局所的及び大域的文脈を併用した日本語同音異義語誤りの訂正

角田 孝昭 †　　乾 孝司 ‡　　山本 幹雄 ‡

† 筑波大学 情報学群 情報メディア創成学類

‡ 筑波大学大学院 システム情報工学研究科

† tsunoda@mibel.cs.tsukuba.ac.jp

1　はじめに

近年、ブログ・マイクロブログ・口コミサイト等のユーザ参加型のウェブサービスにおいて、文書作成を専門とはしないユーザでもコンピュータで文書を作成しウェブ上で公開することが多くなっている。これらの文書は推敲や清書と言った過程を経ずに作成されることが多いため、かな漢字変換誤りやミスタイプ（typo）などの誤りが含まれたまま一般に公開されているケースが散見される [1]。

本研究ではそのような文書に見られる誤りの中でも、特に日本語同音異義語の誤りを取り上げる。同音異義語誤りは、かな漢字変換システムが提示した候補の選択間違いや文書作成者の知識不足によって発生し、以下を例として挙げることができる。文中の下線部が誤りであり、カッコ内が正しい語である。

例 1 多様な学習暦 [歴] を有する入学生に適したカリキュラム。

例 2 今回の戦争における被害は甚大なもので、たとえ停船 [停戦] が実現しても市民生活が回復するには長い年月が掛かるであろう。

本研究は、このような誤りをプログラムによって自動で訂正することを目的とする。例1のような誤りであれば、局所的な情報（「暦」の前の単語「学習」）のみでも同音異義語誤りの指摘・訂正が可能であると考えられる。一方、例2のような誤りは前後数単語の局所的情報だけでは誤りであるかどうかが判定できず、対象としている語からは位置的に遠い「戦争」や「市民生活が回復」と言った大域的な情報も併せて利用することによってはじめて訂正が可能になる。

[1] 例えば、「内臓した」と使われているものは誤変換であると考えることができる。2004年3月の調査では、「内蔵した」「内臓した」をGoogleで検索するとそれぞれ約10万件、約0.3万件（約3%が誤り）であったが [三品他, 2004]、2011年1月現在ではそれぞれ約195万件, 約6.9万件であり（約3.5%が誤り）、状況は以前から改善していない。

既存の手法の中でも、局所的情報を利用してかな漢字変換誤りを検出するものとして、固有名詞や複合語情報 [伊吹他, 1997]、直前の品詞や自立語・近くの名詞 [脇田・金子, 1996]、複合語に含まれる形態素と隣接する単語との間の意味的制約 [奥, 1996]、直前直後の単語及び前後3単語の自立語 [新納, 2000] を考慮するものがある。

一方、局所的情報と大域的情報を併用した手法としては、三品らが提案した3-gramと確率的LSAを併用したものが挙げられる [三品他, 2004]。この手法では、3-gramと確率的LSAをunigram rescaling法 [Gildea and Hofmann, 1999] によって併用することで、併用しなかった場合よりも性能が向上することが示されている。

本稿では、三品らの手法を発展させ、局所的情報として5-gramを、大域的情報として確率的LSAよりも性能が高いとされるLDA [Blei et al., 2003] を用いた同音異義語誤りチェッカーを作成し、その結果訂正性能が向上したことを示す。さらに、単語毎に局所的文脈モデルと大域的文脈モデルの重みの学習を試み、一部の単語に対して有効であることを示す。

2　同音語誤りの訂正手法

2.1　概要

まず、文書中に現れる単語 w に対する同音異義語を w' とする。ただし、候補が複数ある場合は尤度が最大のものを w' として用いる。同音異義語誤りの訂正は、文書中において同音異義語が存在する全ての単語に対して w と w' の尤度を比較し、w' の尤度の方が一定以上高くなれば訂正を行うことを繰り返すことで実現する。

単語 w に対する尤度 $L(w)$ は、5-gramモデルとLDAモデルが与える確率を unigram rescaling 法 [Gildea and Hofmann, 1999] によって併用した次の

式 (1) によって求める。

$$L(w) = \frac{P_{LDA}(w)}{P_{unigram}(w)} \quad P_{ngram}(w) \qquad (1)$$

また、w と w' の尤度の比に対数を取った $d(w, w')$ を次の式 (2) のように定義し、$d(w, w') > 0$ であれば誤りと判定して w を w' に置換する。

$$d(w, w') = \log \frac{L(w')}{L(w)} \qquad (2)$$

しかし、実際に用いる各言語モデルは完全なものではないため、全てを誤りとすると誤検出が多くなってしまう。そこで、訂正閾値 t を設定し、$d(w, w')$ が閾値を超えたらはじめて訂正を行うようにすることで誤検出の防止を図る。

さらに、同音異義語誤りを訂正する際は、同音異義語誤りの訂正に特化した学習に基づいて訂正を行うことでより正確な訂正ができると考えられる。本研究では各単語 w に対して、置換する閾値 t_w を設定する方法 [三品他, 2004] に加え、局所的文脈モデルと大域的文脈モデルの重みの学習を試みた。

2.2 置換候補リストの作成

同音異義語誤りチェッカーを作成するにあたって、「明らかに誤りであり訂正されるべきもの」が対象となるように同音異義語を定義する。同音異義語となりうる語には次の制約を加えた。

- 1 文字以上の漢字を含む。
- 固有名詞では無い。

人名等の固有名詞に関しては、人手でも正誤の判断が困難であるために除外している。さらに、二つの語が同音異義語となるための条件を以下のように設定する。

1. 同一の読みが存在する。
2. ひらがなとカタカナを排除した際に、片方の文字列をもう片方が含まない。

2) の条件は、例えば「締め切り/締めきり/締切」と言った正書法に起因する同音異義語を排除するためのものである。これらの表記揺れの訂正は同音異義語の訂正とは異なる手法で扱われるべきであるため、本実験の対象とはしない。

以上の定義に基づいて毎日新聞コーパス 2004 年版に出現した単語から同音異義語リストを作成し、10,304 組（22,754 語）の同音異義語を得た。さらに、訂正の際に単語の読み情報が無くても同音異義語候補が列挙できるように置換候補リストも作成した。これにより、例えば「降り」と言う単語に対しては「フリ」「オリ」と言ういずれの読みも考慮し、置換候補として「振り, 折り, ...」が考慮されるようになっている [三品他, 2004]。

2.3 N-gram モデルと LDA モデル

局所的文脈モデルは、単語 5-gram モデルを利用する。なお、同音異義語誤りの訂正において後ろの語も重要な手がかりとなることが多いと考えられる。このため、前向き N-gram モデル P_f と後ろ向き N-gram モデル P_b の幾何平均を取った次の式 (3) によって、局所的文脈 h_L（w_i の前後 4 単語）を条件とする単語の出現確率を計算する [三品他, 2004]。

$$P_{ngram}(w_i|h_L) = \sqrt{P_f(w_i|w_{i\ n+1}^{i\ 1})\ \ P_b(w_{i+1}^{i+n\ 1})} \qquad (3)$$

一方、大域的文脈モデルは LDA を利用する。LDA (Latent Dirichlet Allocation; 潜在的ディリクレ配分法）は、Blei らが提案したマルチトピックモデルの一つである [Blei et al., 2003]。確率的 LSA を拡張し、トピックの事前分布にディリクレ分布を導入することで過適応を軽減したものであり、確率的 LSA よりも一般に高い性能であることが知られている。本研究では、与えられた文書全体の単語を大域的文脈とし、大域的文脈 h_G を条件とする単語 w の出現確率 $P_{LDA}(w|h_G)$ を計算している。

2.4 訂正閾値及び各モデルの重み学習

単語によっては大域的情報が訂正において非常に有効な同音異義語の組もあれば（「引く/弾く」等）、逆に訂正を妨げてしまっているものも存在する（「以外/意外」等）[三品他, 2004]。このような単語に対しては、大域的文脈モデルの影響度を調整できるようにすることで性能が向上すると考えられる。

影響度の調整を考慮した場合の単語 w の尤度を、以下の式 (4) で定義する。これは unigram rescaling の式 (1) を各モデルに重みを適用できるよう改変したものであり、式中の $\ _w$ は単語 w に対する大域的文脈モデルの重みであり、特に $\ _w = 0$ の場合は大域的文脈モデルを全く考慮しない場合となる。

$$L(w) = \frac{{P_{LDA}}^{\alpha_w}(w)}{P_{unigram}(w)}\ {P_{ngram}}^{(1\ \ \alpha_w)}(w)\ \ (0 \quad _w \quad 1) \qquad (4)$$

各単語の訂正閾値及び重みの決定は、誤りが混入している学習用データにおいて、単語 w に対する $\ _w$, t_w を変化させ、当該単語の F 値 [2] を最大化させる組を採用することにより実現する。具体的には、$\ _w$ を $0 \quad _w \quad 1$ の間で 0.1 ずつ変化させ、さらに t_w を $-7.5 \quad t_w \quad 5.0$ の間で 0.1 ずつ変化させて最大となる $\ _w, t_w$ を選択することによる（t_w の範囲はヒュー

[2] データに含まれる同音異語誤りの数を M、チェッカーが誤りとして訂正した数を D、チェッカーが正しく訂正できた数を C とすると、F 値は再現率 $R = C/M$、適合率 $P = C/D$ を使って $F = 2PR/(P + R)$ で求めることができる。

リスティックに決定した）。ただし、パラメータを変えても F 値が同一になる場合は、「訂正の必要が無いものを訂正しなかった回数」と「チェッカーが正しく訂正できた数」の和を最大化する $_w, t_w$ の組を優先的に採用する。

3 実験

3.1 データ及びモデルの学習

訂正閾値と各モデルの重みの学習や、性能評価を行う際は、人間が実際に作成したテキストを対象とすることが望ましいが、誤りを含むテキストを大量に収集し人間の手により正解データを逐一作成することは困難である。

そこで、同音異義語リストに基づいて、コーパス中で置換可能な語の 1% をランダムに選択して置換を行ったデータを、「閾値と各モデルの学習用」及び「テスト用」の二種類用意する（一つの語に対して複数の語に置換可能な場合は、最も unigram 確率が高いもので置換する）。加えて、閾値と各モデルの重みの学習用のデータには、同音異義語リスト中に存在する各単語に対してそれぞれ 1 回の誤りが含まれるように置換する。

N-gram モデルの学習には SRI Language Modeling Toolkit 1.5.10 [Stolcke, 2002] を用い、補間法として Modi ed Kneser-Ney discounting を利用している。また、LDA モデルの学習には plda 3.0 [Wang et al., 2009] を用いた。

以上のモデルと学習用データを用いて、2.4 節の方法で閾値・重みを単語毎に求めた。それぞれのモデルの作成に使用したコーパス等の情報を表 1 にまとめた。

3.2 結果と考察

まず、単語別に訂正閾値や各モデルの重みを設定せず、全ての単語で同一の閾値を適用した場合の結果から示す。図 1 は、閾値を変えたときの再現率（Recall）と適合率（Precision）の変化をグラフにしたものである。破線が 5-gram のみ、実線が LDA と 5-gram を併用した場合の訂正性能を示し、曲線状にある白抜きの四角と丸がそれぞれ F 値が最大となる点（72.25%、72.47%）である。また、参考までに 3-gram のみ、LDA と 3-gram を併用した際の訂正性能を二点鎖線及び一点鎖線で示す。同一の閾値を採用している場合は 5-gram の方が性能が高くなり、さらに LDA を併用することでわずかに高い性能となることが分かるが、あまり大きな違いは見られない。

次に、全単語の大域的文脈モデルの重み $_w$ を 0 及び 0.5 で固定して各単語の訂正閾値のみを学習させたモデルと、各単語の局所的・大域的文脈モデルの重みと訂正閾値を学習させたモデルに基づいて訂正を行わせた結果を示す（なお、 $_w = 0$ の場合は $P_{LDA}{}^{\alpha_w}(w) = 1$ になるので大域的文脈を全く考慮しない場合となる）。表 2 にそれぞれの再現率、適合率、F 値を示す。また、図 1 上の塗りつぶされた丸が $_w = 0.5$、アスタリスクが $_w = 0$ に固定したものであり、四角が $_w$ を単語に併せて変化させたものである。いずれの場合も、単語別に閾値を設定すると全体で閾値を同一にするよりも性能が向上している。一方、各モデルの重みを調整させても全体の訂正性能にはあまり影響していないことが分かる。

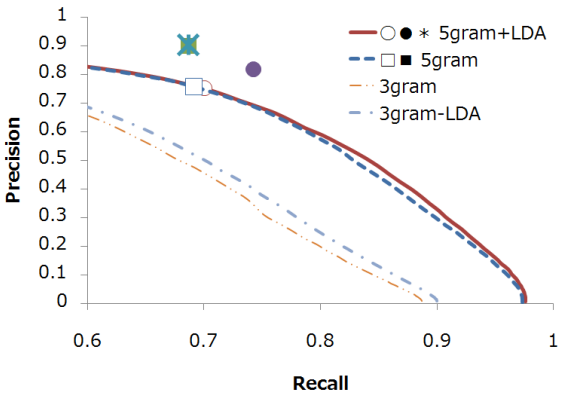

図 1: 同一の訂正閾値を適用した場合の性能比較

表 2: 各モデルの重み調整による訂正性能の変化

$_w$	再現率 (%)	適合率 (%)	F 値 (%)
0〜1	68.66	90.30	78.01
0	68.71	89.90	77.89
0.5	74.22	81.85	77.85

以上より、各モデルの重み調整を加えた結果は全体の訂正性能だけを見るとあまり良くなっていないと言える。この理由として、重みを学習するデータにおいて出現頻度が低い単語に対しては過適応により性能が悪化している可能性等が考えられる。しかし、局所的情報のみでも訂正性能が既に高い単語に対しては、大域的文脈情報を加えてもそれ以上の改善が見込めない。図 2 は、 $_w = 0$、すなわち局所的文脈のみを利用した場合における単語ごとの F 値の分布である。ここで既に F 値が高い語（図の左側にある語）ほど大域的文脈を加えても効果が見込めないため、F 値が低い単語に絞って、各モデルの重み調整による影響を調べた。

図 3 は、局所的文脈のみを利用した場合において、F 値が低い順に一定の割合で選択した単語集合に対して性能を比較したものである。破線が局所的文脈のみ（ $_w = 0$）、実線が局所的文脈と大域的文脈の重みを調整した場合（0 $_w$ 1）の訂正性能を示す。この図より、局所的文脈のみで F 値が低い単語に対しては

表 1: 実験条件

パラメータ	コーパス	補足
N-gram モデル	毎日新聞コーパス 1998 年-2000 年版	
LDA モデル	毎日新聞コーパス 1998 年-2000 年版	トピック数 200 で学習
同音異義語リスト	毎日新聞コーパス 2004 年版	
訂正閾値・重み学習データ	毎日新聞コーパス 2004 年版半年分（前期）	各単語 1 回 + 1% の割合で ランダムに誤りを混入（34,003 語）
テストデータ	毎日新聞コーパス 2004 年版半年分（後期）	1% の割合で ランダムに誤りを混入（33,362 語）

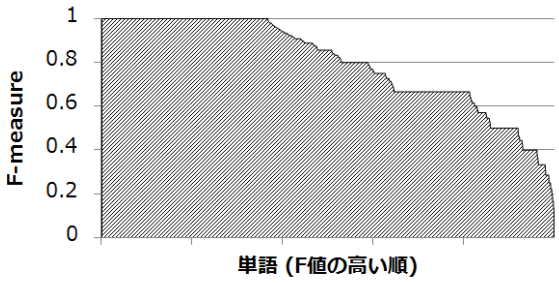

図 2: 局所的文脈のみを利用した場合における単語別の F 値の分布

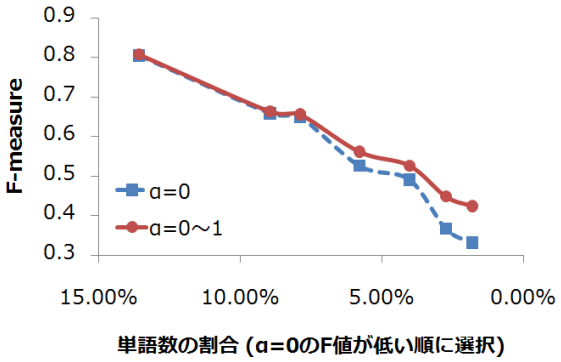

図 3: 局所的文脈のみで性能が低い語に対する重み調整の影響

大域的文脈を併せて利用することによって性能が向上していることが分かる。以下の表に、F 値が向上した単語の例を示す。

現れ	主な候補	局所的文脈のみ F 値（%）	重みを調整 F 値（%）	α_w
形成	形勢 京成	25.00%	83.33%	0.7
投手	党首 当主	30.77%	62.50%	0.3
終始	終止 修士 収支	44.44%	83.33%	0.7

4 おわりに

本稿では、局所的文脈として 5-gram モデル、大域的文脈として LDA モデルを利用することで、テストデータに対して F 値 72.47%と高い性能で同音異義語誤りを訂正できることを示した。さらに訂正において、単語ごとに訂正閾値と各モデルの重みを導入して性能向上を試みた結果、本手法では全体の性能向上にはあまり寄与しないものの、一部の単語に対しては有効性が認められることが分かった。

今後は、より同音異義語誤りに適した局所的・大域的文脈モデルの混合方法の研究をすることで、実用的な同音異義語誤りチェッカーの開発に繋げていきたいと考えている。

参考文献

David M Blei, Andrew Y Ng, and Michael I Jordan, 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, No. 4-5, pp. 993–1022.

Daniel Gildea and Thomas Hofmann, 1999. "Topic-Based Language Models Using EM," in *PROCEEDINGS OF EUROSPEECH*, pp. 2167–2170.

Andreas Stolcke, 2002. "SRILM - An extensible language modeling toolkit," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pp. 901–904.

Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y. Chang, 2009. "PLDA: Parallel Latent Dirichlet Allocation for Large-Scale Applications," in Andrew V. Goldberg and Yunhong Zhou eds. *AAIM '09 Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management*, Vol. 5564 of Lecture Notes in Computer Science, pp. 301–314.

伊吹潤・徐国偉・斉藤孝広・松井くにお, 1997 年. 「校正支援システム Joyner における表記誤りの訂正方式」,『情報処理学会研究報告. 自然言語処理研究会報告』, 第 97 巻, 第 4 号, pp.153–160.

奥雅博, 1996 年. 「日本文推敲支援システム REVISE における複合語同音異義語誤りの検出および訂正支援手法」,『電子情報通信学会論文誌. D-II, 情報・システム, II-情報処理』, 第 79 巻, 第 11 号, pp.1836–1846.

三品拓也・貞光九月・山本幹雄, 2004 年. 「確率的 LSA を用いた日本語同音異義語誤りの検出・訂正」,『情報処理学会論文誌』, 第 45 巻, 第 9 号, pp.2168–2176.

新納浩幸, 2000 年. 「表記情報をデフォルトの証拠として用いた決定リストによる同音異義語の誤り検出」,『情報処理学会論文誌』, 第 41 巻, 第 4 号, pp.1046–1053.

脇田早紀子・金子宏, 1996 年. 「変換ミスチェッカーのための辞書生成」,『情報処理学会研究報告. 自然言語処理研究会報告』, 第 96 巻, 第 7 号, pp.27–32.

毎日新聞社, CD-毎日新聞 1998 年版-2000 年版, 2004 年版, 日外アソシエーツ.