日本語校正支援システム (Joyner) の研究について (4)
－正解語辞書作成－

2J－7

徐 国偉　伊吹 潤　中村 直人　松井 くにお
富士通研究所

1　はじめに

ワープロで日本語文書を作成する時に、仮名漢字変換ミスや思い込みなどよく誤りが発生する。Joynerは入力文から誤りを探し出し、正解語辞書を照合することによって誤りを修正する。誤り訂正を行なうためには大量の正解語が不可欠である。われわれは新聞記事のデータベースから一旦正解語候補を抽出して、それに対して絞り込みを行なうことによって正解語辞書を作成した。

本論文では、正解語候補の自動抽出と絞り込みによる正解語辞書の作成方法について述べる。

2　自動抽出

まず、抽出対象の新聞記事に対する前処理を行ない、それから記事に対する正解語候補の自動抽出を行なう。

2.1　対象データに対する前処理

正解語作成は対象データが絶対正しくなければいけないので、新聞記事データ（毎日新聞のＣＤ－毎日新聞'９１版、ＣＤ－毎日新聞'９２版）を対象としている。記事は見出しと本体などから構成されるが、見出しにはよく「第３２回毎日芸術賞決まる」のような、助詞の省略表現がある。このような表現は正解語候補の抽出を間違えさせるので、見出しを抽出対象外にした。また、新聞記事本体の「人事」記事に対しても対象外にした。本システムは著名人を対象にしているが、「人事」記事に記述しているのは組織のトップの人事移動ではなくて、組織内部の移動という理由からである。

2.2　正解語候補の自動抽出

われわれは以下のものを正解語候補とする。

1. 名詞連続
「湾岸‖戦争」、「多‖国籍‖軍」のような名詞連続

2. 動詞連用形と名詞の連続
「振れ‖止め金‖具」のような連用形と名詞の連続

毎日新聞９１年、９２年の記事（年間約１０万記事）に対して、日本語名詞句抽出ルーチン[1]を用いて、正解語候補の自動抽出を行なった。抽出された正解語候補の総語数は3,497,740語で、異なり語数は822,134語である。表１は抽出された、頻度つきの上位１０個の正解語候補である。

表1．上位10個の正解語候補

正解語候補	頻度	正解語候補	頻度
宮沢‖首相	5840	ブッシュ‖大統領	4132
湾‖岸‖戦争	5824	日本‖時間	4115
海部‖首相	4881	国連‖平和‖維持‖活動	4012
政治‖改革	4878	ゴルバチョフ‖大統領	3909
関連‖記事	4678	毎日新聞‖社	3590

3　正解語辞書作成

以下に抽出された膨大な正解語候補に対する分析と絞り込みについて述べる。

3.1　正解語候補の増加具合

正解語候補はどんなふうに増えるのか、一定期間に達してから増加が落ちるかどうかを調べるために、われわれは毎日新聞データに対して、年ごとに１２群に分割して、群を加えることに従って正解語候補の異なり語数を計算した。図１は正解語候補の増加グラフである。横軸は群数を表し、縦軸は正解語候補の異なり語数を表す。

この結果から抽出対象を加えることにつれて正解語候補が増加することが分った。この結果は林大氏などの結果[2]と一致している。

3.2　正解語候補頻度

抽出された膨大な正解語候補を絞るために、正解語候補の頻度計算を行なった。図２は頻度グラフを表す。横軸は語数、縦軸は頻度を表す。頻度グラフを見ると、

Japanese spelling corrector(Joyner)(4)
Guowei XU, Jun IBUKI, Naohito NAKAMURA, Kunio MATSUI
Fujitsu Laboratories Ltd.
1015 Kamikodanaka, Nakahara-ku, Kawasaki-shi 211 Japan

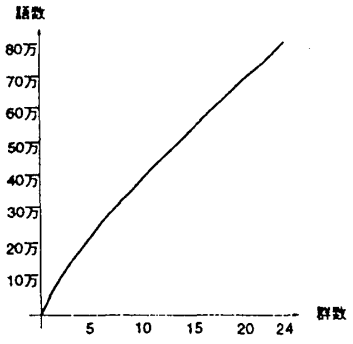

図 1: 正解語候補増加

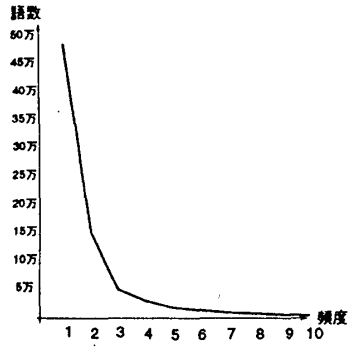

図 2: 正解語候補頻度

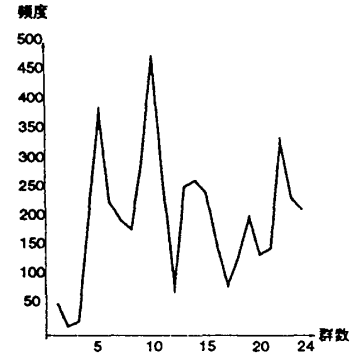

図 3: 「政治改革」の頻度の時間変化

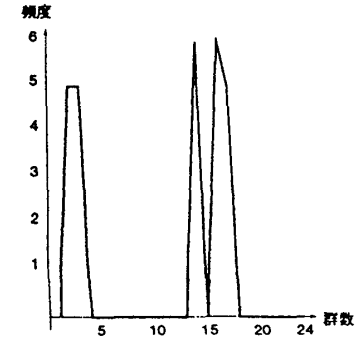

図 4: 「政府予算」の頻度の時間変化

頻度 3 以降になると語数が急激に減ることが分かった。頻度によって必要な正解語に絞ることができる。

3.3 正解語候補頻度の時間変化

新聞に出ている単語は時期によってバラツキがある。頻度だけで正解語を絞ることは、折角抽出した正解語候補を失ったり、最近全然使われていなくて、捨ててもいいような語を入れる恐れがある。そこで、われわれは正解語候補に対して頻度の時間変化を取ってみた。図 3、図 4 と図 5 はそれぞれ「政治改革」、「国家予算」と「イラク兵」の頻度の時間変化グラフである。横軸は群別、縦軸は頻度を表す。グラフから「政治改革」は 2 年間を通して高頻度で出現し、「国家予算」は特定の月に繰り返し出現し、「イラク兵」は湾岸戦争という特定の時期以降に全く出現しないことが分かった。この単語頻度の時間変化を利用して、正解語を作成する時に古い語の削除と正解語辞書のメンテナンスが出来る。

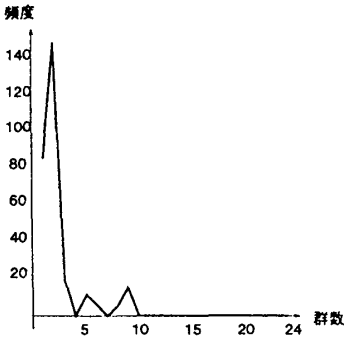

図 5: 「イラク兵」の頻度の時間変化

4 まとめ

正解語候補の自動抽出と抽出された正解語候補に対する分析を行なった。これらの分析を基に正解語候補から正解語辞書を作成した。実際に校正システムで使った結果を分析し、正解語辞書の改良を行なう予定である。

今後の課題としては名詞句以外の句を正解語として取り入れる研究が挙げられる。また、抽出データの対象を新聞記事以外に広げる必要がある。

最後に、本研究のために抽出ルーチンを提供してくれた西野文人研究員に感謝する。

参考文献

[1] 西野：" 日本語テキスト分類における特徴抽出", 情処研報ＮＬ No.112(1996)

[2] 林他：" 図説日本語－グラフで見ることばの姿－", 角川書店 昭和５７年