1J－6

自然言語処理による新しい日本語組版をめざして

松本　高幸† 佐藤　慎司† 筧　捷彦‡
†早稲田大学大学院理工学研究科 ‡早稲田大学理工学部

1　はじめに

ここで紹介するものは、従来のべた詰めの日本語組版とは違う、新しい日本語組版の方法である。

日本語文書の組版は、一部の例外を除いてべた詰めで組まれてきた。これに対して、欧文では単語毎にスペースが入れられ、単語が行で分割される場合は、ハイフネーションが入れられる。

我々が日本語文書を読み書きするとき自然に文節を意識しており、改行するときは、文節の切れ目で改行するのが自然である。しかし、印刷されたものは、一見して文節というものが分からない。

そこで、日本語を文節単位で考え、2章で欧文・手書き文書の特徴を、3章でいくつかの新しい組版方法を提案する。

2　欧文組版と手書き文書の特徴

2.1　欧文の特徴

1. 行分割は、ハイフネーション処理等の特別の場合を除いて単語中は、行分割地点となりえない。
2. 欧文文字は、個々の文字固有の字幅を持つ。
3. 隣り合う文字の組合わせにより、文字間を調整する。

2.2　手書き文章の特徴

1. かなは、漢字よりやや小さめである。
2. 品詞では、「助詞」「助動詞」が他の品詞より小さく、平仮名では、「こ」「め」「と」「の」「ろ」「さ」などが小さい。漢字と漢字にはさまれる1文字の平仮名は小さい。
3. 行末は、区切りのよい字句で区切られる。
4. 1つの文の最初と最後の文字は、大きめに書かれる。

Japanese Documents Processing System
by means of Natural Language Processing Method.
Takayuki Matsumoto, Shinji Sato, Katsuhiko Kakehi
Waseda University.

3　新しい日本語組版の提案

2章で上げた、欧文・手書き文書の特徴を参考にして、次の4つの処理を考えた。

3.1　文節間のスペーシング処理

欧文が単語を単位として処理しているように、日本語文を文節単位で考える。そこで、欧文で単語間にスペースが入れられるように、日本語においても文節間にスペースを入れる。

文節間にスペースを入れた例を図1に示す。

日本語文を 文節単位で 考える。

図 1: 文節間にスペースを入れた例

3.2　行分割処理

欧文の行分割処理のように、日本語においても文節間（区切りのよい字句）で行分割処理をおこなう。

この例を図2に示す。

この例を図
2に示す。　⟹　この例を
図2に示す。

図 2: 行分割処理の例

3.3　プロポーショナル処理

1文字1文字を全角幅に並べる組方に対し、各文字それぞれの幅に応じて文字を並べるプロポーショナル処理をおこなう。

ただし、漢字に関しては、すべて同じ文字幅として扱った。

プロポーショナル処理有無の例を図3に示す。

有　プロポーショナル処理
無　プロポーショナル処理

図 3: プロポーショナル有無の例

3.4 文字サイズ処理

2.2であげた、手書き文書の特徴を参考にして文字サイズを変える。

文字サイズの変更例を図 4 に示す。

よろしくお願い致します

図 4: 文字サイズの変更例

4 処理システムの構成

文書処理システムとして TEX を用い、各整形規則に対応する TEX の命令を文書に挿入することで実現する。

4.1 文節の切り出し

最初に与えられた文書に対して、形態素解析を行なう。そして、文節を次のように定義し文節の切り出しをおこなう。

文節 := 自立語の列 + 付属語

文節の切り出し結果を図 5 に示す。

文節 (普通名詞) の (名詞接続助詞)
切り出し (動詞) 結果 (副詞的名詞) を (格助詞)
図 (普通名詞)5(数字) に (格助詞)
示す (動詞)。(句点)

図 5: 文節の切り出し例

4.2 整形規則

4.2.1 文節間のスペース

文節区切りの前後の文字の組合せ（同字種、異字種）により、文節間に入れるスペース 2 つに分ける。

4.2.2 行分割

TEX には、行分割を制御する \penalty という命令がある。

そこで、文節間で行分割をおこなうために、文節間にこの命令を入れる。

4.2.3 プロポーショナル

平仮名、カタカナの各文字の前後で詰める幅を定義したファイルを用意し、このファイルを参照し、各文字間の詰め幅を指定する。

4.2.4 文字サイズ

文字サイズを、「大大」、「大」、「標準」、「小」、「小小」の 5 つを定める。

漢字を大サイズとし、それ以外の字種を標準サイズとする。手書き文書の特徴 2. の品詞、文字に関しては、1 つサイズを小さくし、特徴 4. の最初と最後の文字に関しては 1 つサイズを大きくする。

5 組版結果

3.1,3.2,3.3 の処理を同時におこなった例を次にあげる。

我々が 日本語文書を 読み書きする とき 自然に 文節を 意識 して おり、改行する ときは、文節の 切れ目で 改行する のが 自然である。

6 まとめ

印刷結果を見ると、文節が自然な切り方とは言えないが、文節の切れ目での改行は、うまくいっている。

すでに入力済みの文書に対して、形態素解析をおこない、文節を切り出しているので、システムが求めた文節と文書を書いた人の考える文節とは、完全に一致しないのは当然である。

我々が文書を入力する際、かな漢字変換から返された文節の切り目や品詞などの変換結果の正誤を判定し、間違っていれば正しく直しながら文書を入力している。

しかし、これらの情報は変換を決定した途端、すべて捨てられてしまう。これらの情報を入力文書と共に保存しておけば、書き手の文節の切れ目など正しく情報を得ることができる。

そこで今後、文書を入力する部分から、この論文で述べた処理までをおこなう文書処理システムを研究したい。

参考文献

[1] 「JIS ハンドブック情報処理（ソフトウェア・符号編)」p.727-p.738, 日本規格協会,1995.

[2] D.E.Knuth, "The TEXbook", Addison-Wesley,1984 （斉藤信男監修, 鷺谷好輝訳, 「TEX ブック」, アスキー出版局,1989）.

[3] アスキー出版局技術部責任編集,「日本語 TEX テクニカルブック」, アスキー出版局,1990.