利用過程で得られる言語情報を活用する音声言語処理システム

森　信介[†1]　前田浩邦[†2]

学習に基く音声言語処理は一定の成果を上げるているが、多くの研究は、データの集収方法を含めた全体戦略を欠いている。本発表では、有用な音声言語処理システムを一般ユーザーに提供し、その利用過程で得られる音声言語に関する情報を活用し更なる精度向上を図るという新しい音声言語処理パラダイムを提案する。このパラダイムに沿った具体例として、未知語も変換候補として列挙する仮名漢字変換システムを挙げる。これを一般ユーザーに供することにより、様々な分野において、単語境界と読み情報が付与された文の断片が得られる。このデータから学習することで、単語分割や読み推定、あるいは音声認識などの音声言語処理の精度向上が自動的に実現される。

Speech and Language Processing System Exploiting Information Given by Users

Shinsuke MORI [†1] and Hirokuni MAETA[†2]

Data-driven speech and language processing (SLP) systems proved a success to some extent. Many researches, however, do not mention an entire strategy including data preparation. In this paper, we propose a novel paradigm in which useful SLP systems are provided to users and improved by the data obtained through their use. As an example, we propose an input method for Japanese which can enumerate substrings contained in a raw text as conversion candidates. By providing this SLP system to public users we are able to obtain sentence fragments annotated with word boundary information and pronunciation. The data allow us to improve, without any cost, the accuracies of SLP systems such as word segmentation, pronunciation estimation, speech recognition, etc.

1. はじめに

学習に基く音声言語処理は一定の成果を上げるているが、多くの研究は、データの集収方法を含めた全体戦略を欠いている。実際、ある問題を設定し、コストを投じてその問題の入力と出力の組の例を作成し、機械学習を適用して一定の精度で入力から出力を推定した結果の報告は枚挙に暇がない。規則による方法からこのような方法への移行により、アルゴリズムとデータが分離された。その結果、言語処理システムの構築において、ある程度の分業や並行作業が可能になり、さらに分野適応も容易になった。しかしながら、データをどのように集収あるいは作成するのかという課題が依然として残されたままである。

この言語資源の収集を、有用な音声言語処理システムの一般への提供により解決するというのが本論文の提案である。音声言語処理システムの利用過程で得られる情報は、利用結果から得られる情報よりもはるかに多い。実際、本論文で具体例とする仮名漢字変換システムは、未知語も変換候補として列挙することが可能で、この利用過程の情報には、未知語の単語境界やその読みが含まれる。これらは、完成された文には含まれない情報である。

インターネットが十分に普及した現在、ありとあらゆる分野の大量のテキストが利用可能であるとの誤解があるが、実際には、企業内での活動報告書やカルテなど、インターネットではアクセスできない分野の文も多数ある。このような分野での音声言語処理にも高い需要があり、このような需要に短い開発期間で応えるためにも、活動報告書やカルテの執筆過程において副次的に産生される情報を蓄積しておくことは、公開・非公開にかかわらず非常に重要である。つまり、音声言語処理が、個人や団体が日々行う言語活動を補助することは、二重の意味で重要である。

このような考察に基づいて、本論文では、有用な音声言語処理システムをまず一般ユーザーに提供し、その利用過程で得られる音声言語に関する情報を活用し更なる精度向上を図るという新しい音声言語処理パラダイムを提案する。このパラダイムに沿った具体例として、未知語も変換候補として列挙する仮名漢字変換システムを挙げる。これを一般ユーザーの利用に供することで単語境界と読み情報が付与された文の断片が得られる。このデータを用いることで、仮名漢字変換や音声認識などの生成系、あるいは単語分割や読み推定など解

†1 京都大学 学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University
†2 京都大学 理学部
Faculty of Science, Kyoto University

析系の音声言語処理の精度向上がコストなしで実現できる可能性を示す。

2. 確率的言語モデルによる仮名漢字変換

確率的言語モデルによる仮名漢字変換[1)] は、キーボードから直接入力可能な記号 $\mathcal{Y}$ の正閉包 $\boldsymbol{y} \in \mathcal{Y}^+$ を入力として、日本語の文字 $\mathcal{X}$ の正閉包である変換候補 $(\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \cdots)$ を確率値 $P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})$ の降順に提示する。この第 1 因子 $P(\boldsymbol{y}|\boldsymbol{x})$ は仮名漢字モデルであり、日本語文 $\boldsymbol{x}$ が与えられたときのキーボードからの入力の記号列 (読み) の確率を表す。第 2 因子 $P(\boldsymbol{x})$ は、確率的言語モデルであり、日本語文字列 $\boldsymbol{x}$ の出現確率を表す。

確率的言語モデルとしてはマルコフ性を仮定する単語 n-gram モデルが、仮名漢字モデルとしては単語毎の独立性を仮定るすモデルが一般的に用いられる。以下では、これらを順に説明する。

2.1 単語 n-gram モデル

単語 n-gram モデルは、文を単語列 $\boldsymbol{w}_1^h = w_1 w_2 \cdots w_h$ と見なし、これらを文頭から順に予測する。

$$M_{w,n}(\boldsymbol{w}) = \prod_{i=1}^{h+1} P(w_i|\boldsymbol{w}_{i-n+1}^{i-1}) \tag{1}$$

この式の中の w_i $(i \leq 0)$ は、文頭に対応する特別な記号であり、w_{h+1} は、文末に対応する特別な記号である。完全な語彙を定義することは不可能であるから、未知語を表す特別な記号 UW を用意する。未知語の予測の際は、まず、単語 n-gram モデルにより UW を予測し、さらにその表記 (文字列)$\boldsymbol{x}_1^{h'}$ を以下の文字 n-gram モデルにより予測する。

$$M_{x,n}(\boldsymbol{x}_1^{h'}) = \prod_{i=1}^{h'+1} P(x_i|\boldsymbol{x}_{i-n+1}^{i-1}) \tag{2}$$

この式の中の x_i $(i \leq 0)$ は、語頭に対応する特別な記号であり、$x_{h'+1}$ は、語末に対応する特別な記号である。したがって、未知語は以下のように予測される。

$$P(w_i|\boldsymbol{w}_{i-n+1}^{i-1}) = M_{x,n}(w_i)P(\mathtt{UW}|\boldsymbol{w}_{i-n+1}^{i-1})$$

2.2 仮名漢字モデル

仮名漢字モデル $P(\boldsymbol{y}|\boldsymbol{x})$ は、入力記号列と日本語文との確率的対応関係を記述する。あらゆる可能な日本語文に対応する入力記号列の確率を推定することは不可能であるから、日本語文を単語に分割し、単語と入力記号列との対応関係がそれぞれ独立であると仮定する。このとき、単語列 $\boldsymbol{w}$ が与えられたときの入力記号列 $\boldsymbol{y}$ の仮名漢字モデル M_{kk} による出現確率は以下の式で表される。

$$M_{kk}(\boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^{h} P(\boldsymbol{y}_i|w_i) \tag{3}$$

ここで、入力記号部分列 $\boldsymbol{y}_i$ は単語 w_i に対応する入力記号列であり、以下の条件を満たす。

$$\boldsymbol{y} = \boldsymbol{y}_1 \boldsymbol{y}_2 \cdots \boldsymbol{y}_h$$

確率 $P(\boldsymbol{y}_i|w_i)$ の値は、単語ごとに入力記号列が付与されたコーパスから以下の式を用いて最尤推定することで得られる。

$$P(\boldsymbol{y}_i|w_i) = \frac{f(\boldsymbol{y}_i,\ w_i)}{f(w_i)} \tag{4}$$

この式中の $f(e)$ は、事象 e のコーパスにおける頻度を表す。

未知語に対する変換モデルは提案されておらず、文献 1) の仮名漢字変換器は単に入力記号列の平仮名部分を片仮名にして返す[⋆1]。これは、確率的言語モデルの文字集合 $\mathcal{X}$ 上の未知語モデル $M_{x,n}(\boldsymbol{x})$ を入力記号集合 $\mathcal{Y}$ 上の未知語モデル $M_{y,n}(\boldsymbol{y})$ に置き換えることで実現される。

2.3 確率的言語モデルによる仮名漢字変換

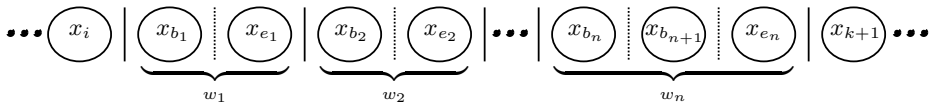
上述の単語 n-gram モデルと仮名漢字モデルを融合するにあたり、単語の定義が両モデルで一致している必要がある。この条件の下、確率的言語モデルによる仮名漢字変換器は、変換候補を以下の値の順に列挙する。

$$P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x}) = \prod_{i=1}^{h} P(\boldsymbol{y}_i|w_i)P(w_i)$$

$$P(\boldsymbol{y}_i|w_i)P(w_i) = \begin{cases} P(w_i|\boldsymbol{w}_{i-n+1}^{i-1})P(\boldsymbol{y}_i|w_i) & \text{if } w_i \in \mathcal{W} \\ P(\mathtt{UW}|\boldsymbol{w}_{i-n+1}^{i-1})M_{y,n}(\boldsymbol{y}_i) & \text{if } w_i \notin \mathcal{W} \end{cases}$$

ここで $\mathcal{W}$ は確率的言語モデルの語彙を表す。

⋆1 文献 1) によると、約 33.0%の未知語が片仮名列とすれば正しい変換となる。



$$f_r(w_1^n) = P_i(1-P_{b_1})P_{e_1}(1-P_{b_2})P_{e_2}\cdots(1-P_{b_n})(1-P_{b_{n+1}})P_{e_n}$$

図 1 確率的単語分割コーパスにおける単語 n-gram 頻度

3. テキストの部分文字列の利用

前節の仮名漢字変換器は、学習コーパスに出現し、対応する入力記号列が与えられている表記 (単語) のみが変換候補になる。この制限を取り払うために、サブワード モデルによる候補の列挙と確率的単語分割コーパスを用いた文脈の記述により語彙をテキストコーパスの部分文字列全てに拡張する方法が提案されている[2)] 。この拡張により、閲覧したウェブページや受け取ったメールなどに含まれる単語を変換候補として列挙することが可能となる。

3.1 サブワード モデル

文字を単位とするサブワード モデルは、まず、ある表記 $w = x_1x_2\cdots x_m$ に対応する入力記号列を各文字 x_i の入力記号列 y_i の連接とし、次に、その出現確率 $P(\boldsymbol{y}|w)$ を各文字に対応する入力記号列が一様に出現すると仮定して、以下のように計算する。

$$P(\boldsymbol{y}|w) = P(\boldsymbol{y}|x_1x_2\cdots x_m) = \prod_{i=1}^{m}\frac{1}{|\boldsymbol{\mathcal{Y}}_{x_i}|} \tag{5}$$

ここで、$\boldsymbol{\mathcal{Y}}_x$ は文字 x に対応する可能な入力記号列の集合であり、単漢字辞書を検索することで得られる。例えば、$\boldsymbol{\mathcal{Y}}_{日} = \{\text{か}, \text{じつ}, \text{にち}, \text{にっ}, \text{ひ}, \text{び}\}, \boldsymbol{\mathcal{Y}}_{テ} = \{\text{て}\}, \boldsymbol{\mathcal{Y}}_{レ} = \{\text{れ}\}$ であり

$$P(\text{にってれ} \mid \text{日テレ}) = \frac{1}{|\boldsymbol{\mathcal{Y}}_{日}|}\frac{1}{|\boldsymbol{\mathcal{Y}}_{テ}|}\frac{1}{|\boldsymbol{\mathcal{Y}}_{レ}|} = \frac{1}{6}\times\frac{1}{1}\times\frac{1}{1} = \frac{1}{6}$$

となる。

3.2 文脈の記述

サブワード モデルが列挙する単語候補を適切に選択するために、その文脈を適切に記述する必要がある。このためには、仮名漢字変換を適用する分野のコーパスから言語モデルを推定することが望ましい。これを実現するために、単語分割情報がないテキストコーパスから文献 3) の方法を用いて推定した単語 n-gram モデルを用いる。この方法では、テキストコーパスの各文字間 x_ix_{i+1} に単語境界確率 P_i を付与し、確率的単語分割コーパスとし、単語 n-gram 確率を期待頻度から計算する (図 1 参照)。単語境界確率は、単語分割済みコーパスから学習した最大エントロピー法による点推定器[4)] によって与えられる。

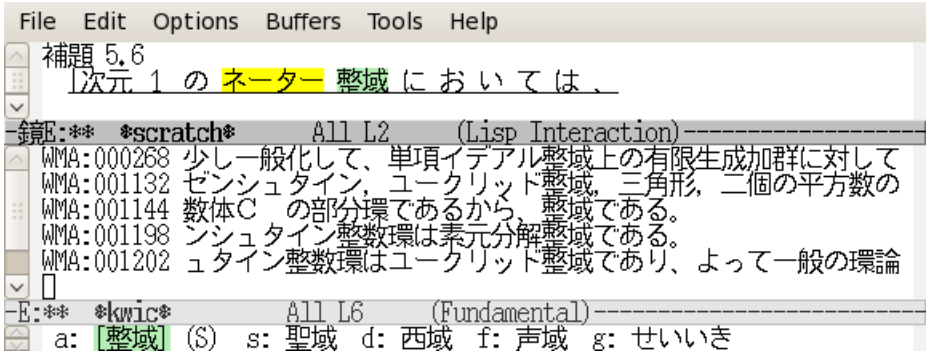


図 2 変換候補としてのテキストの部分文字列 (「整域」) の提示例 (Wikipedia の数学関連のページから作成したモデルを利用し文献 6) の 96 ページ 18 行目を入力)

テキストコーパスの部分文字列も候補にする仮名漢字変換においては、単語分割済みコーパスから推定した言語モデル P_g(式 (1)(2) 参照) とテキストコーパスから推定した言語モデル P_r を以下のように補間して用いる。

$$P(w_i|H_i) = \lambda_g P_g(w_i|H_i) + \lambda_r P_r(w_i|H_i) \tag{6}$$

この式中の H_i は、単語 w_i を予測する際の履歴である。λ_g と λ_r は補間係数であり、削除補間[5)] によって求める。

3.3 無限語彙の仮名漢字変換システム

テキストコーパスの部分文字列も候補にする仮名漢字変換は、式 (4) と式 (5) で表記の候補をその生成確率とともに列挙し、式 (6) で与えられる言語モデルの確率を掛けることで得られる文全体での生成確率の降順に変換候補を提示する。

各変換候補には、それが式 (4) によるか式 (5) によるかの情報を付与することができる。図 2 の例では、入力記号列「せいいき」がテキストの部分文字列「整域」に対応すると推測して提示されている。候補選択のための最も下の行では、テキストの部分文字列であることを示す「(S)」が付与されている。また、編集領域の直下には、当該候補文字列のテキスト

表 1 一般分野のコーパス

コーパス	単語分割	文数	単語数	文字数
現代日本語書き言葉均衡コーパス	人手	33,147	899,025	1,292,249
日経新聞の記事	人手	9,023	263,427	398,570
英語表現事典の例文	人手	14,754	189,915	254,436
産経新聞の記事	自動	8,335,449	–	60,065,893

での KWIC(Key Word In Context) が表示されている。

また、この仮名漢字変換システムは変換ログを自動で収集しており、各入力記号列に対して、最尤の単語列とユーザが選択した単語列を入力記号列と整列された状態で保持する。例えば、図 2 の例では、まず、最尤の単語列の入力記号列との整列結果として以下の情報が得られる。

次元/じげん 1/1 の/の ネーター/ねーたー 整域/せいいき に/に お/お い/い て/ては/は 、/、

さらに、ユーザーが最尤解から単語境界や表記を変更し確定すると、その結果が上記と同様の形式で保持される。

4. 言語資源の自動獲得実験

前節までで説明したテキストの部分文字列も候補として列挙する仮名漢字変換器を作成した。これを文献 4) の第 3 節以外の執筆に実際に利用し、変換ログ収集の実験を行った。本節では、この実験について詳述する。

4.1 実 験 条 件

まず、実験に用いた仮名漢字変換器の主要諸元について述べる。第 2 節で説明した一般分野の確率的言語モデルによる仮名漢字変換のコーパスは以下の通りである (表 1 参照)。

(1) 単語分割済みかつ入力記号列付与済みコーパス

このコーパスの各文は正しく単語に分割されており、各単語には入力記号列が付与されている。単語の定義は、概ね現代日本語書き言葉均衡コーパス[7] の短単位[8] に一致するが、活用語尾の分割などの変更がなされている。このコーパスを用いて、単語境界確率を推定するための最大エントロピー法による点推定器のパラメータを推定した。また、式 (4) の仮名漢字モデルのパラメータもこのコーパスを用いて推定した。

(2) 生コーパス

このコーパスの各文には何の情報も付与されておらず、自動的に単語に分割される。

表 2 学習コーパスに含まれていなかった単語と入力記号列の組

頻度	単語	入力記号列	頻度	単語	入力記号列
35	コーパス	こーぱす	2	連接	れんせつ
10	品詞	ひんし	2	例文	れいぶん
5	形態素	けいたいそ	2	語彙	ごい
4	エントロピー	えんとろぴー	1	未知語	みちご
3	文脈	ぶんみゃく	1	所与	しょよ
3	可読	かどく	1	下線	かせん

これは、各文字間に対して最大エントロピー法による点推定器が出力する単語境界確率を 0.5 と比較することで実行される。これは、言語モデルの統計情報の信頼性を高めるために用いられる。

一般分野の言語モデルには、表 1 のコーパスの全ての文から推定した単語 2-gram モデルを用いた。語彙は、これらのコーパスを 9 つに分割した結果 2 つ以上の部分コーパスに出現する 35,321 単語とした。対応する入力記号列との組の数は 36,901 であった[⋆1]。

適応に用いた生コーパスは、本論文の第 1 著者が 1996 年から 2008 年に渡って第 1 著者として執筆した国内研究会の論文のテキストである。全ては LaTeX を用いて書かれており、そのソースファイルに簡単なパターンマッチを適用することで文を認定した。その結果、2,157 文、109,782 文字からなる生コーパスが得られた[⋆2]。これを前述の単語境界確率推定器を用いて確率的単語分割コーパスとし、適応分野の単語 2-gram モデル (式 (6) の P_r) を推定した[⋆3]。式 (6) の補間係数は、一般分野の出現確率が最大となる値と適応分野の出現確率が最大となる値の平均とした。これは、仮名漢字変換器は、論文の執筆のみならず、メールなどの一般的な分野の作文にも用いる必要があるためである。なお、適応分野のモデルや補間係数の動的変更は今後の課題である。

4.2 得られた変換ログ

文献 4) の執筆の結果、2,395 の文断片が得られた。この一部を付録 A.1 に提示する。1,965 の候補断片がそのまま選択されており、第 1 候補確定率は 82.05%であった。確定結果の諸元は以下の通りである。

⋆1 公開版の語彙は 10 万語程度となる見込みである。

⋆2 公開版では、医療や法律やなど様々な分野の生コーパスや、World Wide Web のクロール結果を用いた複数の分野適応モデルを用意する予定である。

⋆3 実際には倍率 8 の擬似確率的単語分割[9] である。

平均単位数: 4,686
平均文字数: 7,774
平均入力記号数: 11,683
ここでいう「単位」とはユーザーが確定したときの単語境界情報によって決定される文字列であり、必ずしもコーパスの基準に一致しない。獲得された単位のうち、一般分野の言語モデルにおける未知語、すなわち、適応分野の生コーパスの部分文字列として変換候補に挙げられた単位は延べ 69 個であった。このすべてを頻度とともに表 2 に示す。これら 12 個の異なり単語候補のうち「形態素」と「未知語」以外の 10 個はコーパスの単語の基準に合致し、それぞれの入力記号列も適切である。したがって、適切な単語とその入力記号列が自動的に獲得されているといえる。「形態素」は「形態」と「素」の連接であり、「未知語」は「未知」と「語」の連接である。このように、単語の連接も獲得されている。なお、その入力記号列は適切である。

4.3 獲得された言語資源の利用

獲得された単語候補列とその入力記号列は、音声言語処理のための言語資源とみなすことができる。すなわち、ユーザーによる確定結果は、文断片でありかつ誤りを含む可能性があるが、単語分割済みかつ入力記号列付与済みコーパスとみなすことができる。これを用いて、仮名漢字変換の言語モデルや仮名漢字モデルを推定することが可能であり、変換精度の向上が見込まれる[10]。また、入力記号列は読みに非常に近く、ある程度の確度で読みに変換可能である[11]。したがって、音声認識の精度向上に応用可能である。例えば、大学教員による論文等執筆時の変換ログを用いた講義の音声認識や、ある企業の部署の構成員間でやりとりされるメールや資料の執筆時の変換ログを用いた会議書き起こしシステムの実現が考えられる。また、音声合成のための読み推定の精度向上にも貢献する。さらに、自動単語分割などの言語解析にも利用可能である。文献 4) では、内部の単語境界情報が欠落している複合語 (単語の連接) や単語列のリストを参照することで自動単語分割の精度向上が図れることが示されている。変換ログを用いてこれらを実証することが次の課題である。

5. おわりに

本論文では、自然言語処理におけるデータ収集の重要性を背景に、利用過程で得られる音声言語に関する情報を活用し更なる精度向上を図るという新しい音声言語処理パラダイムを提案した。このパラダイムに沿った具体例として、未知語も変換候補として列挙する仮名漢字変換システムを実際の利用に供し、単語境界と読み情報が付与された文の断片をコストなしで得られることを示した。得られた文の断片は言語資源として有用であり、仮名漢字変換や音声認識などの生成系、あるいは単語分割や読み推定など解析系の音声言語処理の精度向上がコストなしで実現できる可能性を示唆する。

一般化すれば、未知の言語現象に対しても頑健な音声言語処理システムを作成し、これを一般ユーザーに実際に利用してもらうことで、その音声言語処理システムのみならず他の音声言語処理システムの精度が向上するということである。すなわち、本論文で提案する枠組みは、人との広義の「対話」により言語を学習していく計算機を実現する第一歩になる。

謝　辞

コーパスと辞書の単位の変更において笹田鉄郎氏から多大なる尽力を頂きました。心より感謝致します。

参 考 文 献

1) 森　信介，土屋雅稔，山地　治，長尾　真：確率的モデルによる仮名漢字変換，情報処理学会論文誌， Vol.40, No.7, pp.2946–2953 (1999).
2) 森　信介：無限語彙の仮名漢字変換，情報処理学会論文誌， Vol.48, pp.3532–3540 (2007).
3) 森　信介，宅間大介，倉田岳人：確率的単語分割コーパスからの単語 N-gram 確率の計算，情報処理学会論文誌， Vol.48, pp.892–899 (2007).
4) 森　信介，小田裕樹：3 種類の辞書による自動単語分割の精度向上，情報処理学会研究報告，Vol.NL193 (2009).
5) Jelinek, F., Mercer, R.L. and Roukos, S.: Principles of Lexical Language Modeling for Speech Recognition, *Advances in Speech Signal Processing*, Dekker, chapter21, pp.651–699 (1991).
6) 堀田良之：可換環と体，岩波書店 (2006).
7) Maekawa, K.: Balanced Corpus of Contemporary Written Japanese, *Proceedings of the 6th Workshop on Asian Language Resources*, pp.101–102 (2008).
8) 小椋秀樹，小磯花絵，冨士池優美，原　裕：『現代日本語書き言葉均衡コーパス』形態論情報規程集，独立行政法人国立国語研究所 (2008).
9) 森　信介，小田裕樹：擬似確率的単語分割コーパスによる言語モデルの改良，自然言語処理， Vol.16, No.4 (2009).
10) 森　信介，小田裕樹：自動未知語獲得による仮名漢字変換システムの精度向上，言語処理学会年次大会 (2007).
11) 秋田祐哉，河原達也：話し言葉音声認識のための汎用的な統計的発音変動モデル，電子情報通信学会論文誌， Vol.J88-DII, No.9, pp.1780–1789 (2005).

付　　録

A.1 変換ログ

以下に、変換ログの一部を掲載する。各は変換ログは 3 行からなる。これらは順に、入力時の時間 (エポックタイムの上位 2 バイトと下位 2 バイト)、システムの最尤解の文断片、ユーザーの選択結果の文断片である。各文断片は、表記と入力記号列と出所の 3 つ組である。出所は、一般言語モデル (IN)、適応分野テキストの部分文字列 (IV)、片仮名 (K)、未知語 (UM) のいずれかである。

```
19088 61253
背景/はいけい/IN の/の/IN 元/もと/IN
背景/はいけい/IN の/の/IN 下/もと/IN

19088 61282
人手/ひとで/IN に/に/IN よ/よ/IN る/る/IN
人手/ひとで/IN に/に/IN よ/よ/IN る/る/IN

19088 61284
単語/たんご/IN 分割/ぶんかつ/IN
単語/たんご/IN 分割/ぶんかつ/IN

19088 61288
コーパス/こーぱす/IV
コーパス/こーぱす/IV

19088 61306
文/ぶん/IN
文/ぶん/IN

19088 61314
構築/こうちく/IN する/する/IN
構築/こうちく/IN する/する/IN

19088 61316
努力/どりょく/IN
努力/どりょく/IN

19088 61352
経験/けいけん/IN 的/てき/IN
経験/けいけん/IN 的/てき/IN

19088 61354
主砲/しゅほう/IN
手法/しゅほう/IN

19088 61365
自動/じどう/IN たん/たん/UM 五/ご/IN 分割/ぶんかつ/IN
自動/じどう/IN 単語/たんご/IN 分割/ぶんかつ/IN

19088 61367
システム/しすてむ/IN
システム/しすてむ/IN

19088 61370
構築/こうちく/IN
構築/こうちく/IN

19088 61374
試み/こころみ/IN
試み/こころみ/IN

19088 61385
多数/たすう/IN あ/あ/IN る/る/IN
多数/たすう/IN あ/あ/IN る/る/IN

19088 61400
最近/さいきん/IN
最近/さいきん/IN

19088 61408
自然/しぜん/IN 言語/げんご/IN 処理/しょり/IN
自然/しぜん/IN 言語/げんご/IN 処理/しょり/IN

19088 61417
さまざま/さまざま/IN な/な/IN
様々/さまざま/IN な/な/IN

19088 61431
分野/ぶんや/IN に/に/IN 適/てき/IN おう/おう/UM さ/さ/IN れ/れ/IN て/て/IN い
/い/IN る/る/IN
分野/ぶんや/IN に/に/IN テキ/てき/K おう/おう/UM さ/さ/IN れ/れ/IN て/て
/IN い/い/IN る/る/IN

19088 61435
適用/てきよう/IN さ/さ/IN れ/れ/IN て/て/IN い/い/IN る/る/IN
適用/てきよう/IN さ/さ/IN れ/れ/IN て/て/IN い/い/IN る/る/IN

19088 61456
特許/とっきょ/IN 開示/かいじ/IN 所/しょ/IN
特許/とっきょ/IN 開示/かいじ/IN 所/しょ/IN

19088 61459
初/しょ/IN
書/しょ/IN

19088 61464
自動/じどう/IN 翻訳/ほんやく/IN
自動/じどう/IN 翻訳/ほんやく/IN

19088 61476
```

裁判/さいばん/IN 記録/きろく/IN
裁判/さいばん/IN 記録/きろく/IN

19088 61482
自動/じどう/IN 作成/さくせい/IN の/の/IN ため/ため/IN の/の/IN
自動/じどう/IN 作成/さくせい/IN の/の/IN ため/ため/IN の/の/IN

19088 61484
音声/おんせい/IN 認識/にんしき/IN
音声/おんせい/IN 認識/にんしき/IN

19088 61488
言語/げんご/IN
言語/げんご/IN

19088 61489
モデル/もでる/IN
モデル/もでる/IN

19088 61496
用い/もちい/IN る/る/IN
用い/もちい/IN る/る/IN

19088 61512
医療/いりょう/IN 文章/ぶんしょう/IN
医療/いりょう/IN 文章/ぶんしょう/IN

19088 61517
情報/じょうほう/IN 抽出/ちゅうしゅつ/IN
情報/じょうほう/IN 抽出/ちゅうしゅつ/IN

19088 61532
現在/げんざい/IN
現在/げんざい/IN

19088 61539
自動/じどう/IN 単語/たんご/IN
自動/じどう/IN 単語/たんご/IN

19088 61540
分割/ぶんかつ/IN
分割/ぶんかつ/IN

19088 61541
システム/しすてむ/IN
システム/しすてむ/IN

19088 61564
一般/いっぱん/IN 的/てき/IN な/な/IN 分野/ぶんや/IN の/の/IN コーパス/こーぱす
/IV から/から/IN 構築/こうちく/IN さ/さ/IN れ/れ/IN て/て/IN お/お/IN り/り/IN
一般/いっぱん/IN 的/てき/IN な/な/IN 分野/ぶんや/IN の/の/IN コーパス/こーぱす
/IV から/から/IN 構築/こうちく/IN さ/さ/IN れ/れ/IN て/て/IN お/お/IN り/り/IN

19088 61578
上記/じょうき/IN の/の/IN よう/よう/IN な/な/IN
上記/じょうき/IN の/の/IN よう/よう/IN な/な/IN

19088 61580
上述/じょうじゅつ/IN
上述/じょうじゅつ/IN

19088 61586
さまざま/さまざま/IN な/な/IN
様々/さまざま/IN な/な/IN

19088 61600
分野/ぶんや/IN の/の/IN 文/ぶん/IN を/を/IN
分野/ぶんや/IN の/の/IN 文/ぶん/IN を/を/IN

19088 61606
適切/てきせつ/IN に/に/IN 単語/たんご/IN 分割/ぶんかつ/IN
適切/てきせつ/IN に/に/IN 単語/たんご/IN 分割/ぶんかつ/IN

19088 61636
とりわけ/とりわけ/IN 、/、/IN
とりわけ/とりわけ/IN 、/、/IN

19088 61644
対象/たいしょう/IN 分野/ぶんや/IN 特有/とくゆう/IN の/の/IN
対象/たいしょう/IN 分野/ぶんや/IN 特有/とくゆう/IN の/の/IN

19088 61645
単語/たんご/IN
単語/たんご/IN

19088 61648
表現/ひょうげん/IN
表現/ひょうげん/IN

19088 61673
周辺/しゅうへん/IN
周辺/しゅうへん/IN

19088 61678
制度/せいど/IN の/の/IN 低下/ていか/IN
精度/せいど/IN の/の/IN 低下/ていか/IN

19088 61682
著し/いちじるし/IN い/い/IN
著し/いちじるし/IN い/い/IN

19088 61709
単語/たんご/IN
単語/たんご/IN