ngramモデルと単純な誤り検出ルールを用いた日本語スペルチェッカ

森輝彦　山本幹雄
筑波大学

1 はじめに

単純な誤り検出ルールに基づく誤り候補（および訂正候補）の生成部分と、候補の中から統計的言語モデル(*n*gramモデル)を用いて指摘するべき誤り箇所を絞りこむ部分から構成される日本語スペルチェッカの検討を行った。

*n*gramモデルに基づく統計的言語モデルは、対象言語を含むあらゆる文字列に対して「対象言語らしさ」を与える確率モデルである[北99]。*n*gramモデルは対象言語の正例コーパスから推定されるため、日本語の誤った記述に関する知識をまったく含んでおらず、誤りを直接同定するのには向かない。しかし、音声認識システムでの成功[鹿野他01]から推測すれば、２つの似た文があった場合、どちらがより日本語らしいかといった比較には有効に使えると想像される。また、規則による文法の記述に比べて、はるかに適用範囲が広いことも特徴である。これらの特徴を生かして、上記のような日本語のスペルチェッカを検討した。

ルールで検出された誤り候補は*n*gramモデルで検証される。このため、スペルチェック特有の知識として人間が書く検出・訂正ルールは網羅性を中心に考えることができるようになり、比較的容易に単純なルールとして追加することができる。本報告では、上記提案手法の概要と、留学生の書いた日本語作文を対象とした日本語スペルチェッカの試作・評価について報告し、本手法が有効に検出できる誤りとそうでない誤りを明らかにする。

2 *n*gramモデルを用いた日本語スペルチェッカ

提案する日本語スペルチェッカの処理手順は次のようになる。

(1) 網羅性の高い単純な誤り検出・訂正ルール（文字列マッチングと置換）を用い、誤り候補部分（と訂正候補）を検出する。
(2) 統計的言語モデルによって訂正候補と元の文の確率を比較し、訂正候補の確率がしきい値以上に高くなった場合に、誤りであったと判断・指摘する。

例えば、留学生がよく間違えるテ形の促音脱落を検出・訂正するために、「て」という文字があればその前に「っ」を入れてみるというルールを考えることができる。入力文の例とルールを適用した訂正候補は次のようになる。

入力：つかれてねむて しまった。
訂正候補1：つかれって ねむて しまった。
訂正候補2：つかれてねむって しまった。

それぞれの文の確率を*n*gramモデルで計算し（実際には文字あたりの確率になるように幾何平均する）、しきい値以上に訂正候補の確率が大きければ誤りと指摘・訂正する。「て」の前の促音が実際脱落している場合は訂正候補の方が高い確率となる可能性が高い（訂正候補2)。また、「て」の部分がもともと正しければ「っ」の挿入された文は正しくない日本語となる可能性が高い(訂正候補1)。この場合、訂正候補文の確率が低いため、誤りと指摘する可能性は低くなると期待できる。

このように、提案手法は*n*gramモデルにその精度を依存しているため、*n*gramモデルがモデル化していない大局的な言語現象に関する誤り検出には無力であると予想される。以下では、留学生の書いた日本語作文に対するスペルチェッカの試作と評価を述べるが、これは留学生の書いた日本語作文には活用の誤り等の局所的な誤りが多く含まれ、本システムの特徴を活かすことができると考えたためである。

3 留学生向け日本語スペルチェッカ

3.1 誤り検出・訂正ルール

検出・訂正ルールは、実際の留学生が書いた作文を観察して作成した。誤りは実に様々であるが、比較的多く出現し、単純なルールで広い

範囲に渡って訂正候補を出力できる誤りとして、「テ形・タ形の活用誤り」と「濁音・清音の誤り」について試作システムに組み込んだ。以下では、それぞれの種類の誤りを検出するために使用したルールについて述べる。

その他、本手法の検出能力を越えると思われるが、検証のために助詞の誤りのうち「を」を「が」と書いている部分の検出・訂正ルールも組み込み4節で他のルールと共に評価した。

3.1.1 テ形・タ形の活用誤り検出・訂正ルール

動詞のテ形の活用は、音便の規則が複雑なためため、留学生にとって特に学習が困難な活用である。例えば、次のような誤りである。

誤り：つかれてねむてしまった。
正解：つかれてねむってしまった。

誤り：その飲み物を飲った。
正解：その飲み物を飲んだ。

テ形の活用規則を表1に示す[益岡・田窪92]。留学生は規則を適用すべき動詞を間違える。この観点より留学生の作文データ中のテ形に関する誤りを分類したものが表2である。分析データとして、筑波大学留学生センターから提供していただいた留学生の作文データ(90人934文）を用いた。表中の各行が本来適用するべき活用規則を表しており（左端）、各列が実際留学生が適用した規則である（上の行）。各規則は表1中の活用の際接続する文字列で表現した。子音動詞であれば、最後の子音を削除して規則の文字列を接続する。ただし、「して」の規則を誤って「って」となるべき動詞（末尾がt,r,wの子音動詞）に適用したものは、語幹末尾の子音に"*i*"を付加した活用をしている。例えば、「売って」を「売りて」としている例である。留学生は存在しない規則も誤って作り出しており、それらも同様な記法で記述した。「っで」「んて」「で」である。その他のものは「その他」とした。

表2より、特に多い誤りは「って(語幹末尾がt,r,wの子音動詞)」と「て(母音動詞)」の規則の取り違えである。これだけで、全誤り109個のうちの72個を占める。[小森]によれば、「動詞テ形」の形成の誤りのうち60.7%が促音の脱落誤り、3.6%が促音の挿入誤りであると報告している。脱落・挿入の比率は大きく異なるが、促音誤りが多いことは今回の調査とも一致する。具体的な誤り検出・訂正ルールは以下のようになる。

テ形促音脱落誤り：「て」→「って」
テ形促音挿入誤り：「って」→「て」

この他、表2で少なくとも1例は間違っている活用に関して検出・訂正するルールを作成した(ただし、「その他」の規則を用いているものは一般性はないと考え今回作成しなかった)。例えば、「って」となるべきところを「んで」と活用してしまう誤り検出・訂正ルールは以下のようになる。

テ形その他ルールの例：「んで」→「って」

このようなタイプのルールを表2に従って、17個作成した(「って」に対する「して」の適用誤りに関するルールは、実際には2種のルールになる）。このグループのルールは誤りが多くないのでまとめて評価する(「テ形その他」と呼

表1　テ形・タ形活用規則

子音動詞については、語幹末尾の子音を削除して、活用規則の列の文字列を接続する。

動詞の種類（語幹末尾の子音）		テ形・タ形活用規則
子音動詞	s	して，した
	k	いて，いた
	g	いで，いだ
	m,n,b	んで，んだ
	t,r,w	って，った
母音動詞		て，た

表2　留学生のテ形活用誤り

正解＼適用した規則（存在する規則）	して	いて	いで	んで	って	て	適用した規則（存在しない規則） っで	んて	で	その他
して	(50)					1				1
いて		(9)			1					
いで			(1)							
んで	1			(58)			1	4	1	3
って	4	4		3	(235)	47	1	1	3	
て		1		3	25	(181)	1		2	1

※括弧の中は正しい適用数

ぶ)。テ形だけでなくタ形についてもほぼ同様な検出・訂正ルールを作成した。

3.1.2 濁音誤り

濁音と清音を取り違える誤りも多い。例えば以下のような例である。

誤り：ほうしをかふった。
正解：ぼうしをかぶった。

このタイプの誤りをおかす原因は、発音に関する誤解や、単なるタイプミスなど様々な理由が考えられる。しかし、今回はあまり分析的に考えずに、ひとまとめにして濁音-清音間の単純な置き換えルールを作成した。濁音になることのできる文字は「か行，さ行，た行，は行」の各文字であるため、相互に置き換え可能とすると40個の検出・訂正ルールができる。このうち、データ中に実際誤りが確認できた19個を選んで濁音-清音誤り検出ルールとした。

3.2 統計的言語モデル

統計的言語モデルは、毎日新聞1995年版を用いて、文字単位の3gramモデルと4gramモデルをCMU/Cambrigde SLM Toolkit [Rosenfeld95]を用いて作成した。スムージング手法はGood-Turing discountingを用いたBack-off法[北99]である。留学生の漢字能力は様々であるので、今回の作文データにも単純な漢字を平仮名書きしているケースも多い。このため、学習データとしては、新聞記事をそのまま使ったもの(以下「漢字」)の他、茶筌によってすべて平仮名に変換したテキスト(以下「平仮名」)と、元の新聞記事と平仮名に変換した記事を混ぜたもの(以下、「混合」)の3種類の学習データを用いて、性能を比較した。

4 留学生の作文データを用いた性能評価

4.1 評価データ

分析に用いた留学生の作文データに対して、誤りに関するタグを付与し、これを正解データとしてシステムの性能評価を行った。ただし、前節で述べたルールによって検出・訂正できる誤りのみに着目し、その他のタイプの誤りや複数の原因が複合して生じた誤りで判断に苦しむような誤りは評価対象から除いた。特に助詞の誤りは係り先の動詞と連動している場合が多く、単純に助詞が誤りとは判断できない場合が多かった。

4.2 言語モデルの違いによる性能評価

図1は、テ形・タ形の促音挿入誤り検出・訂正ルール（「って→て」と「った→た」）を使った場合で、*n*gramモデルの種類(3.2節)を変えて実験した再現率-適合率のグラフである。再現率は、評価データ中の全誤りのうちシステムが正しく指摘した誤りの割合である(必ずしも訂正結果が正しくなくともよいとした)。適合率は、システムが誤りと指摘した箇所のうち、実際に誤りであった箇所の割合である。確率を比較する際のしきい値の値を変化させることで再現率と適合率を変化させ、横軸に再現率、縦軸に適合率をプロットした。

混合テキストを用いて4gramとしたモデルが今回使用したデータに対しては最も高性能であった。今回のモデルは文字を単位としているため4gramでもスパースネスの問題は生じていない。また、評価データを書いた留学生の漢字の学習レベルが様々であったため、混合テキストがよかったと推測される。また、このモデルを用いたとき、テ形・タ形の促音挿入誤りに対する再現率が92%のとき適合率が89%と、高性能であることが分かる。

4.3 検出・訂正ルールごとの性能評価

前節で最高性能となった混合テキストで訓練した4gramモデルを用い、各ルール(グループ)ごとの再現率-適合率を計算したグラフを図2に示す。また、F値(適合率と再現率の調和平均)が最大になる点の再現率と適合率を表にしたものが表3である。表3の中で、「濁音-清音(しきい

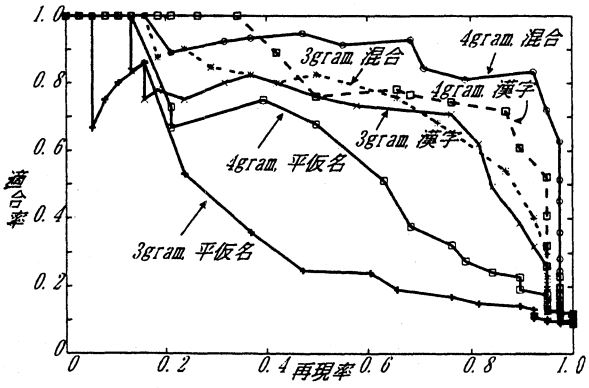

図1 各モデル毎の再現率-適合率グラフ
(テ形・タ形促音挿入誤り検出・訂正ルール)

値最適化)」とある行は、濁音-清音に関する各ルール毎にF値最大となるしきい値を個別に調整した結果を合わせて計算した値である。

図2より、テ形・タ形の促音挿入誤り、テ形の促音脱落誤り、特定の濁音-清音誤りについては、再現率約80%のとき適合率約80%以上と、高い性能であることが分かる。タ形の促音脱落誤りに対する性能が悪いが、指摘誤りを見ると、「たくさんいた」を「たくさんいった」に誤って訂正しているような例が多い。このように平仮名書きのテキストでは、正しい活用に促音を挿入すると他の動詞に変ってしまいどちらも局所的には正しい活用となってしまうことも多い。また、「テ形その他」の性能も高くないが、理由の分析は今後の課題である。

濁音-清音の誤りに関しては、特定の文字についてはかなり高性能である。代表例として、図2・表3には「ほ→ぼ」の書き換えルールの性能を載せた。しかし、すべてのルールを用いた平均はそれほどよい性能ではない。これは、

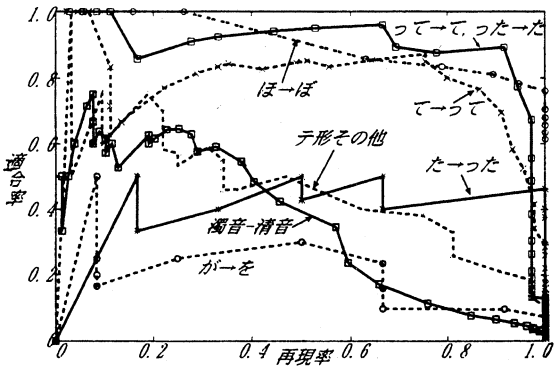

図2 各ルールに対する再現率-適合率グラフ

表3 F値最大点における適合率・再現率

誤り種類	適合率	再現率
テ形促音脱落	76%(39/51)	83%(39/47)
タ形促音脱落	46%(6/13)	100%(6/6)
テ・タ形促音挿入	89%(33/37)	92%(33/36)
テ形その他	38%(24/63)	75%(24/32)
濁音-清音(ほ→ぼ)	76%(19/25)	100%(19/19)
濁音-清音(19種)	55%(30/55)	38%(30/79)
濁音-清音(19種)しきい値最適化	66%(67/101)	85%(67/79)
助詞(が→を)	30%(6/20)	50%(6/12)

すべてのルールのしきい値を同じ値にして変化させ再現率-適合率を算出したためである。ルールごとに最適化を行うことにより(表3「濁音-清音(しきい値最適化)」の行)、再現率85%,適合率66%とよい性能が得られることが分かる。

助詞については、予想通り低い性能しか得ることができなかった。「が→を」以外のルールについても、検出結果の目視による印象評価を行ったが、さらに悪い結果となっている。

訂正結果まで正しいときにだけ正解とする実験も行った。この場合、全体的に約5%～20%の適合率の低下が見られた。より詳しい分析は今後の課題である。

5 おわりに

*n*gramモデルと単純な誤り検出ルールを用いた日本語スペルチェッカを提案し、留学生向けのシステム開発と評価について報告した。評価実験より、活用誤り等の局所的な誤りには比較的高精度に誤り検出・訂正が行えるが、助詞の誤り等の大局的な文情報を必要とする誤りには予想通り精度が悪いことが明らかになった。

今後は、日本語母語話者に対しても本手法が使えるか否かを検討していきたい。

謝辞

留学生の日本語作文データ利用の許可を与えて下さった筑波大学留学生センター西村よしみ教授、衣川隆生講師に感謝いたします。

参考文献

[北99] 北研二：確率的言語モデル，東京大学出版会，1999.

[鹿野他01]鹿野他：音声認識システム，オーム社，2001.

[小森] 小森早江子：「5. 分析例：作文データの中に見られる「動詞て形」の誤用分析」，http://langue.hyper.chubu.ac.jp/komori/JCSS98sec5.html

[益岡&田窪92] 益岡隆志、田窪行則：基礎日本語文法 -改訂版-，pp.14-16，くろしお出版，1992.

[Rosenfeld95] R.Rosenfeld: "The CMU Statistical Language Modeling toolkit and its use in the 1994 ARPA CSR evaluation", In Proc. ARPA Spoken Language System Technology Workshop, pp.47-50, 1995.