誤りを含み得るコーパスからの校正支援用データの整備

伊吹 潤, 西野 文人
ibuki,nisino@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同音語誤り等の検出・校正を行なう校正支援システムの開発を行なっており、システムでの処理に必要な共起データを新聞コーパスから抽出している. ところが校正の実験／評価を進める内に新聞コーパス中には同音語誤りが相当数含まれており, これがシステムの誤り検出能力の低下を招いていることがわかった.

我々はコーパス内での共起データの分布を統計的に処理して誤りの疑いの大きい部分を抽出し、それを人手で検証することによってコーパス中の誤りの除去を行なった. その結果、システムの誤り検出能力を向上させると共に実際の誤り例の収集を行なうことができた. 本稿ではどのように誤り候補を新聞コーパスから検出したかを示し, 実際に見つかった誤りの傾向を分析する.

2 共起データを利用した同音語誤りの検出

同音語誤りは単にテキストを辞書と照合するだけでは検出できず、文脈から用法の誤りの可能性を検出しなければならない. そのためには複合名詞における隣接単語間, あるいはかかり受け関係にある単語同士の共起情報を手がかりとして使うことがよく行なわれる.

例えば「安全補償」という単語の構成要素の「安全」と「補償」は正しい単語であり、辞書との照合だけでは誤りは検出できない. だが「安全」と「補償」は隣接しないが、「補償」の同音語である「保障」なら「安全」と隣接可能という共起情報を用いることによって「安全保障」に校正することができる.

2.1 共起データの整備

コーパスからの共起データの抽出

同音語誤りを広くカバーするためには共起データが大量に必要となるが、これを人手で整備するには限界がある. このためにコーパス・データから自動的に共起データを抽出することが広く行なわれている. 対象とするコーパスとしては現在のところ新聞記事が一般的と考えられる. これは最新のテキストが継続的に供給されていること、用字用語の基準が明記された均一の品質のテキストから構成されることによる.

誤り校正に利用可能な共起データへの絞り込み

同音語誤りの検出に対する利用を考えた場合, コーパスから抽出した共起データに対して更に検討すべき項目がある.

仮名漢字変換においては共起データの頻度の比率を利用して複数の同音語候補から優先度順位を決定すればよいのに対して, 校正支援では単に頻度比だけからどの候補が誤りかを断定することはできない. 例えば「運行を」と「再開する」の間の共起頻度が「運航を」と「再開する」の間の共起頻度よりかなり多いとしても「運航を」と「再開する」の共起データ例が少しでもあるならば誤り以外の可能性が残るため「運航を再開する」を誤りとはできない.

つまり誤り検出に共起データを利用できるためには, 特定のキーワードと同音群内の各単語との共起データを集めたセットの中で群内のいずれかの単語に共起データが欠落している（頻度が0である）ことが必須条件となる. この条件を満たす共起データのセットを禁止パターンと呼ぶこととする.

データ整備の観点から言えば禁止パターンに適合

しない共起データのセットを削除することで更にデータ量を絞り込むことができる.

3 新聞記事中の誤りの校正システムに対する影響

3.1 我々の校正支援システムの概略

我々が開発している校正支援システムは隣接単語の情報, 述語とそれにかかる格要素の間の共起データを利用して同音語の校正を行なっている [1]. 共起データ整備に対しては新聞コーパスのデータを抽出対象としている. 容量と効果のバランスを考えて共起データに対して頻度閾値を設けてそれにより絞り込みを行なっており、その後で禁止パターンの条件による2番目の絞り込みを行なっている.

3.2 コーパス中の誤りによって引き起こされた問題

ここで校正実験においてコーパス中に充分な生起例があり、共起データがあるはずなのに誤りの校正ができなくなる例が出てきた. その大きな原因は新聞記事中の誤り部分から抽出される共起データだった. 誤りを含む共起データについてコーパス中での検索を行ない件数を調べた結果を下に示す.

共起データ（正解）	件数	コーパス内での生起例
成果を挙げる（上げる）	9	成果を挙げる には至らず
勝を上げる（挙げる）	3	趙が初の一 勝を上げた
新興財団（振興）	4	北区文化 新興財団
遺骨収拾（収集）	6	遺骨収拾
社内広告（車内）	4	地下鉄の 社内広告

表 1: 毎日新聞コーパス内の誤り例

これらの誤りは発生件数が頻度閾値を越えたために共起データとして第1段階の絞り込みの後に残り、共起データが禁止パターンから外れたために誤り検出ができなくなっていた.

3.3 現行の枠組での対処方法と問題点

こうした誤りは校正処理の再現率の低下を引き起こす点で大きな問題となる. 誤りの件数のはそれほど多くないと考えられるので頻度閾値を現行の値より増加させれば誤りの影響は除けるであろう. しかし共起データの頻度分布を見ると頻度の低い部分に大多数のデータが分布しており（例えば頻度4以下の部分に全体の60%以上のデータがある）、ほとんどのデータが少ない頻度しか持たない. この状況で一様にデータ選別の閾値を上げれば抽出されるデータが大きく減少してしまう. 共起データの量を確保しつつ誤りの影響をできる限り除くためには誤りの影響が大きい部分のみにターゲットを絞って誤りの候補を検出する必要がある.

4 誤りの除去に向けての検討

同音語誤りの発生確率がどの同音語群についてもほぼ一定だと仮定すると単語自体の生起頻度が多い程, 誤りの生起頻度も高くなり、閾値を越える可能性が高くなるはずである. このため, まずコーパス中での生起頻度の高い同音語群に対象を絞ることとする.

また誤り例の観察からも誤りの発生頻度の絶対値は非常に低いことが想定される. このため共起データの頻度の上限（現行の頻度閾値程度）を設けて対象を絞る.

更に本来禁止パターンを示すような共起データセット中では共起データ内で相対的頻度が非常に少ない部分があるはずであり, その偏りの度合は誤りの混入によっても大きく変化しないと考えられるので、これも誤り候補を絞り込むための手がかりとして使う.

同音群内での分布の偏りの評価尺度としては t-score を利用する. ここで同音群中には単語 h_1, h_2 の2つが存在し、kw を同音語群と共起するキーワードとしよう. t-score は h_1,kw の同時出現（共起）の確率と h_2,kw の同時出現確率の相違を表す値として定義され、以下の計算式で表される [2].

$$t = \frac{p(kw|h_1) - p(kw|h_2)}{\sqrt{\sigma^2(p(kw|h_1)) + \sigma^2(p(kw|h_2))}}$$

5 除去作業の枠組

誤りの除去作業の枠組の細部について説明する.

抽出対象のコーパスはデータ整備の際と同様毎日新聞 CD-ROM5 年分の記事本文を対象としている. 又作業の大枠は 1) 同音語の範囲の設定 2) 共起データからの誤り候補の抽出 3) コーパス中の本文との照合による正誤の判定, という手順によって行なった. 以後各段階について説明する.

同音語の範囲の設定　以下の手順でテキスト中に頻出する同音語を選択した.

1. 名詞類約 7 万語, 述語約 5 千語について漢字, 平仮名で構成される単語中で読みと品詞が同じものを同一のグループにまとめ, 同音群とした.

2. それから更に抽出対象コーパス中で同音群内に単語単体での生起頻度が 500 以上のものを含んでいること, 名詞については更に文字数が 2 以上という条件によって名詞同音群 451 群, 述語同音群として 87 群を得た.

共起データからの誤り候補の抽出　対象とした共起データと同音語誤りの種別は 1) 隣接名詞同士（名詞） 2) 格要素と述語の共起（名詞） 3) 格要素と述語の共起（述語）, の 3 種であり、各々の場合について以下の手順で誤り候補の抽出を行なった.

1. コーパスから対象とする同音語群について共起データを抽出する

2. 件数が 10 件以下かつ同音群内での tscore の値が -2.1 以下のデータに更に絞り込む

6　作業結果と解説

抽出された誤り候補と実際に見つかった誤りについて表 2 に示す.

種別	共起データ種別数	誤り種別数
名詞連続部分	836	57
格要素／述語（名詞）	398	2
格要素／述語（述語）	1114	32

表 2: 作業結果

- 又述語／格要素の共起データは述語と格要素の双方を制約するはずだが、実際に検出される誤りはほとんど述語の同音語誤りばかりだった.

- 又副詞と述語の共起ではほとんど誤りは見つからないし、たとえ tscore の値がかなり高くとも校正に有用と判断できそうな共起パターンはなかった.

- 誤り候補中の実際の誤りの比率は 1 割以下しかないが、原因として考えられることを挙げる.

  – 数字, 記号類等が共起対象になった部分が誤り候補としてかなり残っており、かつほとんど誤りは見つかっていない. これらを抽出対象から外すだけでかなりの適合率の向上が見られるはず.

  – 誤り候補を見た限りでは同音語の群によって誤りの生起状況はかなりの違う.（例えば「話す／離す」等は全く誤りが検出できなかった）

7　同音誤りの傾向

同音語誤りを検出する際にどのような範囲を同音語とするかは誤り検出の再現率／適合率のバランスをとる上で重要な要因である. しかし、この範囲を実際に出現する誤りの分布に基づいて実証的に検討した文献はほとんどない. 名詞については共通する文字を持つ同音語同士が誤りやすいという仮説に対して既存の単語から一文字を置き換えて作った非語をコーパス中で検索して実証する例 [3] があるのに留まる.

7.1　名詞の同音誤りの傾向

- 検出された誤りについては件数的には意味的に類似したもの同士の間違いが目立つ. これは思い込みにより継続的に誤りを引き起こすためと思われる.

- しかし誤りの種類の点から見ると一見して誤りということが判るようなケアレス・ミスが優勢となる.

- その中では見た目が似ている単語（例えば共通の文字を持つもの）が最も多くを占めるが、それ以外について少数ながら誤りの可能性があることがこの結果からわかる.

- 共通文字をもつ名詞間

  収拾／収集 (6), 課程／過程 (2), 新興／振興 (6)

  債権／債券 (6), 一体／一帯 (3) , 指名／氏名 (2)

  国際／国債 (3), 社内／車内 (2), 史料／資料 (3),

  協会／教会 (4),

- 意味的な類似度の高いもの

  食料／食糧 (14), 主席／首席 (6), 障害／傷害 (5)

- それ以外

  経常／計上 (1), 後退／交代 (1), 構想／抗争 (1), 期間／機関 (2)、現象／減少 (1), 生涯／障害 (1)

7.2 述語の同音誤りの傾向

記事ハンドブック中には誤りやすい形容詞, 形容動詞の群についての記述がかなりあるが、実際の新聞記事内には形容詞, 形容動詞はほとんど出てこない. このために形容詞, 形容動詞についての共起データは非常に少なく、コーパスからの誤り検出もほとんどできていない. 形容詞類の使い分けに関する情報を集めるためには別のコーパスを必要とするだろう.

また動詞の中でも動詞の種類によって誤りやすいものと誤りがまったく検出されないものの区別がはっきりしていることも特徴と言える. 基本的には次のような動詞が誤りやすいと言えよう.

- 意味的な類似度の高い述語間

  聞く／聴く (36), 変える／代える (15), 打つ／撃つ (13), 現れる／表れる (8), 写る／映る (5)

- 抽象度の高い述語間

  上げる／挙げる (22), 合う／逢う (14), 上がる／挙げる (6), 犯す／冒す (2)

8 まとめ

- 新聞記事は校閲部のチェックを経てきたはずではあるが、少なくともCD-ROM上の記事5年分に対象を広げることによって誤りの生起回数は校正支援用のデータ利用には見逃せないほど大きくなることが確認された.

- 単純な条件によって100種類近くの誤りを見つけることができたことから考えると, 今回の作業の対象外の部分（頻度の低い同音語, 単独名詞類etc）にも誤りが含まれる可能性がある. 今後は他の共起データの利用や複数の共起データの総合によって誤りの検出対象を広げてゆく予定である.

- 今後は得られた正誤の情報を付加してコーパス上での誤り実例へのリンク付のコーパスを構築してゆく予定であり、これらを利用すれば校正システムの能力についてより現実的な評価をすることが可能となろう.

- コーパス中に実際に存在する同音語誤りの傾向についての情報が得られた. 今後はこれらの誤り傾向を同音語群データに対してフィードバックしてより精密な誤りモデルの構築を行ない, 適合率の高い校正指摘を目指してゆく.

最後に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に感謝する.

参考文献

[1] 伊吹潤他： “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”, 言語処理学会第4回年次大会予稿集,pp.626-629(1998)

[2] Uri Zernik： “Lexical Acquisition”, Lawrence Erlbaum Asscociates ,pp.125-128(1991)

[3] 新納浩幸： “誤りやすい同音異義語の収集”, 情処研究報告,98-NL-126-1,pp.1-8(1998)