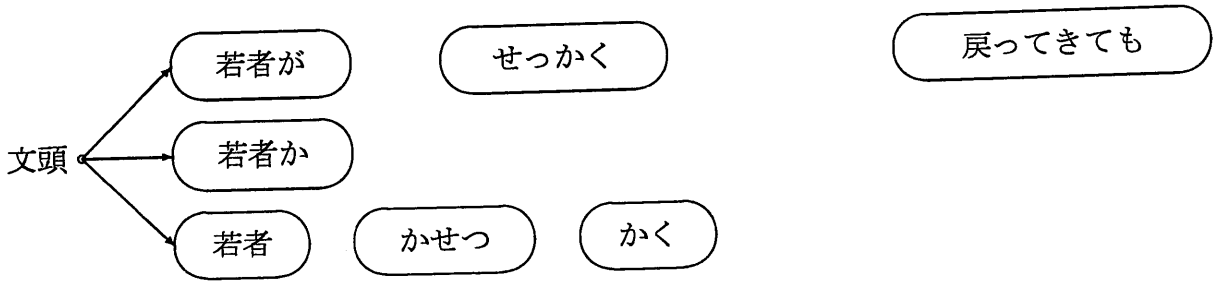
最尤文節列を選択することは、文節の接続を考えたグラフの最適なパスを求めることに相当し、グラフの最短経路探索アルゴリズムの手法によりコストが最小になるパスを求めることができる。図 3.14、3.15にパスの求め方を示す。各文字位置 i(文頭から数えて何文字目か) に抽出された文節が複数存在する。(図 3.14(a)) i を文頭を表す 0 から 1 ずつふやしていき、文字位置 i が末尾である文節が存在したら、その文節の後に接続可能である文節にパスをつなぐ。例えば、図 3.14(b) は、$i = 0$ の場合であり、文頭から接続可能である「若者が」「若者か」「若者」という 3 つの文節に、文頭からパスがつながれる。(c) は $i = 2$ の場合であり、2 文字目が末尾である文節「若者」から接続可能な文節「かせつ」にパスがつながれる。

探索の途中で単語・品詞が等しい文節を通る複数のパスがある場合には、その時点のパスのコストを求め、コストが最小のパスだけ残して他のパスは削除する。図 3.14(d) の例では、「若者が／せっかく」と「若者か／せっかく」の 2 つのパスが同じ「せっかく」という文節を通る。この時点での 2 つのパスのコストを比較すると、「か」より「が」のほうが認識評価値が小さいので、「若者が／せっかく」のパスがコストが小さくなり、文節列候補として残る。「若者か／せっかく」のパスは削除される。同様に図 3.15(f) では「若者が／せっかく／戻ってきても」と「若者／かせつ／かく／戻ってきても」の 2 つのパスのコストを比較し、文節数から「若者が／せっかく／戻ってきても」のパスが文節列候補として残ることになる。以上の処理を操り返し、最終的に図 3.15(g) に示す「若者が／せっかく／戻ってきても」というパスが最尤文節列となる。

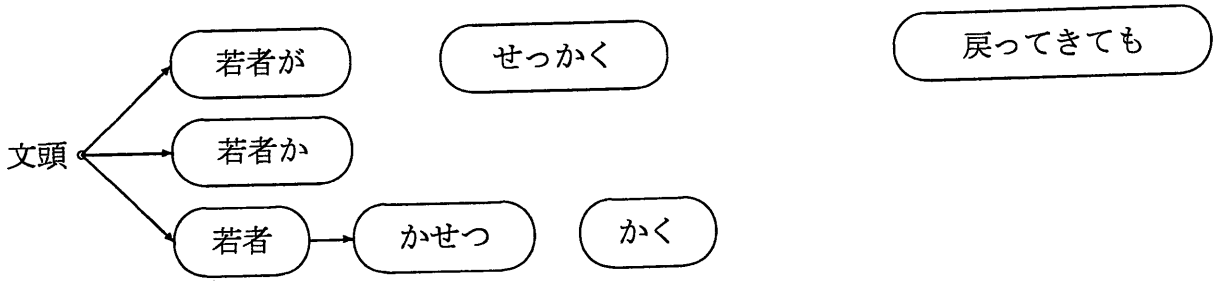
候補文字：若者がせっかく戻ってきても
か つが



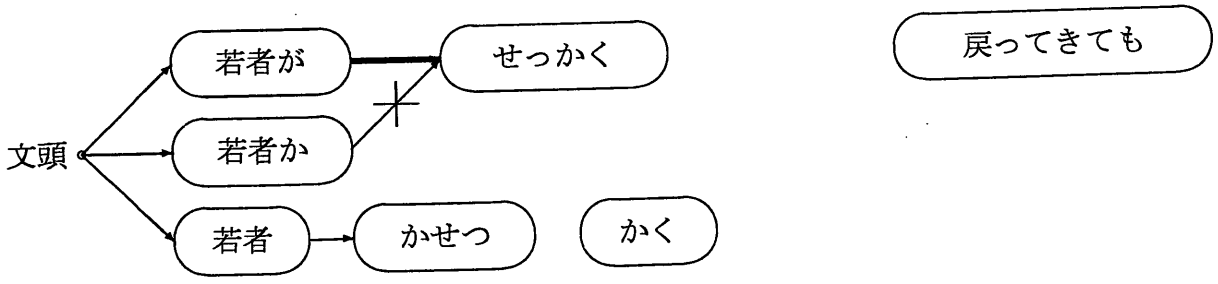
(a) 抽出された文節



(b) $i = 0$

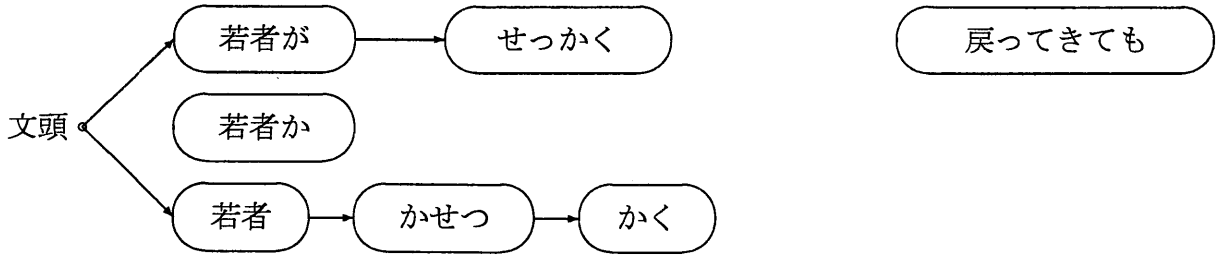


(c) $i = 2$

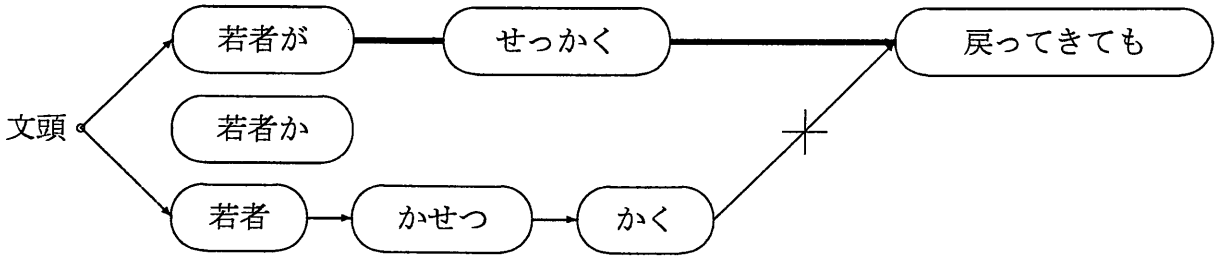


(d) $i = 3$

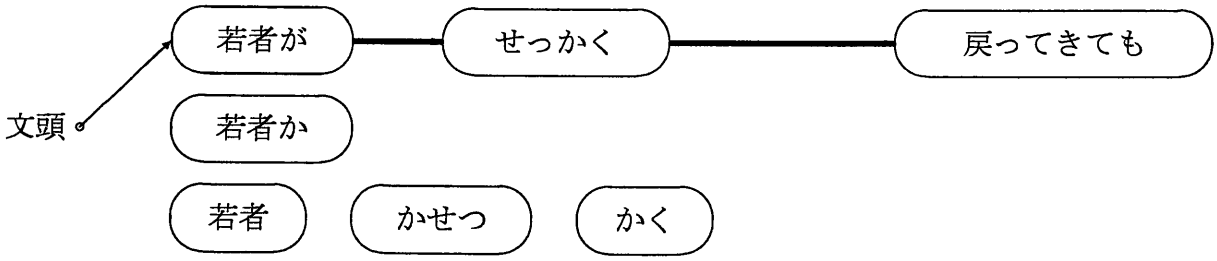
図 3.14: 最尤文節列選択 1



(e) $i = 5$



(f) $i = 7$



(g) 選択された最尤文節列

図 3.15: 最尤文節列選択 2

最尤文節列の探索に用いたグラフの最短経路探索手法を記号を用いて形式的に表現すると次のようになる。

最尤文節列探索アルゴリズム

照合範囲の長さを lとし、各文字位置 $i(0 \leq i \leq l)$ において n 個の文節候補が抽出されているとき、i 文字目の j 番目 $(0 \leq j \leq n)$ の文節候補を W_{ij} と表す。

最尤文節列選択とは文節列として $S = W_a, W_b, \ldots, W_z (W_a, W_b, W_z$は接続可能な文節候補) を選択したときその文節列から決まるコストを算出し、その値が最小となる文節列を最尤文節列とする作業である。

まず、使用する記号の意味を述べる。

各ノード D を以下の 5 項目で $\langle L, \overline{D}, W, C, H \rangle$ のように表す。

- L:文節の位置を文節の末尾文字の位置で表す。
- $\overline{D}$:ノード Dに至る直前のノード
- W:文節
- C:文節 Wまでの累積コスト
- H:文節 Wの品詞データ
  $(D_{ij}, \overline{D_{ij}}, C_{ij}, H_{ij}$ は文字位置が i の j 番目の文節候補 W_{ij} の各要素)

同一位置 i をもつノード集合を $NS(i)$ とする。また、**step 1** で使用する記号として次のものがある。

- $\emptyset$:空集合
- $D_\emptyset$:直前のノードが存在しない
- $W_\emptyset$:長さ 0 の仮想的文節
- B:文頭であることを表す

α, β, γ は式3.1で述べたものである。

step 1 $NS(i) \leftarrow \emptyset (i = 1, \ldots, l)$.
$NS(0) \leftarrow \langle 0, D_\emptyset, W_\emptyset, 0, B \rangle$.
$i \leftarrow 0$.

step 2 $NS(i)$ に属するすべてのノード (D_{ij}) について、

- D_{ij} と同じ W, H をもつノード ($\in NS(i)$) の中で最小の累積コスト (C) をもつものを選択しその値を C_{min} とおく。
- $C_{ij} > C_{min}$ ならば D_{ij} を $NS(i)$ から削除する。

step 3 $i = l$ ならば終了。
$i < l$ ならば i について **step 4** を行う。

step 4 生成された N_w 個の文節 ($W_k(k = 1, \ldots, N_w)$) のすべてについて、

- $NS(i)$ に属するノード
($D_{ij} = \langle i, \overline{D_{ij}}, W_{ij}, C_{ij}, H_{ij} \rangle$) を一つ選択 (すべて選択し終ったならば **step 4** を終了)。
- 次のようなノード D を対応するノード集合 $NS(i+len)$ に追加する (ここで len は W_k の単語長で $i + len \leq l$ を満たしていること)。
  - $L \leftarrow i + len$.
  - $\overline{D} \leftarrow D_{ij}$.
  - $W \leftarrow W_k$.
  - $C \leftarrow \alpha + \beta \cdot P + \gamma \cdot E + C_{ij}$.
    - α は文節数がひとつ増えたことによるコストの増加を表すもの
    - P:文節 W_k が未登録語または 1 文字語であれば 1、そうでなければ 0
    - E:文節 W_k の認識評価値
      (文節列を構成する候補文字の 1 位候補との距離比の合計)
  - $H \leftarrow W_k$ の品詞.

step 5 $i \leftarrow i + 1$ として **step 2** から繰り返す。

以上の手続きの結果得られる $NS(l)$ の中のノードから直前のノード $\overline{D_{ij}}$ を逆にたどればコストが最小である最尤文節列を求めることができる。

3.4 後処理実験

3.3節で構築した文字認識後処理システムの性能を評価するため、以下の実験を行った。

3.4.1 実験方法

認識させる文書データとして朝日新聞の社説 (1985 年のもの) から 50 編を使用する。1 編分をレーザプリンタより出力し、出力イメージを 400dpi でスキャナより読みこむ。この実験では文字の切り出し誤りの影響を除くため、図 3.16 のように格子状に文字を配置し、1 文字の文字領域を与えて実験を行った。この 1 文字ごとの文字画像を第 2章で述べた方向線素特徴量を用いて認識し、各文字の 10 位候補までの文字と認識評価値を後処理システムへの入力とする。

今回実験に用いた文章は新聞の社説であり人物や土地名などの固有名詞や専門用語など多くの未登録語が含まれている。そこで3.3.6で述べた未登録語処理がどの程度効果をあげているのかを調べるため、未登録語処理をしない場合と未登録語処理をした場合とで認識率がどのように変化するかを調べた。

また、個別認識率の低い文書でも後処理の効果があるかどうかをみるため、次の５つのデータを使用して同様の実験を行った。それぞれ社説 10 編分を使用する。

1. 10pt、明朝体、印刷は鮮明なもの
2. 1. を 1 回コピーしたもの
3. 1. を 2 回コピーしたもの
4. ゴシック体で印刷したもの
5. 5pt の文字で印刷したもの

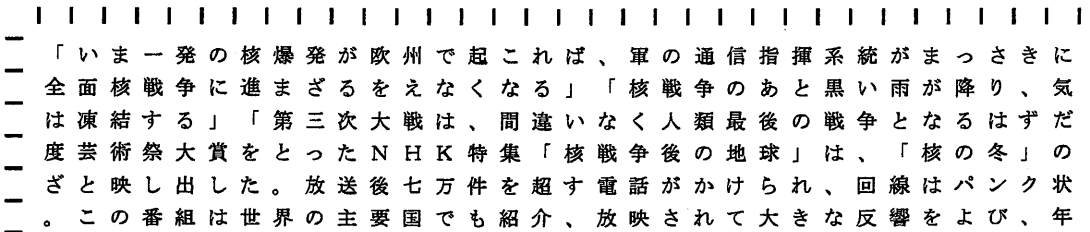


図 3.16: 認識データ例 (明朝体)

3.4.2 実験結果

個別文字認識での認識率と後処理を行った後での認識率とを表 3.1に示す。後処理前の認識率が 98.46% と比較的高い場合の実験であったが、個別文字認識における誤認識の 87.6% を正しく修正することができた。

図 3.17は社説 50 編それぞれについて後処理前と後処理後の認識率を表したものである。横軸は後処理前の認識率、縦軸は後処理後の認識率であり後処理の効果が大きいほど点線の上部に点が位置することになる。後処理前認識率 97% 台のデータでも後処理後の認識率 99%以上に向上していることがわかる。

後処理前認識率の低いデータの後処理結果は表 3.3、図 3.18のようになった。それぞれに認識率を向上しているが、個別認識で誤認識をどれだけなくすことができたかの割合を見てみると 2 回コピーしたものと 5pt のものが低い。これは全文字についてかすれやつぶれがひどく、候補中に正解文字が入らない場合が非常に多いためである。

表 3.1: 後処理結果

後処理前認識率	98.46% (誤り : 1126/73199)
後処理後認識率	99.70% (誤り : 219/73199)
訂正できた文字数	986 (誤認識の 87.6%)
訂正できなかった文字数	140 (誤認識の 12.4%)
誤訂正した文字数	79

表 3.2: 未登録語処理の効果

	未登録語処理なし→未登録語処理あり		
処理後認識率	99.56 %	→	99.70 %
訂正できた文字数	939	→	986
訂正できなかった文字数	187	→	140
誤訂正した文字数	131	→	79

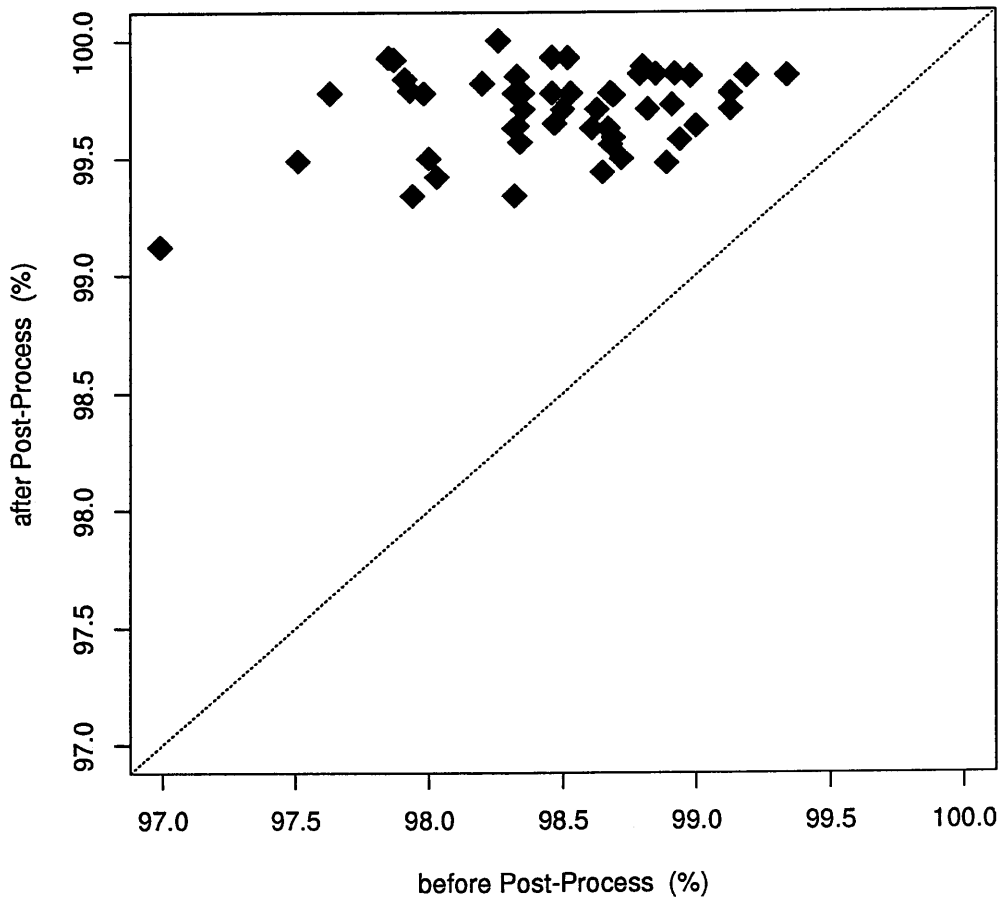


図 3.17: 後処理前後の認識率の変化

表 3.3: 後処理前認識率が低い場合の後処理効果

	明朝、10pt	1 回コピー	2 回コピー	ゴシック体	5pt
後処理前認識率	98.83%	97.84%	93.14%	95.02%	83.76%
後処理後認識率	99.58%	98.90%	94.89%	97.62%	88.26%
訂正できた文字数	69	98	185	239	471
(誤認識文字中の割合)	74.76%	60.10%	34.42%	61.68%	37.72%
誤訂正した文字数	11	15	49	36	120

(総文字数：7801 文字)

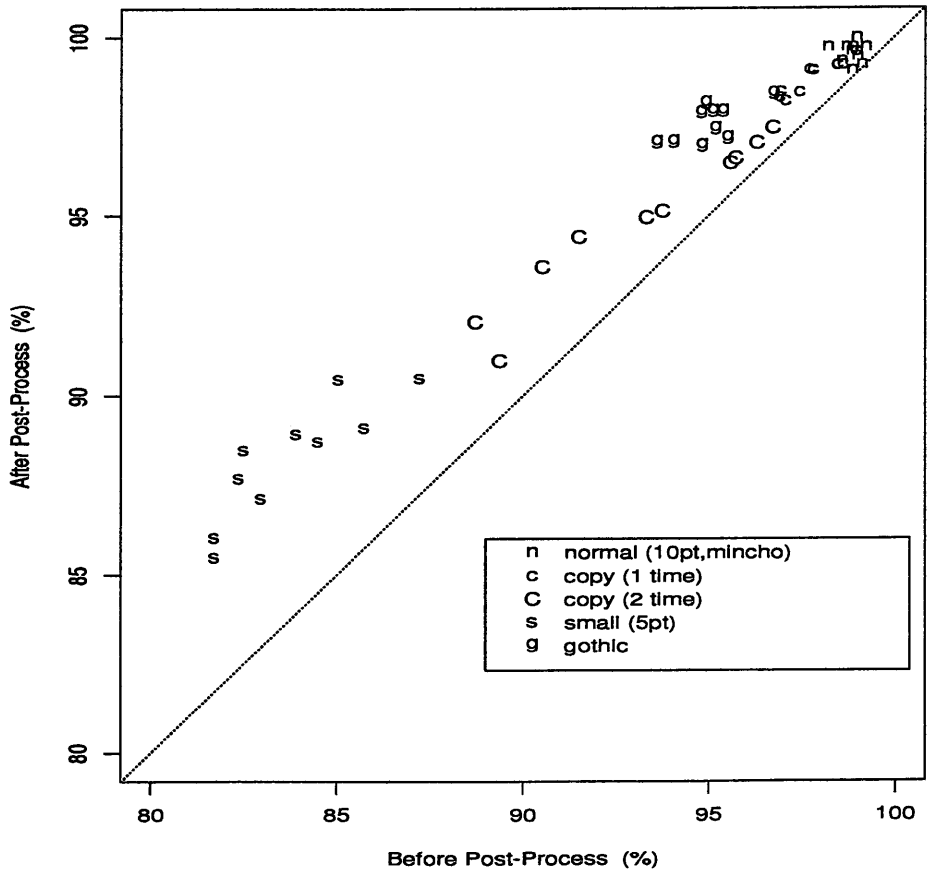


図 3.18: 後処理前認識率が低い場合の後処理前後の認識率の変化

3.4.3 訂正できなかった文字の考察

実験において個別認識での誤りを正しく修正できなかったり、個別認識では正しい結果を出していたのに後処理によって誤った結果に誤修正したりした個々の事例についてその原因や対処法について考察する。

3.4.3.1 未登録語

単語辞書に登録されていない片仮名語に対して未登録語処理を行なったが、

正解	認識結果
ニクソン	→ ニクノン
ＰＬＵＳ	→ ＰＩＵＳ

のように同一文字種間で誤認識してしまった場合には「ニクノン」を「ニクソン」に修正することは不可能となる。未登録語処理で単語とみなした文字列については、確実に単語であるという確証がないので、実際に使用する場合にはユーザにどのような文字列を未登録語として扱ったのか警告を出すといった処理が必要になる。

未登録語処理の悪影響としては「レーニン＝ゴルバチョフ」(等号) を「レーニンニゴルバチョフ」(片仮名の「に」) と誤修正する例があった。

また、「問題がらみで」「体力づくり」などの正解文節があるとそれぞれ

後処理前認識結果	：	後処理後認識結果
問題がらみで	→	問題からみで
体力づくり	→	体力つくり

と誤修正してしまう場合があった。このように２つの語が結びついて複合語になるとき、あとにくる語の最初の音が濁音になることを連濁と呼ぶ [13]。連濁は、２つの語の結び付きが強い場合に起こるが、連濁が起こらないからといって結び付きが弱いわけではない。連濁が起こる起こらないは必ずしも規則的ものではなく、連濁を起こす語とそうでない語がある。ただし結び付く語によって、次のような多少の傾向がある。

- 前にくる語の末尾の音が [ン] の場合起きる。(「本箱」・「三階」)
- 動詞と動詞とが結びついて複合動詞となっている場合には少ない。
  (「打ちつける」・「流れ込む」)。
- 後にくる語にすでに濁音が含まれているときは連濁しない。
  (「白壁」・「大風」)。

したがって、単純なルールで誤修正を防ぐのは困難であると考えられる。(前述の「対策づくり」の例があったからといって、名詞のあとの「つくり」をすべて「づくり」に変えようとすると「罪つくり」のような例外があり困難が生じる。)

3.4.3.2 複数の表記法があるもの

同じ単語を表記する場合にも複数の表記が存在する場合がある。今回の実験で現れたのは

文章での表記 : 辞書に登録されている表記
ウオーズ → ウォーズ

という例があった。漢字で登録されている単語が平仮名で表記された場合のために自立語平仮名辞書を作成したが「破たん」「かけ足」のように一部だけ平仮名で表記されていると単語として抽出することができなくなる。さらに平仮名で表記された部分が前後の文字と組み合わさり誤った文節を生成することもある。

複数の表記法があるものへの対策として考えられる手法は次のものがある。

1. 表記のあらゆる可能性を単語辞書に登録する
2. 片仮名表記の変化に関するルールを導入する
   「ウオ」 → 「ウォ」等
3. 漢字辞書と仮名辞書を使い、任意の漢字を平仮名表記に変換できるルールを導入し、一部仮名表記の単語にも対応できるようにする

しかし、1 は辞書の探索に、2,3 はルールから異表記を求めるのに時間がかかるため、後処理の計算速度が問題となる。

3.4.3.3 記号

出現する記号が「、」「。」などの句読点や『「』『」』などの括弧のみに限定されていれば簡単な規則を与えることにより処理できるが、新聞の社説だけでもこれらの他に「・」「—」「…」「～」などの記号が使われている。

記号は個別文字認識したとき、類似文字に他の記号が入る場合が多い。例えば次のようなものは文字パターンが類似したものである。

- 「。」「.」「・」「°」
- 「—」「ー」「-」
- 「)」「》」「〉」「>」

記号同志の類似文字の場合、言語情報を用いようにも文章を書く人により記号の使用方法が様々であるため、記号に関する規則を特定するのが難しい。一部の記号を見分けるためには、個別認識の段階で、文字パターンの位置情報などから限定するのが有効である。例えば、「一(いち)」と「—(ダッシュ)」は言語情報から判断するのがよいが「.(ピリオド)」と「・(センタードット)」はその位置から判断するのがよいと思われる。

3.4.3.4 候補外文字

候補として選択された文字のなかに正解がない場合、正しい文節を抽出できない。候補文字選択の過程で認識評価値の比の閾値が的確でなかったなどの理由も考えられるが、激しいかすれやノイズなどによって正解文字の尤度がかなり低くなることも考えられる。単語照合する際に、候補文字を組合せたもののすべてが辞書の単語と一致する完全一致のときだけを候補単語とせず、ある程度の割合で候補文字が単語辞書と一致していれば類似単語として単語候補に抽出するなどの方法が考えられている [14]。しかし、この類似検索アルゴリズムは住所や商品名などの単語の長さが長い場合には有効であるが、一般的な日本語単語を対象とした場合には、単語長が短いので類似単語が非常に多く検索されてしまうという問題がある。そこで、短い単語は完全一致による照合を行い、比較的長い単語にのみ類似検索を行うという方法も考えられている[15][16]。しかし、一般文の類似検索は処理速度と精度の面であまり効果はあがってはいない[7]。

一般に入力文書の品質が悪くなると、候補文字の中に正解がはいらない場合が増えると考えられる。今回の実験でも 2 回コピーしたデータや 5pt の文字のデータなどは候補外文字が非常に多く生じた。今回の実験では個別認識での候補文字を最高で 10 位までしかとっていないので、もっと多くの候補文字を後処理の対象とすることも考えられるが、激しくつぶれたものなどは、少しぐらい対象とする候補を増やしたところで正解が含まれるものでもない。候補数を増やしたときに累積認識率がどう変化するかを調べたのが図 3.19である。認識率の高いものは 10 位程度、低いものに関しても 30 位程度までになると頭打ちの状態になっており、対象とする候補数を増やしても無駄に誤った文字を対象としてしまうことになる。対象とする候補文字を単に増やすだけでは、後処理の効果がそれほど上がるものとは考えられない。

低品質の文字に対して、どのような認識結果 (候補文字) の傾向があらわれるかは個別認識する認識系の特性に依存するため、後処理の効果を更に高めるためには、認識系の特性を考慮することが必要になると考えられる。

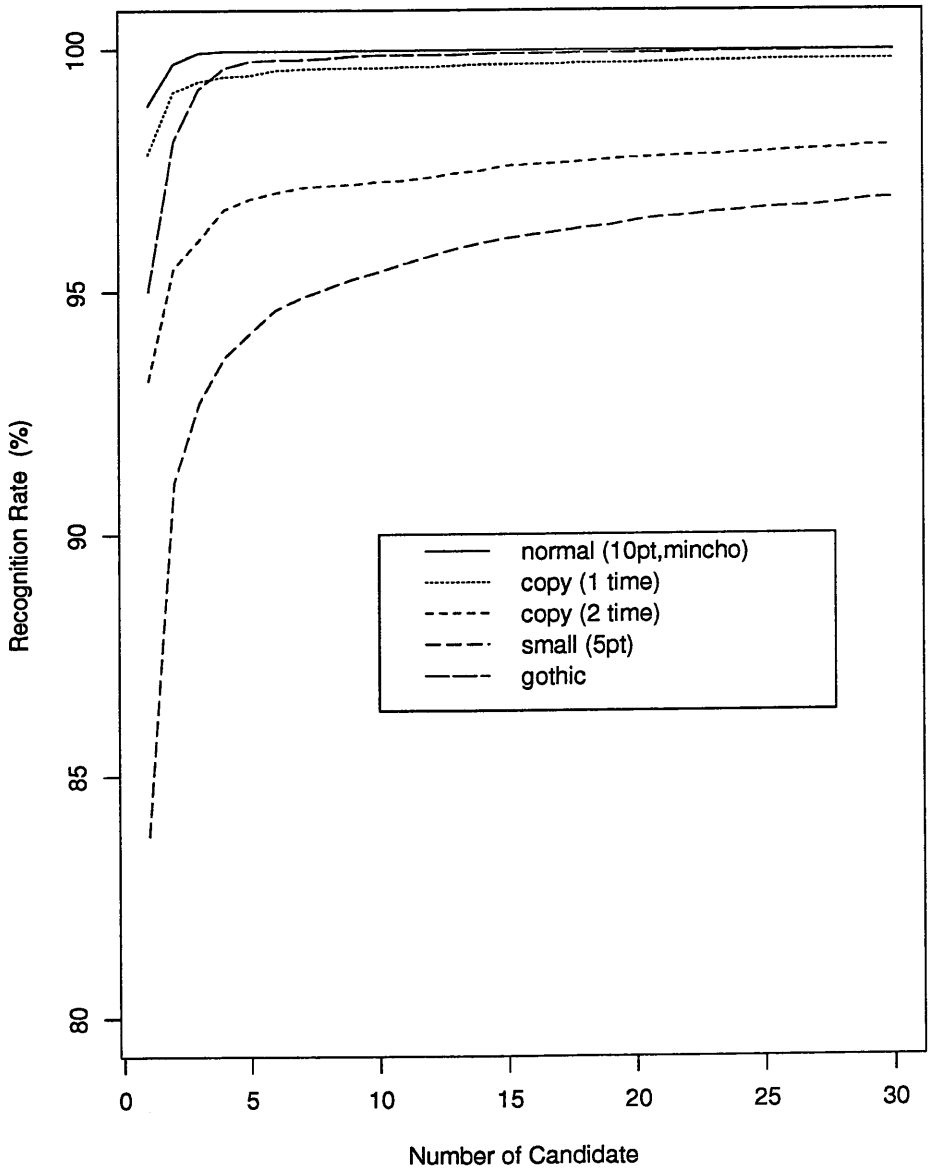


図 3.19: 候補数による累積認識率

3.4.3.5 接頭語・接尾語

今回の実験では直接精度を落す原因とはならなかったが、「～的」「～式」「初～」などの単語があった場合、その前後の文字と接続しているという情報は今回のシステムではもっていない。文節レベルでの処理で解決しようとすると、接続可能な単語を知識としてもつ、あるいは接続したものを単語辞書に登録するなどの方法が考えられる。

3.5 まとめ

言語情報を活用して、個別文字認識結果を日本語の文章として正しいと考えられる文字列へと修正する文字認識後処理システムについて述べた。まず、後処理システムに使用する日本語の知識として単語辞書・文法の構造を定めた。文字認識後処理システムの構成をまとめると次のようになる。

- 個別文字認識の認識評価値による候補文字選択
- 句読点を区切りとした照合範囲決定
- 辞書主導型の単語照合
- 単語接続可能性・文節成立性を用いた文節生成
- 片仮名・アルファベット・数字に対応した未登録語処理
- 最小文節法と個別認識評価値とを融合した最尤文節列選択

以上の構成により文字認識の高精度化が達成された。例えば、実際の文章を入力とした後処理の実験では個別認識における誤りの 87.6% を修正し、個別認識率を 98.46 % から 99.70 % へ向上させることができた。

第 4 章

文字切り出しを含めた文字認識への言語情報の活用

4.1 まえがき

文書認識を行う場合、文字の切り出しという処理が必要になる。これは多くの文字が書かれている文書の中から１文字ごとの文字領域を抽出する操作であり、基本的には連結した黒画素領域を抽出するという処理によって行われる。第1章で述べたように日本語の文書には分離文字や半角文字、アルファベット、種々の記号など、その文字ピッチが一定ではない文字が混合したものがあり、黒画素の連結情報からだけでは正確な切り出しは難しい。

第3章で、正しく切り出された文字の文字認識結果に対する後処理法を述べた。実際の文書認識システムへの応用を考えた場合、文字の切り出しを誤ると、この方法では正しく修正できないだけでなく誤修正をしてしまうケースも多くなると考えられる。これは文字の伸長・縮退などにより正解の単語と照合できずに文字数の異った不正解の単語を抽出してしまうことがあり得るからである。

ピッチの自由度の高い文書認識の高精度化のためには、文字ごとの認識結果を利用する方法が有効であると考えられる。この方法は文字ピッチが統合可能な複数の文字候補パターンに対し、イメージ上で統合した場合の認識結果とそれぞれのパターンに対する認識結果とを比較し、統合・非統合を決定するものである。

そこで本章では、認識結果を利用した文字切り出し方法に第3章で述べた文字認識後処理法を組込むことによって認識の高精度化をはかる方法について述べ、実験により評価する。

4.2 認識結果を利用した文字切り出し

ここではまず従来用いられている認識の評価値を利用する (言語情報は用いていない) 切り出しについて解説する。

4.2.1 行抽出

文書イメージから文字行を抽出する行抽出は、基本的に次のように行なわれる。まず読み取り範囲内のイメージの上端から 128 ラインずつを一時的に取り込み、行方向のヒストグラムを計算する。ヒストグラムの計算結果を参照し、ある閾値よりも大きい値が連続して見つかった場合にはこれを行の上端とみなす。次にこの上端から 128 ラインを一時的に取り込み、行方向のヒストグラムを計算する。計算結果を上から見ていき、ヒストグラムの値が 0 となるラインを探し、そのラインまでを 1 行とみなす。以後、この 1 行のイメージに対し以下の文字領域抽出を行う。

4.2.2 文字領域抽出

文字行として取り込まれたイメージに対し、以下の処理を行う。まずイメージの各列を縦方向にサーチしていったときに黒画素の全くない領域を区切りとし、各区間ごとに行ヒストグラムを計算する。これにより黒画素の塊が検出できるので、これをもとに平均文字サイズ (幅・高さ) を求める。すなわち各黒画素の塊の横方向縦方向の長さを 1 行の高さと比較し極端に違わない場合は、この塊をひとつの文字とみなす。文字とみなされた塊の横方向と縦方向の長さの平均を、この行の文字の平均文字幅・平均文字高さとみなす。

そして、行内の領域のうち、横幅が平均文字幅と比較して極端に大きいものがある場合 (しかもそれが引用の横線などの文字以外の要素ではない場合) は文字間接触があるとみなし、まず接触文字分割を行う。これを行うには黒画素の塊ごとの行ヒストグラムの他に列ヒストグラムが必要なので、まずこれを求める。最初に列ヒストグラムの各列ごとにその近傍を見て微分係数を求め、極小となる位置を探す。極小となる位置において、その列ヒストグラムの値がある閾値よりも小さいとき、その列を文字の区切りとみなす。以上により文字の区切りを定めてもまだ横幅が極端に大きい領域がある場合、その領域の左端から平均文字幅の 1/2 の整数にあたる列を求め、その近傍で極小となる列にラベルを付ける。そしてラベルの付けられた列を左から順に見ていき、そのヒストグラムの値がある閾値よりも小さいか、あるいは前のラベルが文字区切りとみなされなかった場合、そのラベルの位置を文字区切りとみなす。この操作によりすべての領域が平均文字幅程度 (またはそれ以下) に分割される。

次に、平均文字幅・平均文字高さと比較して、幅・高さともに小さい領域を探し、この

領域の存在する位置を考慮し句読点の可能性の高い場合は句読点フラグをつける。さらに平均文字幅から最大文字幅を定め句読点フラグのついていない領域について、隣接する領域をいくつか統合しても幅が最大文字幅以下になるときは、分離文字フラグをつけて認識後に統合処理を行うものとする。従って、分離文字フラグのついた文字は、各領域ごとに個々の文字であるとみなしたものと、それらをまとめてひとつの文字とみなしたものを両方とも認識部にまわす。

なお、句読点や幅の小さい文字となりうる文字は決まっているので、句読点フラグのついた文字は句読点のみの辞書、分離文字フラグのついた文字は幅の小さい文字のみの辞書に切り替えて認識を行う。

4.2.3 分離文字統合

文字であるとみなされる領域ごとの認識結果から、どのように文字を組合せていけば確からしい文字列となるのかを判定するのが分離文字統合である。ここでは認識結果を参照し、評価値の和が最小になるような組み合わせを見付けることで行う。具体例として入力されたイメージを図 4.1に示す。このとき黒画素の連続性からだけではAからIまでの9個の領域に分割されることになる。そこで最大文字幅を考えるとCとD、EとF、DとEがそれぞれ統合の可能性がある (統合した領域をそれぞれJ,K,Lとする)。これらの領域がどのように連結しているかを示したのが図 4.2のグラフである。このグラフの最適パスを見付けることが分離文字統合をすることに相当する。A〜L の各領域を文字とみなして認識した結果 (1 位候補) と評価値を図 4.1に示す。パスの評価は各領域を文字とみなして認識した結果の評価値をもとに以下のように求める。領域 i の 横幅を W_i、分離数 (黒画素の連結性を見たときの領域数) を n_iとし、第一候補の評価値を V_i とする。連結情報のグラフのパス $Path$ の評価値 E_{Path} を、

$$E_{Path} = \sum_{i \in Path} V_i \cdot W_i \cdot K^{n_i} \tag{4.1}$$

と定め、この評価値が最小となるパスを選択する (K は定数 ($K \leq 1$)。図 4.1の例では「軍事化に反対する」が認識結果として得られる。

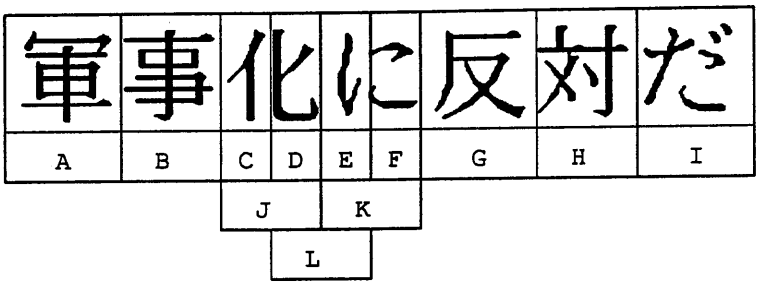


図 4.1: 入力イメージ

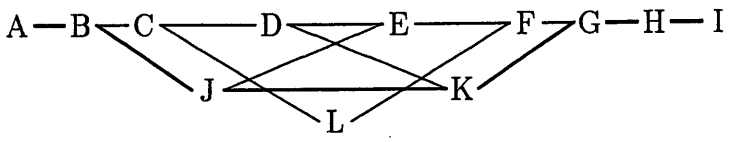


図 4.2: 各領域の連結情報

	候補文字	評価値		候補文字	評価値
A	軍	6666	G	反	6868
B	事	7337	H	対	5066
C	イ	13335	I	だ	5874
D	ヒ	14871	J	化	3930
E	|	33470	K	に	3570
F	こ	4546	L	‖	20901

表 4.1: 各領域の認識結果

4.3 言語情報を活用した文字切り出し

言語情報を用いて文字の認識を行う場合、まず文字領域抽出までは4.2と同様に行う。分離文字統合では各領域ごとにその下位候補までを考え、その組合せで日本語の単語・文節になる部分を抽出する。ここで行う下位候補の選択・単語照合・文節抽出の手法は第3章で述べた手法をそのまま用いる。図4.1の例において、下位候補文字の選択を行ったものが図4.3に示すものであり、これから抽出される文節は次の通りである。

軍事・軍事化・軍事化に・仏に・ごご・二反・反対・反対だ

抽出された文節から最尤文節列を選択する方法は、基本的には第3章で述べた方法である。ここで用いる各文節に与えられるコスト関数は4.2.3の式(4.1) から求められる値を使用する。

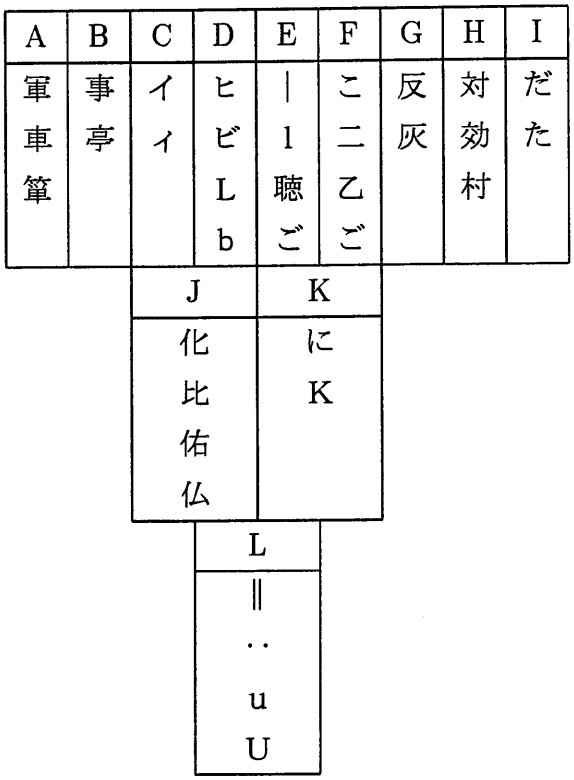


図 4.3: 各領域における候補文字

4.4 文書認識実験

実際の文書イメージにおいても言語処理を用いて認識の高精度化がはかれるかどうかを確かめるために以下の認識実験を行った。

4.4.1 実験方法

入力データとして朝日新聞の社説 20 編 (1 編 1000〜2000 文字) を TEX により出力した文書を 400dpi でスキャナより読み込み認識を行う。文書の文字サイズとしては 5pt, 7pt, 8pt, 9pt の 4 種類を用意した。

前述の認識結果のみを用いた場合と言語情報を活用した場合とで認識率を比較する。

4.4.2 実験結果と考察

認識結果のみを用いた場合と言語情報を活用した場合との認識率の差は表 4.2、図 4.4のようになった。ここでいう認識率とは以下のようにして求めたものである。

(正解文字数 ／ 入力文書中に出現する文字数) × 100

表 4.2: 文書認識結果

size	言語処理なし → 言語処理あり	
5pt	77.98 % →	83.58 %
7pt	96.15 % →	97.33 %
8pt	97.68 % →	98.17 %
9pt	98.67 % →	98.98 %

切り出し誤りを修正した例としては「イ」「ヒ」と分離したものを「化」、「ノ」「レ」と分離したものを「ル」と正したものが見られた。

一方、登録された名詞の後に接続した助詞の「に」は単語接続可能性から正しく文節の一部として抽出できるが、未登録語に接続した「に」の場合には文節の一部としては判定されず左と右に分離した画像がそれぞれ「む」と「こ」に認識され「むこ (婿)」という単語を抽出してしまう例などが見られた。このようなものは認識評価値を見ることにより正しくないものとして排除できると考えられるが、文節としての確からしさの閾値をどこに設けるかは大きな問題である。

4.5 まとめ

文書認識をする場合に、文字切り出しの段階から言語情報を活用して文字認識の高精度化をはかるため、第3章で構築した後処理システムを拡張した。これにより実際の文書の認識においても認識率を向上させることができた。

しかし、切り出された各文字領域での下位候補も含めた認識結果を考慮するために、抽出される文節候補の数が増大し、正確な文節列選択も困難になると考えられる。とくに未登録語などの影響によって文節を正確に抽出できない場合、「文節と考えられるが認識での評価は低い文字列」と「文節としては成立しないけれど認識での評価は高い文字列」のどちらを選択するかは非常に難しい。この問題を含めて文字列の評価方法を改良していくことが今後の課題となるであろう。

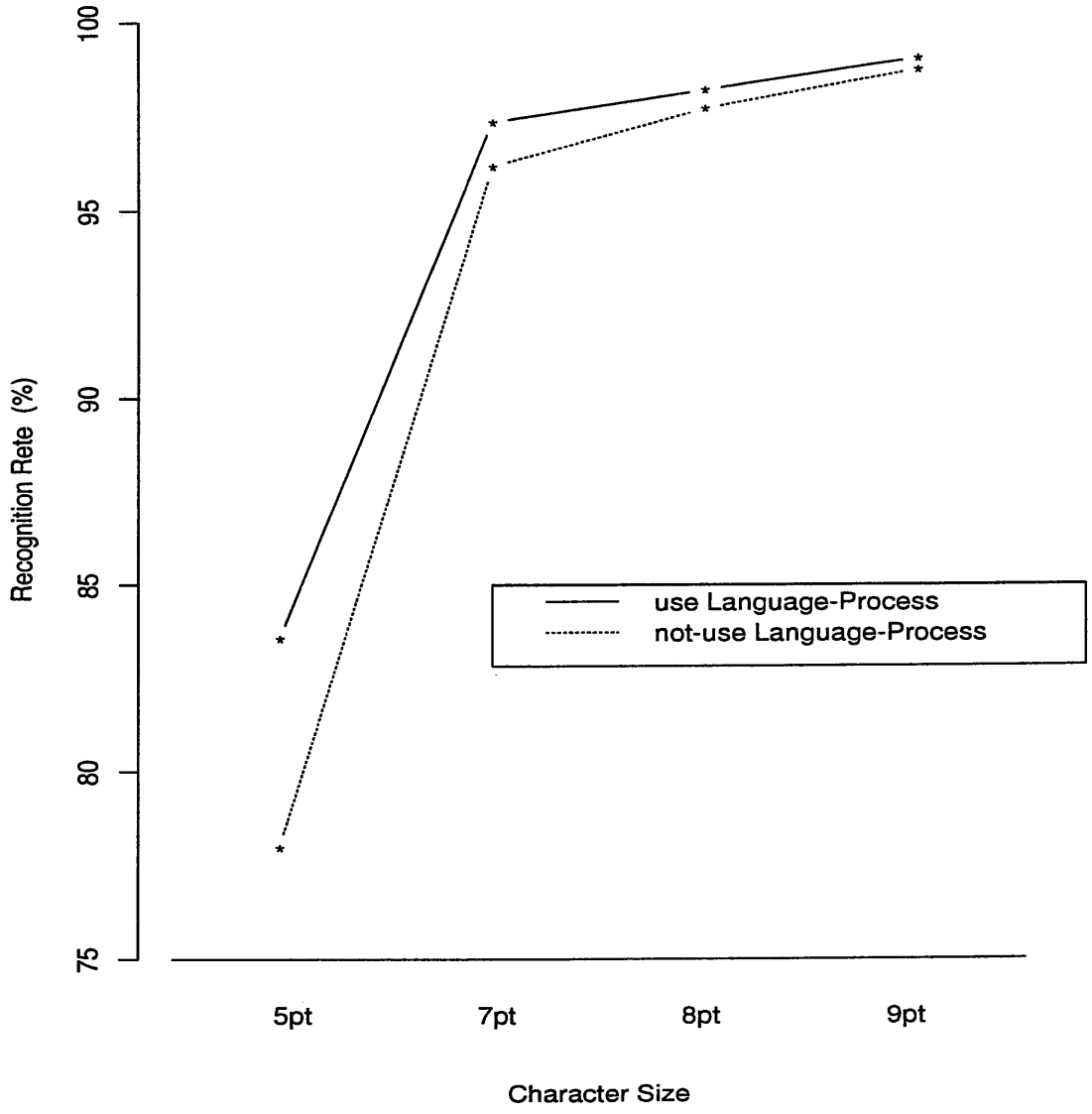


図 4.4: 文字サイズごとの言語処理による認識率の変化

第 5 章

結論

5.1 本研究の成果

文書の計算機入力の完全自動化を目指す文字認識に、言語情報を活用することによって認識精度を高めることを目的とし、以下のことを行った。

- 第2章では、個々の文字を認識するために用いる方向線素特徴量による認識手法について述べた。この手法により出力される候補文字と評価値が、本研究で構築する文字認識システムへの入力となる。

- 第3章では、個々の文字を認識した結果に対し、日本語の言語情報と個別認識の評価値とを活用することによって文字認識の誤りを修正する、文字認識後処理システムを構築した。使用した言語情報は日本語一般について成り立つものであり、対象となる日本語文の範囲を狭くするものではない。また実際に後処理の実験を行い、個別認識率を 98.46 % から 99.70 % に高精度化できることを確かめた。未登録語処理についても効果をあげていることがわかった。個別認識率が低いものについては、後処理をすることにより認識率を向上させることはできたが、個別認識率が高いものに比べて誤りを修正する割合は低くなることがわかった。

- 第4章では、一般文書の認識に対応できるように文字切り出しの段階から言語情報を組込むことで、切り出し・文字認識・言語処理を統合した高精度文字認識システムを構築した。切り出されていない文書の認識実験により言語情報を用いない認識システムよりも認識率を 0.3% ～ 5.6% 向上させることができた。

5.2 今後の課題

言語情報の活用により文字認識の精度を向上させることはできたが、実用的にはまだ十分な精度とはいえない。

まず本研究で構築したシステムは、日本語標準語の文法に従った文書でないと効果がでない。話し言葉や古文が引用された文書などを対象とする場合、個別認識結果を誤修正してしまうことが多くなると考えられる。この問題に対処するためには、辞書の登録語の追加、文体に応じた文法の整備などが必要になる。また、多くの人が認識システムを使用することを考えると、辞書への単語の登録や、新たな文法の追加などが効率よくできるシステムづくりが必要である。現段階では文法の追加は容易ではなく、辞書や文法規則の構造にも改善の余地があると考えられる。

また、個別認識率の低い文書に言語情報を活用した場合、個別認識での評価が低いものに無理に言語情報をあてはめたために、個別認識結果を誤修正してしまうことが多い。あらゆる文章に対して信頼できる言語情報を与えることは難しく、言語情報にたより過ぎないようにするため、本研究では認識評価値と言語情報とのふたつの要素から最終的な認識結果を決定している。しかし、文章らしさを評価する評価式やパラメータの値は、経験的に決定しているに過ぎない。認識システムを様々な文書に適用するには、言語情報を整えるとともに、個別認識系の特性にあわせて言語情報と認識評価値とをうまく融合させることが重要である。

謝辞

本研究を進めるにあたり、全般的な御指導を賜わりました東北大学工学部 阿曽 弘具教授に心より感謝致します。また、本論文をまとめるに際し貴重な御意見をいただいた東北大学工学部丸岡 章教授、東北大学電気通信研究所 佐藤 雅彦教授に深く感謝致します。

さらに、数多くの御指導、御助言をいただいた東北大学情報処理教育センター大町 真一郎博士、東北大学工学部 成富 敬博士、同じく 後藤 英昭氏、八戸高専 細越 淳一氏に深く感謝致します。

そして、本研究におきましては『九州芸工大自立語辞書ＫＩＤ－Ｊ９４』を使用させていただき、研究を大きくすすめることができました。自立語辞書の使用を御了承してくださった九州芸工大 稲永 紘之氏に深く感謝致します。

最後に、多面に渡り御意見、御協力をいただき、また日頃の生活において御世話になりました東北大学工学部 阿曽研究室の皆様に感謝致します。

参考文献

[1] 浅見直樹：「印刷漢字 OCR が市場に本格参入、一般オフィスの文書入力用に」
日経エレクトロニクス,1988.8.22(no.454),pp.179-184(1988 年 8 月).

[2] 浅見直樹：「OCR 市場が急成長、金融・保険業界で卓上型の大量導入相次ぐ」
日経エレクトロニクス,1988.10.17(no.458),pp.175-178(1988 年 10 月).

[3] (財) 関西文化学術研究都市推進機構：「学術研究支援のための高度情報システムに関する研究」(1989 年 3 月).

[4] Tappert,C.C.,Suen,C.Y.and Wakahara,T.:The State fo the Art in On-Line Handwriting Tecognition, *IEEE Transactions of Pattern Analysis and Machine Intelligence,* Vol.12,No.8,pp.787-808(1990).

[5] 阪倉篤義：「改稿 日本文法の話」 教育出版.

[6] 築島裕、白藤禮幸：「新編 国語の文法」明治書院.

[7] 西野文人：特集「自然言語処理技術の応用 5. 文字認識における自然言語処理」
情報処理,Vol.34 No.10,pp. 1274-1280 (1993 年 10 月).

[8] 印牧直文、中島健造、荒川弘熙：「単語辞書を活用した文字認識法の一検討」
信学技報,PRL81-91,pp.69-76 (1981 年)

[9] 黄瀬浩一、白石忠道、高松忍、福永邦雄：「構文・意味解析を用いた文字認識後処理法」
電子情報通信学会論文誌 (D-II),Vol.J77-D-II,No.11,pp.2199-2209 (1994 年 11 月).

[10] 孫寧、田原透、阿曽弘具、木村正行：「方向線素特徴量を用いた高精度文字認識」
電子情報通信学会論文誌 (D-II), Vol.J74-D-II, No.3, pp. 330-339 (1991 年 3 月).

[11] C. J. Hilditch : "Linear Skeleton from Square Cupboards"
In Machine Intelligence 6, B. Meltzer & D. Michie, Eds., Univ. Press, Edinburgh, pp. 403–420 (1969).

[12] 吉村賢治、日高達、吉田将：「文節数最小法を用いたべた書き日本語文の形態素解析」
情報処理学会論文誌 Jan.1983 Vol.24 No.1 (1983 年 1 月).

[13] 岩淵匡、桜井光昭、武部良明、森田良行：「日本語文法用語辞典」三省堂

[14] Hall, P.A.V. and Dowling, G.R.:Approximate String Mathing,
ACM Computing Surveys, Vol.12, No.4, pp.381-402 (1980).

[15] 高尾哲康、西野文人：「日本語文書リーダ後処理の実現と評価」
情報処理学会論文誌 Nov.1989 Vol.30 No.11 (1989 年 11 月).

[16] 木谷強：「手書き文書の文字認識結果に対する後処理方式」
自然言語処理研究会 86-1 (1991 年).

[17] 伊東伸泰、丸山宏：「OCR 入力された日本語文の誤り検出と自動訂正」
情報処理学会論文誌 May.1992 Vol.33 No.5 (1992 年 5 月).

[18] 久光徹、丸川勝美、嶋好博、藤沢浩道、新田義彦：
「OCR 誤認識後処理の効果について –候補単語抽出方法と動詞活用処理を中心に–」
情報処理学会研究報告 94-NL-104-3 (1994 年 11 月).

[19] 杉村利明：「候補文字補完と言語処理による漢字認識の誤り訂正処理法」
電子情報通信学会論文誌 (D-II) Vol.J72-D-II No.7 (1989 年 7 月).

研究業績

[1] 秋山秀三、阿曽弘具：「単語情報を利用する文字認識手法」
平成 5 年度電気関係学会東北支部連合大会, 1D–13, (平成 5 年 9 月)

[2] 秋山秀三、阿曽弘具：「言語情報を利用する文字認識誤り訂正手法」
平成 6 年度電気関係学会東北支部連合大会, 2D–13, (平成 6 年 8 月)

付録 A

使用付属語一覧

本研究で用いた付属語 (形式名詞、助動詞、助詞、補助用言、活用語尾を含む) およびその活用型・品詞・活用形を次に示す。表の構成は以下のとおりである。

付属語のつづり	活用型	品詞 1	品詞 2	活用形

品詞 1 ・品詞 2 は、それぞれ前に接続する単語による品詞分類、後に接続する単語による品詞分類であり、本研究では品詞 2 のみ使用している。活用型と活用形のコードは表 A.1 のように分類されている。

表 A.1: 活用型および活用形のコード

ア、ワ行五段活用	1
カ行五段活用	2
サ行五段活用	3
タ行五段活用	4
ナ行五段活用	5
マ行五段活用	6
ラ行五段活用	7
ガ行五段活用	8
バ行五段活用	9
上一段活用	10
下一段活用	11
形容詞	12
形容動詞	13
サ変動詞	14
ザ変動詞	15
ラ変動詞	25

活用なし	0
未然形	1
連用形	2
終止形	3
連体形	4
仮定形	5
命令形	6

うち	00	001	001	0
おり	00	001	001	0
こと	00	001	001	0
ため	00	001	001	0
とき	00	001	001	0
ところ	00	001	001	0
はず	00	001	001	0
ほう	00	001	001	0
ほか	00	001	001	0
まま	00	001	001	0
もの	00	001	001	0
よう	00	001	001	0
う	00	051	051	3
う	00	051	052	4
ごとく	00	054	053	2
ごとき	00	054	054	4
させ	00	057	055	1
させ	00	057	056	2
させる	00	057	057	3
させる	00	057	058	4
させれ	00	057	059	5
させろ	00	057	060	6
させよ	00	057	061	6
しめ	00	064	062	1
しめ	00	064	063	2
しめる	00	064	064	3
しめる	00	064	065	4
しめれ	00	064	066	5
しめよ	00	064	067	6
ず	00	069	068	2
ず	00	069	069	3
ざる	00	069	070	4
ざれ	00	069	071	5
ず	00	069	072	5
ずん	00	069	073	5
せ	00	076	074	1
し	00	076	075	2
せ	00	076	075	2
せる	00	076	076	3
せる	00	076	077	4
せれ	00	076	078	5
せろ	00	076	079	6
せよ	00	076	080	6
そうで	00	082	081	2
そうだ	00	082	082	3
そうだろ	00	087	083	1
そうで	00	087	084	2
そうに	00	087	085	2
そうだっ	00	087	086	2
そうだ	00	087	087	3
そうな	00	087	088	4
そうなら	00	087	089	5
そうでし	00	091	090	2
そうです	00	091	091	3
そうでしょ	00	094	092	1
そうでし	00	094	093	2
そうです	00	094	094	3
たら	00	096	098	5
だろ	00	100	099	1
だ	00	100	100	3
だ	00	100	101	4
だら	00	100	102	5
だろ	00	106	103	1
で	00	106	104	2
だっ	00	106	105	2
だ	00	106	106	3
な	00	106	107	4
なら	00	106	108	5
たかろ	00	112	109	1
たく	00	112	110	2
たかっ	00	112	111	2
たい	00	112	112	3
たい	00	112	113	4
たけれ	00	112	114	5
たがら	00	119	115	1
たがろ	00	119	116	1
たがり	00	119	117	2
たがっ	00	119	118	2
たがる	00	119	119	3
たがる	00	119	120	4
たがれ	00	119	121	5
たる	00	122	122	4
たれ	00	122	123	6
だろう	00	124	124	3
てる	00	127	127	3
てる	00	127	128	4
てれ	00	127	129	5
てろ	00	127	130	6
てよ	00	127	131	6
であろう	00	135	132	1
であり	00	135	133	2
であっ	00	135	134	2
である	00	135	135	3
である	00	135	136	4
であれ	00	135	137	5
であれ	00	135	138	6
であろう	00	139	139	3
でしょう	00	140	140	3
でしょ	00	143	141	1
でし	00	143	142	2
です	00	143	143	3
です	00	143	144	4
なかろ	00	148	145	1
なく	00	148	146	2
なかっ	00	148	147	2
ない	00	148	148	3
ない	00	148	149	4
なけれ	00	148	150	5
なる	00	151	151	4
なれ	00	151	152	6
ん	00	154	153	2
ん	00	154	154	3
ぬ	00	154	155	3
ん	00	154	156	4

ぬ	00	154	157	4
ね	00	154	158	5
べから	00	161	159	1
べく	00	161	160	2
べし	00	161	161	3
べき	00	161	162	4
まい	00	163	163	3
まい	00	163	164	4
なさいませ	00	168	165	1
ませ	00	168	165	1
なさいましょ	00	168	166	1
ましょ	00	168	166	1
なさいまし	00	168	167	2
まし	00	168	167	2
なさいます	00	168	168	3
ます	00	168	168	3
なさいます	00	168	169	4
ます	00	168	169	4
なさいますれ	00	168	170	5
ますれ	00	168	170	5
なさい	00	168	171	6
なさいませ	00	168	171	6
ませ	00	168	171	6
なさいまし	00	168	172	6
まし	00	168	172	6
なさら	00	168	310	1
なさろ	00	168	311	1
なさり	00	168	312	2
なさっ	00	168	313	2
なさる	00	168	314	3
なさる	00	168	315	4
なされ	00	168	316	5
なされ	00	168	317	6
みたいだろ	00	177	173	1
みたいで	00	177	174	2
みたいに	00	177	175	2
みたいだっ	00	177	176	2
みたいだ	00	177	177	3
みたいな	00	177	178	4
みたいなら	00	177	179	5
みたいでしょ	00	182	180	1
みたいでし	00	182	181	2
みたいです	00	182	182	3
よう	00	183	183	3
よう	00	183	184	4
ようだろ	00	189	185	1
ようで	00	189	186	2
ように	00	189	187	2
ようだっ	00	189	188	2
ようだ	00	189	189	3
ような	00	189	190	4
ようなら	00	189	191	5
ようでしょ	00	194	192	1
ようでし	00	194	193	2
ようです	00	194	194	3
らしく	00	198	195	2
らしかっ	00	198	196	2
らしゅう	00	198	197	2
らしい	00	198	198	3
らしい	00	198	199	4
られ	00	202	200	1
られ	00	202	201	2
られる	00	202	202	3
られる	00	202	203	4
られれ	00	202	204	5
られろ	00	202	205	6
られよ	00	202	206	6
れ	00	209	207	1
れ	00	209	208	2
れる	00	209	209	3
れる	00	209	210	4
れれ	00	209	211	5
れろ	00	209	212	6
れよ	00	209	213	6
が	00	214	214	0
から	00	215	215	0
で	00	217	217	0
と	00	218	218	0
に	00	219	219	0
の	00	220	220	0
へ	00	221	221	0
や	00	222	222	0
より	00	223	223	0
を	00	224	224	0
か	00	225	225	0
きり	00	226	226	0
ぎり	00	227	227	0
くらい	00	228	228	0
ぐらい	00	229	229	0
すら	00	231	231	0
ずつ	00	232	232	0
だけ	00	233	233	0
だの	00	234	234	0
ながら	00	235	235	0
なぞ	00	236	236	0
など	00	237	237	0
なり	00	238	238	0
のみ	00	239	239	0
ばかし	00	240	240	0
ばかり	00	241	241	0
ほど	00	242	242	0
まで	00	243	243	0
やら	00	244	244	0
こそ	00	245	245	0
さえ	00	246	246	0
しか	00	247	247	0
でも	00	248	248	0
なんて	00	249	249	0
は	00	250	250	0
も	00	251	251	0
が	00	252	252	0
から	00	253	253	0
くせに	00	254	254	0
けど	00	255	255	0
けれど	00	256	256	0
し	00	257	257	0

たり	00	258	258	0
だり	00	259	259	0
つつ	00	260	260	0
て	00	262	262	0
で	00	263	263	0
と	00	264	264	0
ど	00	265	265	0
なり	00	266	266	0
に	00	267	267	0
ので	00	268	268	0
のに	00	269	269	0
ば	00	270	270	0
や	00	271	271	0
ゆえ	00	272	272	0
な	00	273	273	0
ね	00	274	274	0
ね	00	276	276	0
よ	00	278	278	0
さ	00	279	279	0
の	00	280	280	0
いらっしゃいませ	00	283	165	1
いらっしゃいましょ	00	283	166	1
いらっしゃいまし	00	283	167	2
いらっしゃいます	00	283	168	3
いらっしゃいます	00	283	169	4
いらっしゃいますれ	00	283	170	5
いらっしゃいませ	00	283	171	6
いらっしゃいまし	00	283	172	6
い	00	283	281	1
い	00	283	282	2
いる	00	283	283	3
いる	00	283	284	4
いれ	00	283	285	5
いらっしゃい	00	283	286	6
いろ	00	283	286	6
いよ	00	283	287	6
いらっしゃら	00	283	310	1
いらっしゃろ	00	283	311	1
いらっしゃり	00	283	312	2
いらっしゃっ	00	283	313	2
いらっしゃる	00	283	314	3
いらっしゃる	00	283	315	4
いらっしゃれ	00	283	316	5
いらっしゃれ	00	283	317	6
くださいませ	00	290	165	1
下さいませ	00	290	165	1
くださいましょ	00	290	166	1
下さいましょ	00	290	166	1
くださいまし	00	290	167	2
下さいまし	00	290	167	2
くださいます	00	290	168	3
下さいます	00	290	168	3
くださいます	00	290	169	4
下さいます	00	290	169	4
くださいませ	00	290	170	5
下さいませ	00	290	170	5
くださいませ	00	290	171	6
下さいませ	00	290	171	6
くださいまし	00	290	172	6
下さいまし	00	290	172	6
こ	00	290	288	1
き	00	290	289	2
くる	00	290	290	3
くる	00	290	291	4
くれ	00	290	292	5
ください	00	290	293	6
くれ	00	290	293	6
こい	00	290	293	6
下さい	00	290	293	6
くださら	00	290	310	1
下さら	00	290	310	1
くださろ	00	290	311	1
下さろ	00	290	311	1
くださり	00	290	312	2
下さり	00	290	312	2
くださっ	00	290	313	2
下さっ	00	290	313	2
くださる	00	290	314	3
下さる	00	290	314	3
くださる	00	290	315	4
下さる	00	290	315	4
くだされ	00	290	316	5
下され	00	290	316	5
くだされ	00	290	317	6
下され	00	290	317	6
なら	00	314	310	1
なろ	00	314	311	1
なり	00	314	312	2
なっ	00	314	313	2
なる	00	314	314	3
なる	00	314	315	4
なれ	00	314	316	5
なれ	00	314	317	6
よら	00	322	318	1
よろ	00	322	319	1
より	00	322	320	2
よっ	00	322	321	2
よる	00	322	322	3
よる	00	322	323	4
よれ	00	322	324	5
よれ	00	322	325	6
ございませ	00	329	165	1
ございましょ	00	329	166	2
ございまし	00	329	167	2
ございます	00	329	168	3
ございます	00	329	169	4
ございますれ	00	329	170	5
ございませ	00	329	171	6
ございまし	00	329	172	6
あろ	00	329	326	1
あり	00	329	327	2
あっ	00	329	328	2
ある	00	329	329	3
ある	00	329	330	4
あれ	00	329	331	5
あれ	00	329	332	6

おこ	00	337	333	1
おか	00	337	334	1
おき	00	337	335	2
おい	00	337	336	2
おく	00	337	337	3
おく	00	337	338	4
おけ	00	337	339	5
おけ	00	337	340	6
しまわ	00	345	341	1
しまお	00	345	342	1
しまっ	00	345	343	2
しまい	00	345	344	2
しまう	00	345	345	3
しまう	00	345	346	4
しまえ	00	345	347	5
しまえ	00	345	348	6
み	00	351	349	1
み	00	351	350	2
みる	00	351	351	3
みる	00	351	352	4
みれ	00	351	353	5
みろ	00	351	354	6
みよ	00	351	355	6
わ	01	017	010	1
お	01	017	011	1
い	01	017	014	2
っ	01	017	016	2
う	01	017	017	3
う	01	017	018	4
え	01	017	019	5
え	01	017	020	6
え	01	017	021	1
え	01	017	022	2
える	01	017	023	3
える	01	017	024	4
えれ	01	017	025	5
か	02	017	010	1
こ	02	017	011	1
い	02	017	012	2
き	02	017	014	2
く	02	017	017	3
く	02	017	018	4
け	02	017	019	5
け	02	017	020	6
け	02	017	021	1
け	02	017	022	2
ける	02	017	023	3
ける	02	017	024	4
けれ	02	017	025	5
さ	03	017	010	1
そ	03	017	011	1
し	03	017	014	2
す	03	017	017	3
す	03	017	018	4
せ	03	017	019	5
せ	03	017	020	6
せ	03	017	021	1
せ	03	017	022	2
せる	03	017	023	3
せる	03	017	024	4
せれ	03	017	025	5
た	04	017	010	1
と	04	017	011	1
ち	04	017	014	2
っ	04	017	016	2
つ	04	017	017	3
つ	04	017	018	4
て	04	017	019	5
て	04	017	020	6
て	04	017	021	1
て	04	017	022	2
てる	04	017	023	3
てる	04	017	024	4
てれ	04	017	025	5
な	05	017	010	1
の	05	017	011	1
に	05	017	014	2
ん	05	017	015	2
ぬ	05	017	017	3
ぬ	05	017	018	4
ね	05	017	019	5
ね	05	017	020	6
ね	05	017	021	1
ね	05	017	022	2
ねる	05	017	023	3
ねる	05	017	024	4
ねれ	05	017	025	5
ま	06	017	010	1
も	06	017	011	1
み	06	017	014	2
ん	06	017	015	2
む	06	017	017	3
む	06	017	018	4
め	06	017	019	5
め	06	017	020	6
め	06	017	021	1
め	06	017	022	2
める	06	017	023	3
める	06	017	024	4
めれ	06	017	025	5
ら	07	017	010	1
ろ	07	017	011	1
り	07	017	014	2
っ	07	017	016	2
る	07	017	017	3
る	07	017	018	4
れ	07	017	019	5
れ	07	017	020	6
れ	07	017	021	1
れ	07	017	022	2
れる	07	017	023	3
れる	07	017	024	4
れれ	07	017	025	5
が	08	017	010	1
ご	08	017	011	1
い	08	017	013	2
ぎ	08	017	014	2
ぐ	08	017	017	3

ぐ	08	017	018	4
げ	08	017	019	5
げ	08	017	020	6
げ	08	017	021	1
げ	08	017	022	2
げる	08	017	023	3
げる	08	017	024	4
げれ	08	017	025	5
ば	09	017	010	1
ぼ	09	017	011	1
び	09	017	014	2
ん	09	017	015	2
ぶ	09	017	017	3
ぶ	09	017	018	4
べ	09	017	019	5
べ	09	017	020	6
べ	09	017	021	1
べ	09	017	022	2
べる	09	017	023	3
べる	09	017	024	4
べれ	09	017	025	5
る	10	023	023	3
る	10	023	024	4
れ	10	023	025	5
ろ	10	023	026	6
よ	10	023	027	6
る	11	023	023	3
る	11	023	024	4
れ	11	023	025	5
ろ	11	023	026	6
よ	11	023	027	6
かろ	12	006	003	1
かっ	12	006	004	2
く	12	006	005	2
い	12	006	006	3
い	12	006	007	4
けれ	12	006	008	5
だろ	13	307	303	1
で	13	307	304	2
に	13	307	305	2
だっ	13	307	306	2
だ	13	307	307	3
な	13	307	308	4
なら	13	307	309	5
し	14	032	028	1
さ	14	032	029	1
せ	14	032	030	1
し	14	032	031	2
する	14	032	032	3
する	14	032	033	4
すれ	14	032	034	5
しろ	14	032	035	6
せよ	14	032	036	6
じ	15	032	028	1
ざ	15	032	029	1
ぜ	15	032	030	1
じ	15	032	031	2
ずる	15	032	032	3
ずる	15	032	033	4
ずれ	15	032	034	5
じろ	15	032	035	6
ぜよ	15	032	036	6
ら	25	017	010	1
ろ	25	017	011	1
り	25	017	014	2
っ	25	017	016	2
る	25	017	017	3
る	25	017	018	4
れ	25	017	019	5
い	25	017	020	6
れ	25	017	020	6
い	25	017	355	2
ついて	00	000	500	0
つきまして	00	000	500	0
関して	00	000	500	0
関しまして	00	000	500	0
すぎ	00	000	426	1
すぎ	00	000	427	2
すぎる	00	000	023	3
すぎる	00	000	024	4
すぎれ	00	000	025	5
がち	00	000	413	0
す	00	000	501	3
ず	00	000	501	3
でき	00	000	021	1
でき	00	000	022	2
できる	00	000	023	3
できる	00	000	024	4
できれ	00	000	025	5
来	00	290	288	1
来	00	290	289	2
来る	00	290	290	3
来る	00	290	291	4
来れ	00	290	292	5
来い	00	290	293	6

付録 B

接続可能単語・文節成立性一覧

本研究では第3章で述べたように単語の接続可能性・文節成立性という文法規則を用いている。ここではその一覧を示す。

表の構成はB.1の通りである。

(例)

228	くらい	○	格助詞、係助詞、だ、である、であろう、でしょう、です
1	2	3	4

1. 品詞コード

2. 接続される単語

3. その単語で終わったときに文節として成立するかどうか
   (○ → 成立する、× → 成立しない)

4. 接続可能な付属語
   接続する付属語が助動詞のように活用する単語の場合は、その終止形のみを記した。この欄に数字が記入されている単語は、接続可能単語、文節成立性がその数字が示す品詞コードの単語と同様であることを示す。また「×」が記入されているものは接続可能単語がないことを示す。

表 B.1: 接続可能単語・文節成立性一覧表例

B.0 活用なし

1	形式名詞	○	だ、である、であろう、でしょう、です、らしい、格助詞、副助詞、係助詞
416	名詞		1
417	連体詞	○	×
418	接続詞	○	×
419	感動詞	○	と
420	副詞	○	×
214	が	○	×
215	から	○	の、は、か、くらい、ぐらい
217	で	○	の、すら、のみ、ほど、係助詞
218	と	○	の、くらい、ぐらい、すら、だけ、なぞ、など、のみ、ばかし、ばかり、係助詞
219	に	○	くらい、ぐらい、すら、だけ、なぞ、など、のみ、ばかし、ばかり、係助詞、ついて、関して
220	の	○	形式名詞、だ、である、であろう、でしょう、です、らしい、格助詞、副助詞、係助詞
221	へ	○	の、と、も、は
222	や	○	×
223	より	○	は、も
224	を	○	も
225	か	○	を、と、な、も、が、の、は
226	きり	○	だ、である、であろう、でしょう、です
227	ぎり		226
228	くらい	○	格助詞、係助詞、だ、である、であろう、でしょう、です
229	ぐらい		228
231	すら	○	も
232	ずつ	○	×
233	だけ		228
234	だの	○	×
235	ながら	○	も
236	なぞ	○	は
237	など		228
238	なり	○	に、の
239	のみ		228
240	ばかし		228
241	ばかり		228
242	ほど		228
243	まで		228
244	やら	○	格助詞
245	こそ	○	が、は
246	さえ	○	も
247	しか	○	×
248	でも	○	×
249	なんて	○	×
250	は	○	×
251	も	○	が
252	が	○	×
253	から	○	は、だ、である、であろう、でしょう、です
254	くせに	○	×
255	けど	○	も
256	けれど	○	も
257	し	○	×
258	たり	○	係助詞、する
259	だり		258
260	つつ	○	も

262	て	○	なさいます、なさる、の、係助詞、から、いらっしゃいます、いる、いらっしゃる、くださいます (下さいます)、くる、くださる (下さる)、ございます、ある、おく、しまう、みる
263	で		262
264	と	○	も
265	ど	○	も
266	なり	○	×
267	に	○	×
268	ので	○	×
269	のに	○	×
270	ば	○	こそ
271	や	○	×
272	ゆえ	○	に
273	な	○	×
274	ね	○	×
276	ね	○	×
278	よ	○	×
279	さ	○	×
280	の	○	×

B.1 未然形

10	-a (五段)	×	ず、せる、ない、ぬ、れる
11	-o (五段)		3
21	れ (五段)	×	ず、ぬ、ない、まい、よう
426	(上一、下一)	×	られる、させる、ない、ぬ、ず、よう、まい
28	し、じ (サ変)	×	ない、まい、よう
29	さ、ざ (サ変)	×	せる、れる
30	せ、ぜ (サ変)	×	ず、ぬ
3	かろ (形)	×	う
303	だろ (形動)		3
55	させ	×	られる、ない、ぬ、ず、よう、まい
62	しめ		55
74	せ		55
83	そうだろ		3
92	そうでしょ		3
95	たろ		3
99	だろ		3
103	だろ		3
109	たかろ		3
115	たがら		10
116	たがろ		3
132	であろ		3
141	でしょ		3
145	なかろ		3
159	べから	×	ず
165	ませ	×	ん (打消・終)
166	なさいましょ、ましょ、いらっしゃいましょ、くだ (下) さいましょ、ございましょ		3
173	みたいだろ		3
180	みたいでしょ		3
185	ようだろ		3
192	ようでしょ		3
200	られ		21
217	れ		21

281	い		28
288	こ		426
310	なさら、いらっしゃら、くだ(下) さら、なら	×	ず、ない、れる
311	なさろ、いらっしゃろ、くだ(下) さろ、なろ		3
318	よら		310
319	よろ		3
326	あろ		3
333	おこ		3
334	おか		10
341	しまわ		310
342	しまお		3
349	み	×	ず、せる、ない、ぬ、させる、まい、よう

B.2 連用形

12	い(カ行五段イ音便)	×	た、たり、て、てる
13	い(ガ行五段イ音便)	×	だ、だり、で
14	-i(五段)	○	そうだ、そうです、た、たい、なさいます、ます、なさる、ながら、など、こそ、さえ、は、も、たり、つつ、て、すぎ、がち、たい、たがる、てる
15	ん(撥音便)		13
16	っ(促音便)		12
427	語幹(上一、下一)		14
31	し(サ変)		14
4	かっ(形容)	×	た、たり
5	く(形容)	○	など、こそ、さえ、でも、は、も、て
304	で(形動)	○	など、こそ、さえ、でも、は、も
305	に(形動)		304
306	だっ(形動)		4
22	れ(五段)		14
53	ごとく		5
56	させ		14
63	しめ		14
68	ず	○	×
75	し		14
81	そうで(伝聞)	○	×
84	そうで(様態)	○	×
85	そうに		304
86	そうだっ		4
90	そうでし		12
93	そうでし		12
104	で		304
105	だっ		4
110	たく		5
111	たかっ		4
117	たがり		14
118	たがっ		4
133	であり		14
134	であっ		4
142	でし		12
146	なく		5
147	なかっ		4
153	ん(打消)	×	そうだ(伝聞)
160	べく(必然)	○	×

167	なさいまし、まし、いらっしゃいまし、くだ(下)さいまし、ございまし		12
312	なさり、いらっしゃり、くだ(下)さり、なり		12
313	なさっ、いらっしゃっ、くだ(下)さっ、なっ		4
174	みたいで		304
175	みたいに		304
176	みたいだっ		4
181	みたいでし		12
186	ようで		304
187	ように		304
188	ようだっ		4
193	ようでし		12
195	らしく		5
196	らしかっ		4
197	らしゅう	○	×
201	られ		14
208	れ		14
282	い	×	14
289	き	×	14
320	より	○	そうだ、そうです、たい、なさいます、ます、ながら、など、こそ、さえ、は、も、つつ
321	よっ		12
327	あり		320
328	あっ		12
335	おき		320
336	おい		12
343	しまっ		12
344	しまい		320
350	み		14

B.3 終止形

17	-u(五段)	○	そうだ、そうです(伝聞)、なら(断定「だ」仮定)、だろう、であろう、でしょう、べし、まい、らしい、か、など、なんて、が、から(接助)、けど、けれど、し、と、なり、な(終助詞)、ね、よ(間投)、さ、の(準体)
23	る(上一、下一)		17
32	する(サ変)		17
6	い(形容)	○	ながら、そうだ、そうです(伝聞)、だろう、であろう、でしょう、らしい、か、など、が、から、けど、けれど、し、と(接助)、ね、よ、さ、の
307	だ(形動)	○	そうだ、そうです(伝聞)、など、が、から、けど、けれど、し、と、ね、よ
51	う(推量)	○	が、から、けど、けれど、し、と、ね、よ、か(推量)
57	させる		17
64	しめる		17
69	ず	○	と
76	せる		17
82	そうだ(伝聞)	○	が、から、けど、けれど、し、と、ね、よ
87	そうだ(様態)		82
91	そうです(伝聞)		82
96	た		6
100	だ(過去)		6
106	だ(断定)		307
112	たい		6
119	たがる		6
124	だろう		51

127	てる		6
135	である		6
139	であろう		51
140	でしょう		51
143	です		51
148	ない		6
154	ん (打消)		51
155	ぬ (打消)		51
161	べし	○	と
163	まい		51
168	なさいます、ます、いらっしゃいます、くだ (下) さいます、ございます		17
314	なさる、いらっしゃる、くだ (下) さる、なる		17
177	みたいだ		82
182	みたいです		51
183	よう		82
189	ようだ		82
194	ようです		51
198	らしい		82
202	られる		17
209	れる		17
283	いる		17
290	くる		17
322	よる		17
329	ある		17
337	おく		17
345	しまう		17
351	みる		17

B.4 連体形

18	-u(五段)	○	形式名詞、みたいだ、ようだ (比況)、ようです、みたいです、が、と、に、の、より (格助)、くらい、ぐらい、だけ、ばかり、ほど、まで (副助)、こそ、さえ、しか、でも、は、も (係助詞)、くせに
24	る (上一、下一)		18
33	する (サ変)		18
7	い (形容)	○	形式名詞、みたいだ、みたいです、ようだ、ようです、が、と、に、の、より、くらい、ぐらい、だけ、ばかり
308	な (形動)	○	形式名詞、みたいだ、ようだ (比況)、ようです、みたいです、が、と、に、の、より (格助)、くらい、ぐらい、だけ、ばかり、ほど、まで (副助)
52	う	○	形式名詞、が、と、に、の
54	ごとき		52
58	させる		18
65	しめる		18
70	ざる	○	18 + を
77	せる		18
88	そうな	○	形式名詞
98	た		7
101	だ		7
107	な	×	ので、のに、ゆえ、の
113	たい		7
120	たがる		18
122	たる		18
128	てる		18
136	である		18

言語情報を活用する
知的文字認識に関する研究

～ 修士論文本審査用ＯＨＰ資料 ～

東北大学大学院工学研究科
電気及通信工学専攻

秋山 秀三

1. 序論

●文字認識

文書入力自動化に向けて認識の高精度化への研究

1文字の文字領域切り出し
↓
個別文字認識

日本語文書認識の問題点

- 多数の類似文字
  (「大」「犬」、「べ」「ぺ」)
- かすれ、つぶれなど低品質文字
- 分離文字 (「化」、「イ」＋「ヒ」)

1文字毎の文字パターンからだけでは正確な認識が困難

● 文字認識後処理

日本語に関する言語情報を活用することにより文字認識でのあやまりを修正

- 適用する文法と対象文書
- 未登録語への対処
- 切り出しミス

● 目的

- 言語情報を活用した文字認識後処理システムの構築
- 文字切り出しの段階から言語情報を活用し文書認識の高精度化をはかる

● 構成

1. 序論
2. 個別文字認識
3. 文字認識後処理システム
4. 文字切り出しを含めた文字認識への言語情報の活用
5. 結論

2. 個別文字認識

※ 方向線素特徴量によるパターン認識

入力パターン

↓

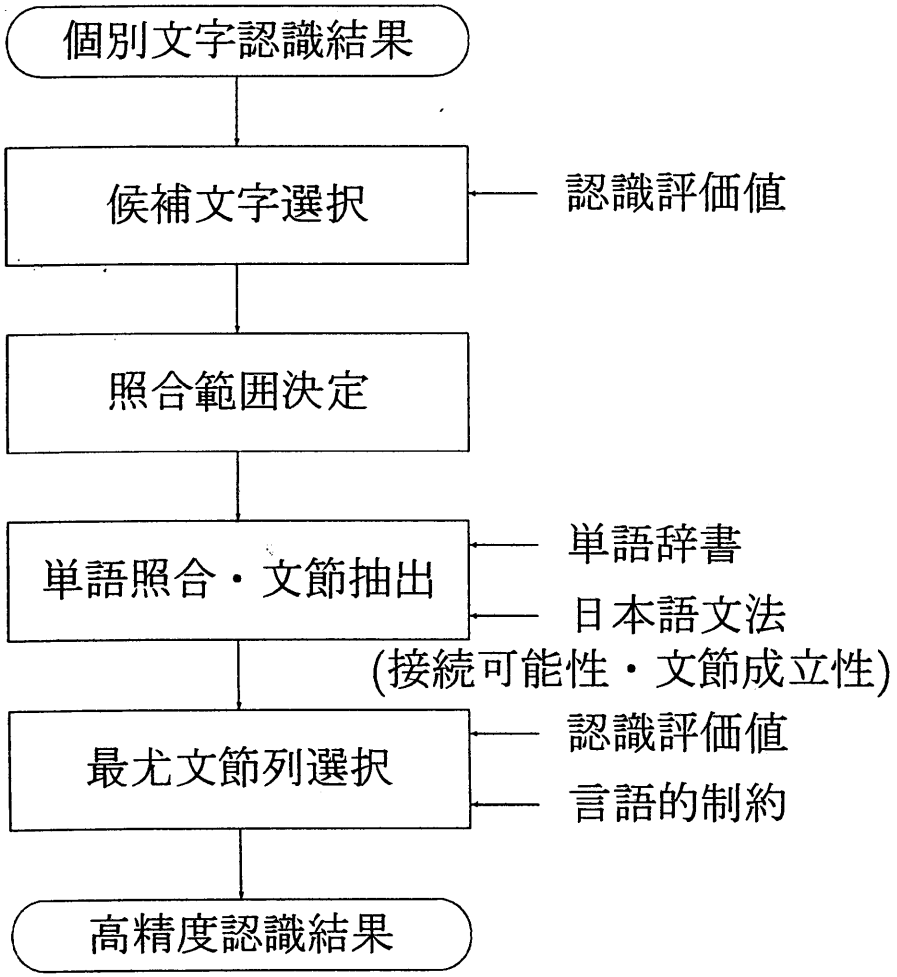
個別文字認識

↓

	候補文字	認識評価値
1位候補	車	7379
2位候補	革	14562
3位候補	草	15751
4位候補	軍	17202
5位候補	牽	17306
6位候補	奉	17687
7位候補	卓	18403
8位候補	華	19463
9位候補	東	19778
10位候補	章	19840
	:	

3. 文字認識後処理システム

● システムの構成



● 候補文字選択

1位：与野党ともにさっぱりやる気をみせない。
2位：兵軒克どｔＫきつばリゃろ氣さ　ぜだぃｏ
3位：写懸覚≧ちｋざりぽつヤね烹な彩廿たレ０
4位：存艱充をおほ古・ぼ●ャ３笥ちゑ壮む～Ｏ
5位：身貯兌上ぉド台勺は０午う汽忘影七を‥○

↓

与野党ともにさっぱりやる気をみせない。
克　ｔ　き　ばリ　ろ　　ぜ
ち　　ぽつ

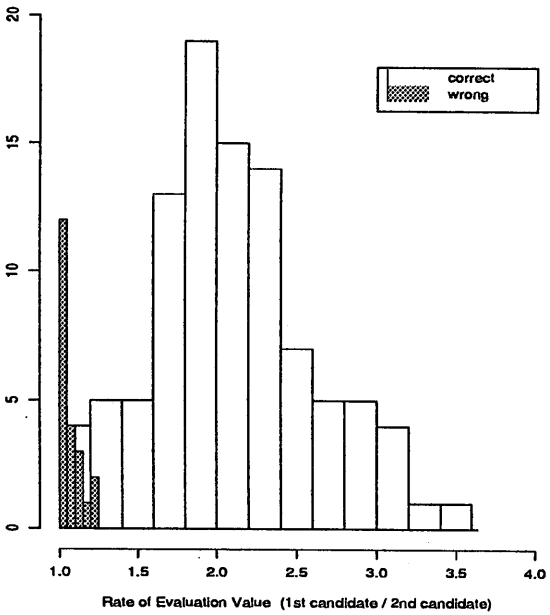
- 後処理の処理量軽減
- 誤った文節の抽出を防ぐ

※ 候補選択の条件
1位候補と2位候補の評価値の比をとり
1位候補が正解かどうかを判断

○：ふ →	ふ 8038	× ：ぶ →	ふ 11306
	ぶ 10480		ぶ 11387
	ぷ 13386		ぷ 13465
評価値の比	1.30		1.01

１位，２位候補の評価値の比の分布例 「諸」

※ 文字毎に違った傾向



下位候補も参照 ← 閾値 → １位候補を正解

↓

１位候補が不正解の場合の正解候補の
評価値の統計より候補を絞り込む

● 単語照合

候補文字

地球珠琲味味 を 石 器 時 代 以 前 に
万 嵐 ツ ッ ソ

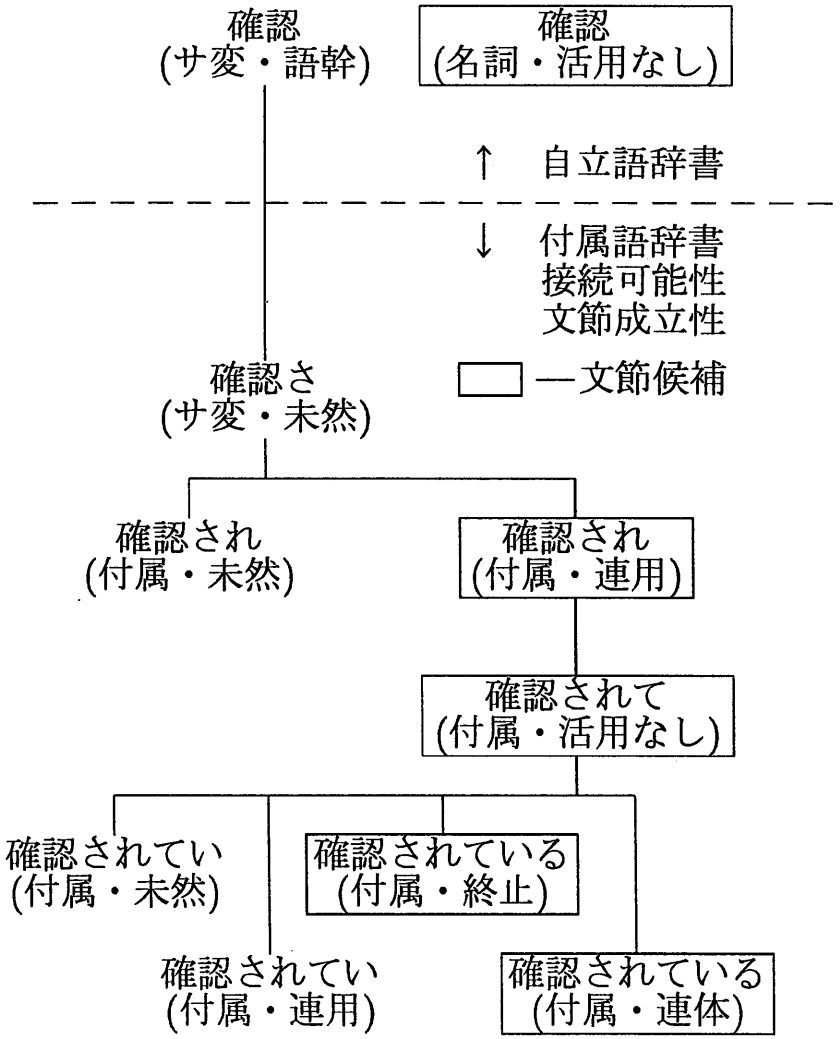
自立語辞書

地 1
地べた 1
地位 1
：
地球 7
地球儀 1
：
地味 1
地味好み 21

単語のつづり・文法情報

● 文節生成

入力：「確認されている」



● 文節抽出

与野党ともにさっぱりやる気をみせない。
つぱ　やろ　　ぜ　い
ば

与野党　名詞-(名詞.活無)
与野党と　名詞-(付属.活無)
与野党とも　名詞-(付属.活無)
ともに　固名-(付属.活無)
もに　名詞-(付属.活無)
さっぱり　副詞-(副詞.活無)
さつ　名詞-(名詞.活無)
つば　名詞-(名詞.活無)
ばり　名詞-(名詞.活無)
ばりや　名詞-(付属.活無)
やる　ラ五-(ラ五.終止)
やる　ラ五-(ラ五.連体)
やる　ラ五-(ラ五.連体)
気　名詞-(名詞.活無)
気を　名詞-(付属.活無)
みせ　名詞-(名詞.活無)
みせ　下一-(下一.連用)
みせない。　下一-(読点.活無)
せな　名詞-(名詞.活無)
ない。　形容-(読点.活無)

※ 文節の持つ情報

文字数、品詞、活用型、活用形、認識評価値、付属語の数、自立語部の文字数 など

● 未登録語処理

※ 誤認識例

イラクと　　ゴルバチョフ
タ　　パテ

同一文字種が続く部分
(カタカナ、アルファベット、数字)
↓
登録語と同様に付属語の接続を考慮
(数字列 ＋ 個、年などの助数詞)

イラクと
タ

イラク	カナ-(カナ.活無)
イラクと	カナ-(付属.活無)
イラ	カナ-(カナ.活無)
ラク	カナ-(カナ.活無)
ラクと	カナ-(付属.活無)
タ	名詞-(名詞.活無)
タと	名詞-(付属.活無)

※ 文節数最小法 (形態素解析の一手法)

文字列を構成する文節の総数が最小になる解釈にしたがい区分する方法
→ 最小文節数 ＋ 1 までに正しい解析が存在する

にほんのれきしをまなぶ

○ にほんの／れきしを／まなぶ
× にほん／のれ／きしを／まなぶ

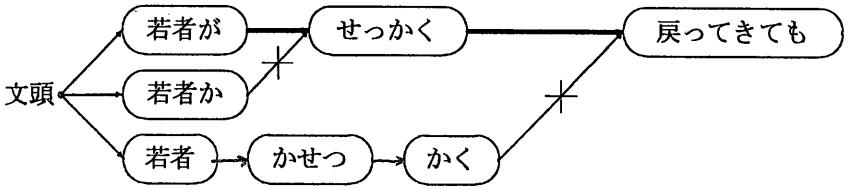
● 文節列選択

文節列のコストを求める

$$C(N,P,E) = \alpha \cdot N + \beta \cdot P + \gamma \cdot E$$

N: 文節数
P: 未登録語、１文字語の単語数
E: 認識評価値
(構成文字の１位候補との距離比の合計)
α, β, γ : 定数

若者がせっかく戻ってきても
か つが



● 実験

- 入力データ
  - 新聞社説 50 編 (1 編 1000～2000 文字)
  - レーザープリンタより出力
  - １文字毎切り出されたイメージとする
- 特徴量
  - 方向線素特徴量

● 使用する日本語知識

- 自立語辞書 (品詞、活用型)
  - 九州芸工大自立語辞書
    KID-J94 より構成
  - 約 15 万単語
  - 読みより作成したかな辞書を併用
- 付属語辞書
  - 助詞、助動詞、活用語尾、形式名詞
- 日本語文法
  - 単語間接続可能性
  - 文節成立性

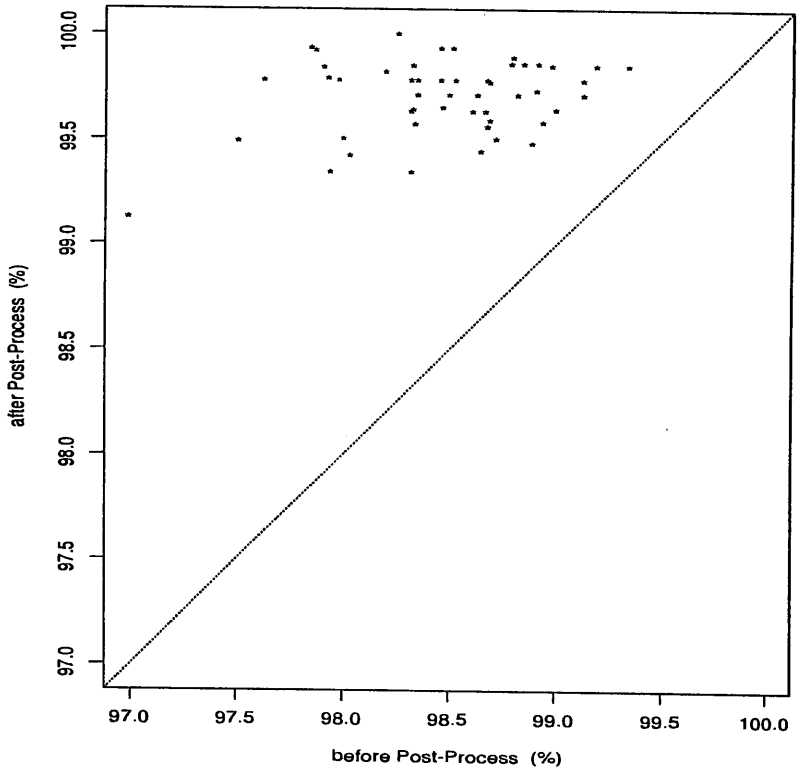
● 結果

処理前認識率	98.46% (誤り：1126/73199)
処理後認識率	99.70% (誤り：219/73199)
訂正できた文字数	986 (誤認識の 87.6%)
訂正できなかった文字数	140 (誤認識の 12.4%)
誤訂正した文字数	79

※ 未登録語処理の効果

処理後認識率	99.56 %→99.70 %
訂正できた文字数	939 → 986
訂正できなかった文字数	187 → 140
誤訂正した文字数	131 → 79

後処理の効果



文書例

衆議院の新しい議長に坂田道太氏が選ばれた。戦後第一回の総選挙以来連続十六回当選の長老だが、年齢は六十八歳と比較的若い。前議長が健康状態を危ぶまれたあげく退陣せざるを得なかっただけに、前尾繁三郎氏以来の六十歳代議長を歓迎したい。　議長として、まず頭に入れておいてほしいのは、わが国の議会政治がいま相当危うい状況にあるという認識である。一言でいえば「国会は国権の最高機関」という憲法の規定が空文化しつつあるということだ。　鈴木内閣の下で第二次臨調（臨時行政調査会）が発足したとき、国会と政党制の外に実質的な政策決定機関が出現したことに対し、「議会政治の権威失墜」「財界や官僚、労組が牛耳るニューコーポラティズム（新職能代表制）」といった論評があった。　最近の政治の状況は、そうした論評を一層裏付ける方向に進んでいる。臨調方式の政策決定システムは臨教審にも持ち込まれた。防衛費のＧＮＰ１％枠見直しへ先導役を務めようとする平和問題研究会など一連の首相の私的諮問機関の存在……。　こうした首相の政治手法の背後には、従来のような国会のあり方では、政治や行政の大きな改革はできっこない、という「総決算的」発想がうかがわれよう。　国会が非力だからといって、別のものにおきかえることには賛成できないが、国民の側からみても、「何もできない国会」というイメージが定着しつつあるのは事実だ。それには三つの原因がある。

第一は、定数是正を怠っていることだ。一昨年十一月の最高裁の違憲状態判決にもかかわらず、法案も審議せず一年を空費し、この夏再び違憲判決が出ることは確実だ。　第二は田中問題への対応である。一昨年十月、田中被告に一審有罪判決があったにもかかわらず、倫理審査会の設置に向けて与野党の話し合いは多少進展したものの、院としての自浄の努力は何もない。　第三は、国会が本来の政策論議の場としての役割を十分に果たしていないことである。その背景には、与野党議員の多くが選挙区への利益誘導に血道をあげざるを得ないという問題があることは明らかだが、国民的論議の必要な問題について国会の指導性が期待できないとしたら、いずれは国会の存在理由そのものが問われてしまうだろう。　すでに、参議院については社会経済国民会議や日経連が参院廃止論を唱えている。衆議院としても事態を深刻に受け止め、国会の活性化に立ち上がらなければならない。　そのための一つの手掛かりは、本会議をもっと真剣な討論の場とすることだ。質問の事前通告や想定問題集などはやめ、質問者も答弁側も自力で渡りあったらよい。　もちろん、議長という立場はリーダーシップをふるうというよりもコンセンサスに乗るポストであり、何かにつけ与野党の合意を必要とする慣行ができている。しかし、国会の現状を厳しく認識すればおのずから発言し、行動する余地はあるはずだ。　坂田氏は自民党内で知性派政策マンとみられてきた。大学紛争当時の文相や三木内閣の防衛庁長官としての手腕を評価する向きは少なくない。文相や防衛庁長官当時、政策実現のために精力的な世論への働きかけを行うなど、世論重視の政治手法をとってきた。国会のかじ取りにあたっても、この持ち味を生かしたらどうか。　衆院議長は５１１人の代議士の代表にとどまらず、全国民を代表するという気概を忘れないでほしい。

1367 文字

認識率　　98.68 % → 99.71 %

● 正しく訂正できない場合

- 未登録語
  - －ニクソン → ニクノン
  - －対策づくり、問題がらみの
    → つくり、からみに誤訂正
- 表記のゆれ
  「駆け足」→「かけ足」
  「破綻」→「破たん」
- 記号　「。」「・」　、「―」「…」
- 候補外文字
  → 激しいかすれ、ノイズによるもの
  → 候補文字選択の際に絞り込みすぎ

※ 連濁

２つの語が結びついて複合語になるとき、あとにくる語の最初の音が濁音になること。(例：「あお–ぞら」)

- ２つの語の結び付きが強い場合に起こる。
- 必ずしも規則的なものではなく、連濁を起こす語とそうでない語がある。
- 傾向
  - 前にくる語の末尾の音が [ン] の場合起きる。(「本箱」・「三階」)
  - 動詞と動詞とが結びついて複合動詞となっている場合には少ない。(「打ちつける」・「流れ込む」)。
  - 後にくる語にすでに濁音が含まれているときは連濁しない。(「白壁」・「大風」)。

4. 文字切り出しを含めた文字認識への言語情報の応用

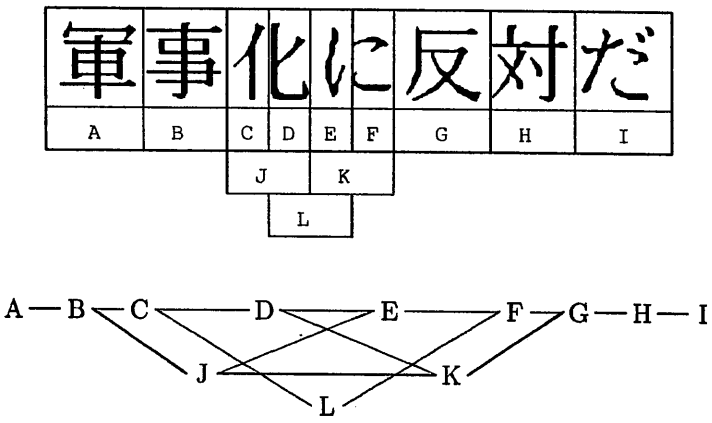
● 言語情報を用いない場合

イメージ入力
↓
ヒストグラムにより行抽出
↓
文字領域抽出
↓
分離文字統合

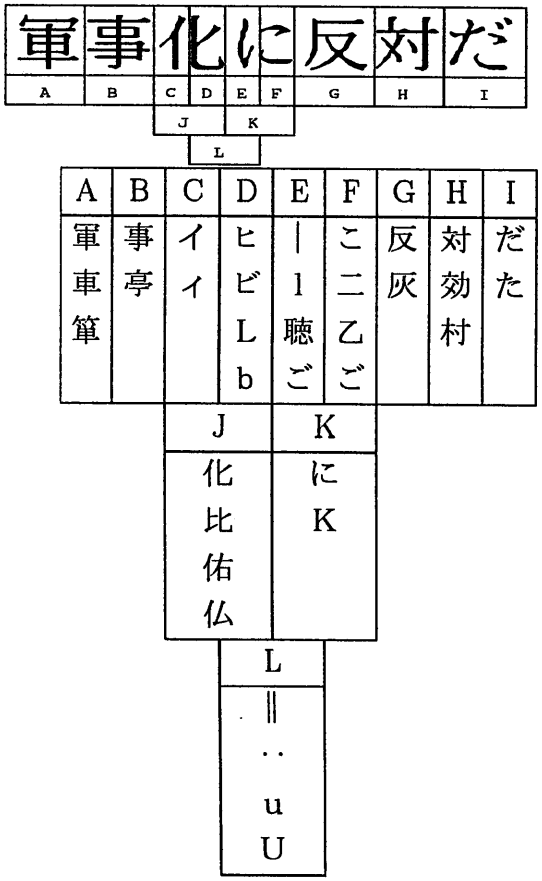
以下の E_path を最小にするパスを選択する。

$$E_{Path} = \sum_{i \in Path} V_i \cdot W_i \cdot K^{n_i}$$

領域 i の W_i:横幅, n_i :分離数, V_i:第一候補の評価値
(K は定数)



※ 言語情報を用いる場合



軍事・軍事化・軍事化に
仏に・ごご・二反・反対・反対だ

以上の抽出された文節候補の E_{Path} を計算し、それに文節数最小法他、言語的制約を与えることにより文節列を決定する。

● 実験

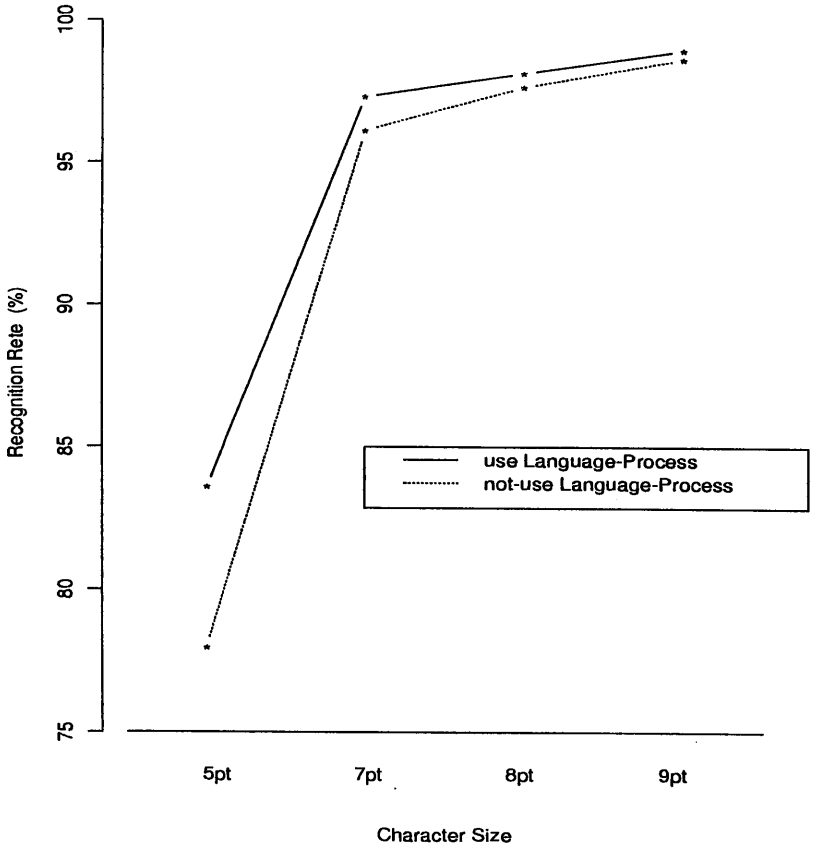
- 入力データ
  - 新聞の社説 20 編
    (1 編 1000〜2000 文字)
  - 5, 7, 8, 9 pt で TEX により出力した文書
- 特徴量
  - 方向線素特徴量

● 結果

size	言語処理なし → 言語処理あり	
5pt	77.98 % →	83.58 %
7pt	96.15 % →	97.33 %
8pt	97.68 % →	98.17 %
9pt	98.67 % →	98.98 %

ノ、レ → ル
イ、ヒ → 化

言語処理による認識の高精度化



5. 結論

● まとめ

- 言語情報を用いた文字認識後処理システムを構築し、認識の高精度化をはかることができた。
- 切り出しの段階から言語情報を適用し、実際の文書認識へ適用できるシステムを構築した。

● 今後の課題

- 正解文字が候補として挙がらなかった場合の対処
  - 候補文字補完
  - 辞書照合方法の改善
- 切り出しと融合させた場合の評価方法の改善
- 単語の登録、文法規則の追加などの容易な「使い易い」システム

144	です		107
149	ない		7
151	なる		7
156	ん		18
157	ぬ		52
162	べき	○	こと、とき、ところ、もの、だ、だろう、であろう、でしょう、です、らしい
164	まい		52
169	なさいます、ます、いらっしゃいます、くだ(下)さいます、ございます		52
315	なさる、いらっしゃる、くだ(下)さる、なる		7
178	みたいな		308
184	よう		52
190	ような		308
199	らしい		52
203	られる		18
210	れる		18
284	いる		18
291	くる		18
323	よる		18
330	ある		18
338	おく		18
346	しまう		18
352	みる		18

B.5 仮定形

19	-e(五段)	×	ば、ど
25	れ(上一、下一)		19
34	すれ(サ変)		19
8	けれ(形容)		19
309	なら(形動)	×	ば
59	させれ		19
66	しめれ		19
71	ざれ		19
72	ず		309
73	ずん		309
78	せれ		19
89	そうなら		309
98	たら		309
102	だら		309
108	なら		309
114	たけれ		19
121	たがれ		19
129	てれ		19
137	であれ		19
150	なけれ		19
158	ね		19
170	なさいますれ、ますれ、いらっしゃいますれ、くだ(下)さいませ、ございますれ		19
316	なされ、いらっしゃれ、くだ(下)され、なれ		19
179	みたいなら		309
191	ようなら		309

204	られれ		19
211	れれ		19
285	いれ		19
292	くれ		19
324	よれ		19
331	あれ		19
339	おけ		19
347	しまえ		19
353	みれ		19

B.6 命令形

20	-o(五段)	○	と (格助)、よ (間投)
26	ろ (上一、下一)		20
27	よ (上一、下一)		と
35	しろ (サ変)		20
36	せよ (サ変)		27
60	させろ		20
61	させよ		27
67	しめよ		27
79	せろ		20
80	せよ		27
123	たれ		20
130	てろ		20
131	てよ		27
138	であれ		20
152	なれ		20
171	なさい、なさいませ、ませ、いらっしゃいませ、くだ (下) さいませ、ございませ		20
172	なさいまし、まし、いらっしゃいまし、くだ (下) さいまし、ございまし		27
317	なされ、いらっしゃれ、くだ (下) され、なれ		20
205	られろ		20
206	られよ		27
212	れろ		20
213	れよ		27
286	いらっしゃい、いろ		20
287	いよ		27
293	くだ (下) さい、くれ、こい		20
325	よれ		20
332	あれ		20
340	おけ		20
348	しまえ		20
354	みろ		20
355	みよ		27

B.7 語幹その他

401	ワ行五段活用	×	活用語尾
402	カ行五段活用	×	活用語尾
403	サ行五段活用	×	活用語尾
404	タ行五段活用	×	活用語尾
405	ナ行五段活用	×	活用語尾
406	マ行五段活用	×	活用語尾
407	ラ行五段活用	×	活用語尾
408	ガ行五段活用	×	活用語尾
409	バ行五段活用	×	活用語尾
410	上一段活用	×	活用語尾
411	下一段活用	×	活用語尾
412	形容詞	×	活用語尾、すぎる
413	形容動詞	×	である、であろう、でしょう、です、らしい、さ、活用語尾、すぎる
414	サ変活用	×	活用語尾、(べき用) す、できる
415	ザ変活用	×	活用語尾、(べき用) ず
416	名詞	○	だ、である、であろう、でしょう、です、らしい、格助詞、副助詞、係助詞 (1 に同じ)
417	連体詞	○	×
418	接続詞	○	×
419	感動詞	○	と (接続助詞)
420	副詞	○	×
423	くる	×	くる、来る
424	する	×	する、(べき用) す
425	ラ変活用	×	活用語尾
428	カタカナ	○	名詞に同じ
429	アルファベット	○	名詞に同じ
430	数字列	○	×
431	読点	○	×
432	句点	○	×
433	とじ括弧	○	名詞に同じ
434	数詞	○	名詞に同じ
435	固有名詞	○	名詞に同じ
436	記号	○	×
437	単語候補なし	○	×
438	文頭	×	名詞に同じ
439	「-の」が付く形容動詞	×	413 +の
500	ついて、つきまして、関して、関しまして	○	の、を、くらい、すら、だけ、だの、なぞ、など、のみ、ばかり、係助詞、だ (断定)、だろう、である、であろう、でしょう、です、みたいだ、らしい
501	(べき用) す、ず	×	べき