同音異義語誤りの校正における各種の共起制約データの有効性の評価

伊吹 潤
ibuki@flab.fujitsu.co.jp
富士通研究所メディア統合研究部

1 はじめに

我々はテキスト中の様々な種類の表記誤りを統一的に扱う校正システムの研究開発を行なっている.

本システムの主要目的は正確な (適合率の高い) 誤り指摘を広い範囲に渡って提供することにあるが, 現在は特に誤り指摘の対象を更に広げる (誤り指摘の再現率を向上させる) ことや, 現状の校正能力においてユーザの再現率や適合率に対する個別の要求に答えることを目標としている.

上記の目標のため、我々は様々な誤りの検証手段をシステムの枠組に組み込んでその評価を行なっている. ここでは対象を同音異義語に限って特に統語的なかかり受けデータ, 文内の共起関連度に関する統計量による検証を比較対照した結果を報告する.

2 本システムにおけるこれまでの取り組み

本校正支援システムは誤り内容の推定 (表記の揺れ, 同音異義語誤り) を行なう仮説生成部と推定された仮説の正誤の検証を行なう仮説検証部から構成される (図1参照).

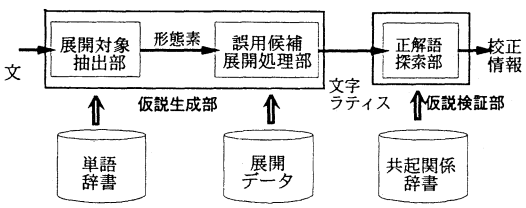


図 1: 本校正支援システムの構成図

仮説生成部の内部ではまず誤りの存在箇所を特定し (展開対象抽出部), 抽出された単語列に対して誤り内容の推定を行ない (誤用候補展開部), その推定結果をラティス構造と呼ばれる2端子グラフの形で提示する. テキストに対してラティスを対応させる処理を誤用候補展開処理と呼ぶ. ラティス構造の例を下図に示す.

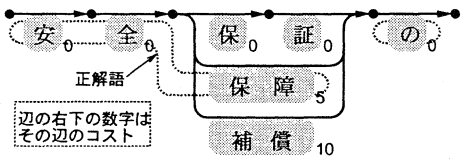


図 2: ラティスの例 (入力: 「安全保証の」)

(ここでは「安全保証の」というテキストに対して「保証」という単語が同音異義語誤りを含む可能性を考えた結果, 「保証」に対して「保障」, 「補償」が綴の新たな候補として原文に付加されている.)

正解語探索部は様々な辞書との照合によってラティス構造中から最適 (コスト最小) の経路を探索する. 一般に同音異義語誤りは単語自身では正誤の判定が不可能であり、このために検証のためには単語辞書以外の新たな単語同士の共起関係に関するデータが必要となる. 本システムの開発に当たっては, 複合名詞の語彙を校正済みのコーパスから自動的に収集しこれを単語辞書と合わせて最適なパスの探索に利用している.

これによって誤り語辞書や誤り検出用のヒューリスティックスによる同音異義語誤りの指摘においては分野や時間に対する依存性が高いために抜け落ちた部分 (固有名詞等) を大きなターゲットとしている. 又本システムによる誤り指摘ではコーパスから

半自動的に整備した複合語辞書に対しては複合語内の同音異義語に対して80% 程度の誤り検出のカバー率（再現率）を得ている.[1]

しかし同音異義語全般に対する対応を考えた場合, 複合語による検証のみでは用言や単独で現れる名詞が対象外となり、再現率の点での問題があるのも事実である. これに対する対応の前に同音語誤りを広く対象として再現率の高い検出を目指した既存のシステムについて見てみる.

3 誤り検出の再現率を上げる試みについて

同音語誤りを対象として再現率の高い検出を目指したシステムの試みとして脇田らのシステム [2] がある. ここでは同音異義語グループ内の特定の単語が出現する際に近傍に共起する単語の情報を利用して誤りの検出と校正を行なう仕組みを提案している. このシステムでは処理のために必要なデータをコーパスから自動収集する枠組を含んでおり, 再現率 80%, 適合率 96% 程度の結果を得ている.

又, 奥ら [3] は複合語内の同音異義語誤りを文字連鎖確率を用いて検出する方式を提案している. ここでは複合語を構成する文字領域内で辞書中に登録されたｎ個の文字連鎖とのマッチングを行ない、辞書に登録されていない文字連鎖部分を誤りとして認定する. この方式においては３文字連鎖に関するデータを利用した場合で 95% 程度の再現率, 適合率としては最高で 77% を得ている.

4 正解語探索部（誤り仮説の検証部）の拡張

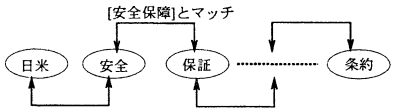
前述の手法では基本的には文内共起や文字列連鎖に関する手がかりを利用しているが, 我々はまず統語的な手がかりによる処理の拡充を行なった. 統語的な裏付けを持つことで、判断基準をユーザにわかり易くすることや、システムの調整における見通しをよくすることを狙っている.

一方で我々は特に再現率において統語的な処理を補うための方策として統計的な枠組の利用を検討しており、このために統計的な検証方式と統語的な検証方式の比較対照を行なった. 以後の節ではシステムに導入された各検証方式について説明する.

4.1 統語的な検証手段

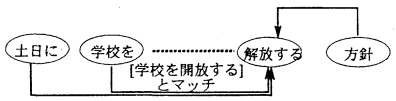
単語 bigram データによる検証

まず我々は名詞連続内での誤りの検出率の向上を図るために複合語をそのまま辞書に登録するのではなく、２単語単位の bigram データとして登録することとした. 検証時には名詞の連続部分に対して前から順に２単語の並びをとり、展開・正解語探索によって bigram 辞書と一致するデータの有無を見ていく. 一致するデータがあれば校正情報として単語に書き込んで行く（下図参照）.



述語と格要素との共起データによる検証

我々は更に, 単独の名詞や述語類を対象とする誤り検出のために述語と格要素間の共起データによるチェックを導入した. ここではテキスト中の述語とその近傍の格要素について同音異義語展開を行ない, 辞書中の共起データとのマッチングを試みる.



- 複数の共起データがマッチする場合は一番内側の格要素を最優先する.
- 係助詞, 埋め込み文にたいしては省略された格マーカーとして「を」・「が」のいずれかを仮説として推定する.

4.2 統計量による検証

まず判定の手がかりを広く求めるために対象単語と同一文中に共起する名詞群を距離に無関係に全てチェックすることとし, 対象とする同音異義語のペアについてコーパスにおいて同一文内の名詞群に対するｔ検定を行ない、この評価値 (t-score) を単語表記と共に辞書に登録した [4].

検証時には対象単語を同音異義語展開した後, 同一の文内の各名詞との共起データを全てチェックして、

その評価値を展開によって得た各表記について集計する．全ての単語に対するチェックが終ったところで、最も評価値の高い表記を正解として選択する．

5 実験環境

対象テキスト

中国新聞の 1996 年 1 月分の記事本文（校正済）から評価対象とした同音異義語ペアの 10 例（表 1 参照）の各々を含む文（単語毎に約 200 文）を抽出して対象テキストとした．

実験の手順

各文集合において対象とする同音異義語について校正支援システムによる正誤の判定を行ない，判定結果の出た文の比率を再現率，判定結果中の正判定の割合を適合率として記録した．

bigram 辞書データの整備

毎日新聞コーパスの'91 年から'95 年までの 5 年分のデータに対し，名詞相当語の連続部分から順に 2 つの連続をとりだし，生起頻度と共に記録した．その中から同音異義語をもち生起頻度が 2 以上のものを辞書に登録した．名詞相当語としては地名，人名，普通名詞，サ変名詞等を想定している．

格共起データの整備

bigram 辞書と同様に毎日新聞コーパスの 5 年分を対象として「名詞相当語＋各助詞＋述語」の連続部分を抽出し，その中から同音異義語をもち生起頻度が 2 以上のものを辞書に登録した．連続した格要素だけに対象を限ることによってデータ中に解析誤りによる不適切なデータが混入することを防ぐことを狙っている．

t-score データの整備

日本経済新聞 CD-ROM91 年,92 年版の 2 年分の記事から対象とする同音異義語のペアについて文内の名詞に対する t 検定を行ない、t 検定の評価値 (t-score) の絶対値が 2.15 以上の名詞を表記,t-score と共に辞書に登録した．これは判定精度 95% 以上を理論的に保証するためである [4]．

6 実験結果と考察

6.1 統語的な処理の拡充による効果

まず統語的な処理についての実験結果と考察を下に示す．

	複合名詞		bigram		格共起		格＋bg	
単語	Rc	Pr	Rc	Pr	Rc	Pc	Rc	Pc
挙げる	0	0	0	0	70	97	70	97
上げる	0	0	0	0	69	95	69	95
解放	33	100	42	100	29	94	64	97
開放	23	100	32	100	26	100	57	100
協議	6	95	20	100	41	98	56	98
競技	29	100	48	99	9	76	55	95
支持	30	100	45	100	32	96	75	98
指示	0	0	0	0	49	93	49	93
公判	22	100	61	98	25	100	75	98
後半	5	100	10	100	6	78	16	91
公園	6	90	51	100	10	86	56	97
講演	0	0	28	100	13	98	40	99
再建	25	100	53	98	22	91	68	97
債権	76	100	81	100	21	94	86	99
史上	34	100	47	100	1	100	48	100
市場	43	99	56	99	12	89	62	97
医師	10	100	19	96	8	83	27	92
意思	11	100	19	100	49	95	66	96
機関	21	98	71	99	14	92	72	97
期間	20	100	22	100	9	87	32	96

表 1: 統語的な制約による訂正結果

(**Pr** は適合率 (%),**Rc** は再現率 (%) である．又**複合名**は複合名詞による検証,**bigram** は名詞 bigram による検証,**格共起**は述語と格要素との共起データによる検証,**格＋bg** は共起データと bigram の組み合わせによる検証を表す.)

- 複合名詞による検証と比べ，名詞 bigram の検証の再現率の向上は明白であり，適合率は 98% 程度で複合語とほぼ同じである．
- 述語の格共起による検証の再現率は品詞に対する変動が大きい（名詞類に対しては平均で 20% の再現率，述語には 70%）．

  適合率は 95% 程度であり、複合名詞や bigram に比べて若干落ちる．これは共起データの不整合（同音群内で一部のみ欠損）による部分が 70% 程度であり、解析誤りによるものがそれに次ぐ．
- 名詞 bigram の処理対象と格共起の処理対象はほとんど重ならず両者を合わせた検証の再現率は平均で 60-70% 程度を期待できる．

- これらの処理の正判定部分に対する簡単なチェックでは処理誤りは見られず, 処理精度は高いと思われる. これらは人手による判定を行なう際に対象を絞り込む際の処理として有効であると考える.

- 対象外となる部分としては「史上」のような副詞,「後半」のように文の主題として記述されたり、数字を含む連鎖 (ex. 後半 30 分) が多くうまく一般化できない例が挙げられる.

6.2 統計量による処理の効果

次に統計量による検証についての実験結果および考察を下に示す.

単語	単純集計 Rc	単純集計 Pr	t-score Rc	t-score Pr	範囲限定 Rc	範囲限定 Pc
挙げる	97	77	97	77	84	76
上げる	93	53	93	76	74	88
解放	96	80	96	80	82	88
開放	98	90	98	90	90	91
協議	95	83	95	90	78	86
競技	96	92	96	88	88	84
支持	96	92	96	94	82	96
指示	91	74	91	72	55	76
公判	97	95	97	98	92	96
後半	79	60	79	67	59	67
公園	97	92	97	92	90	93
講演	93	73	93	81	79	87
再建	96	81	96	82	86	82
債権	100	84	100	94	96	96
史上	97	75	97	64	85	70
市場	94	87	94	94	85	90
医師	97	80	97	83	86	83
意思	93	82	93	83	70	90
機関	98	85	98	93	92	92
期間	90	81	90	83	80	85

表 2: 統計量による訂正結果

(**単純集計**ではマッチした件数を集計した, 又 **t-score** は各データの t-score を集計した, **範囲限定**は後者で更にチェック対象の名詞を対象単語の近傍 (後述) のものに限定したものである.)

- t-score による判定では常に 90% 以上の再現率を得ており、統語的な制約に比べ, 優位にあるが, 適合率では逆に 8 割程度と統語的な枠組に劣る.

- マッチした共起データの件数を単純に集計した場合よりは t-score による重み付けをした場合の方が一般的な傾向としては適合率が向上する.

- 範囲限定では離れた単語による影響を限るために述語の直後で文を分割してその後で判定 (重み付け) を行なった. 結果が改善された部分と改悪された部分が見られ, 処理内容のさらなる検討が必要と思われる.

7 まとめ

統計量に基づく処理は再現率に優れ,90% を超える部分について判断を提供できるが, 判断の精度は 80% 程度となり、統語的な処理の 95% の精度に劣ることが確認された.

今後の研究の方向としてはまず現在の統語的な枠組で未対応の手がかりを理論化して, 既存の統語的, 統計的枠組に組み込むことが挙げられる. 特に複合動詞や, 名詞連続以外の名詞句に対する共起制約についての実験を計画している. 又, 適合率や再現率に対する個別の要求に対処するために, 各検証方式による判断のユーザに対する提示方法や統合して全体としての判断を行なう仕組みについても検討中である.

最後に, 統語的なデータの整備を担当した徐国偉氏, 新聞の校正基準や誤り例, 実際の校閲過程についての詳細なデータや助言を頂いた福岡克氏, 原稿のチェックや技術的な問題点について指摘して下さった西野文人氏に感謝する.

参考文献

[1] 伊吹潤他: "校正支援システム Joyner における表記誤りの訂正方式", 自然言語処理研究報告,97-5,pp.29-35(1993)

[2] 脇田早紀子他: "変換ミスチェッカーのための辞書生成", 自然言語処理研究報告,96-111,pp.27-32(1996)

[3] 奥雅博・松岡浩司: "文字連鎖を用いた複合語同音異義語誤りの検出手法とその評価", 自然言語処理,vol.4 No.3 ,pp.83-99(1997)

[4] Uri Zernik: "Lexical Acquisition", Lawrence Erlbaum Asscociates ,pp.125-128(1991)