意味カテゴリを用いた複合語の同音異義語誤り検定方式

2J-7

奥　雅博
ＮＴＴ情報通信処理研究所

１．はじめに

本稿では、文書中の同音異義語誤りを自動的に検定し、訂正候補を抽出する処理の第１歩として、単語間に意味的な連接関係がある名詞連続複合語を対象とした同音異義語誤り検定方式を提案する。

２．同音異義語誤り検定の現状と問題点

日本語ワードプロセッサの普及により、文書作成が機械化されてきている。これに伴い、文書中に存在する誤りを自動的に検定・訂正しようという試みが行われている[1]-[4]など。日本語ワードプロセッサの入力は、表音文字列を表意文字列に変換する方法が一般的であるので、変換ミスや利用者の思い込みによる同音異義語誤りを生じやすい。このため、同音異義語誤りを自動的に検定することは誤り検定・訂正を行うシステムにおいて重要であると考えられる。

従来の同音異義語誤りの検定方式としては、誤例辞書を用いる方式[2]、ＫＷＩＣを用いる方式[3],[4]がある。前者では、誤り語と正解語をペアで持つ誤例辞書を用いて同音異義語誤りを検定する。後者では、自立語をキーワードとしたＫＷＩＣを基に読みが同じで表記の異なる語を抽出する。しかし、これらの方法では、

①あらかじめ誤例辞書に登録した誤りにしか対処できない（誤例辞書方式）、

②システムが同音異義語誤りを自動的に検定することができない（ＫＷＩＣ方式）、

の問題点があった。

３．同音異義語誤り検定方式

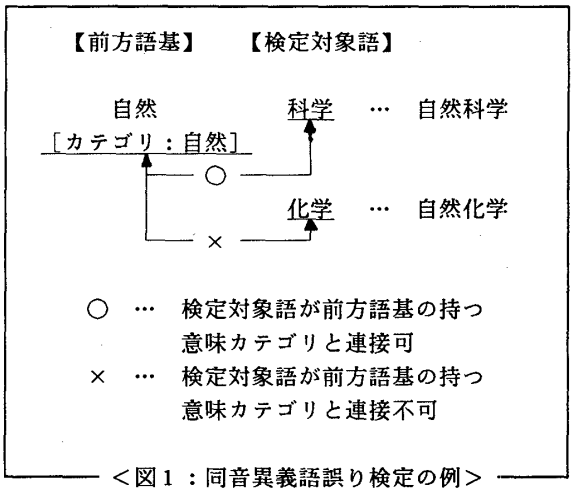
３．１　検定方式の概要

本稿では、上記の問題点を解決した同音異義語誤りを検定する処理を実現する第１歩として、対象を名詞連続複合語に絞り、その検定方式を提案する。

名詞連続複合語は、助詞を介さずに名詞や接辞を直接つなぐことによって構成された１つの句である。「名詞連続複合語において隣接しうる語基間には意味的な制約が存在する」と考え、同音異義語とそれに隣接する単語の持つ意味カテゴリとの連接関係をてがかりにした同音異義語誤り検定方式を考案した。意味カテゴリとは、名詞をその意味によって分類したものをいう。

本方式は、同音異義語の字面と、これに隣接している単語の持つ意味カテゴリとの間の連接可否を調べることによって同音異義語誤りかどうかを検定するものである。

図１に同音異義語誤りの検定例を示す。ここで「科学」と「化学」はともに「かがく」という同じ読みを持ち、同音異義の関係にある。これらの語と単語「自然」からなる名詞連続複合語、「自然科学」、「自然化学」を考える。「自然」は意味カテゴリ“自然”を持っているので、意味カテゴリ“自然”と字面「科学」との連接可否が○、意味カテゴリ“自然”と字面「化学」との連接可否が×であれば、「自然科学」は正しく、「自然化学」の「化学」は同音異義語誤りであると検定することができる。この方式によれば、検定対象語に連接しうる語を意味カテゴリの形でグループ分けしているので、いろいろな同音異義語誤りに対処することができる。



＜図１：同音異義語誤り検定の例＞

このような検定を行うためには、同音異義語の字面ごとに、連接しうる意味カテゴリを収集すればよい。ここでは、この連接関係を意味連接辞書に記述している。

３．２　意味連接辞書

意味連接辞書には、検定対象語の前方にくる単語の意味カテゴリとの連接を記述した前方意味連接辞書と、後方のそれを記述した後方意味連接辞書の２つがある。図２に前方意味連接辞書の内容例を示す。両方の意味連接辞書の各レコードは主に次の３つのフィールドを持つ。

・読み
・字面
・連接可否
　各意味カテゴリとの連接状況としていずれかの値を取る。
　○　…　この字面にのみ連接しうる
　×　…　この字面には連接しない
　△　…　この字面以外に同音異義の関係にある他の字面にも連接しうる

A Detection Method for Homonym Errors in Japanese Compound Nouns using Semantic Relation between a Homonym and Its Surrounding Words
Masahiro OKU
NTT Communications and Information Processing Laboratories

読み	字面	各意味カテゴリの連接可否			
		自然	組織	気体	生命
かがく	科学	○	△	×	○
	化学	×	△	○	×

＜図２：　前方意味連接辞書の内容例＞

例えば、図２からは、意味カテゴリ“自然”を持つ単語は「科学」には連接するが（連接可否○）、「化学」には連接しない（連接可否×）ことがわかる。

意味連接辞書は実際の文書に含まれている名詞連続複合語を調査することによって作成した。

４．評価実験

本同音異義語誤り検定方式の有効性を確認するために評価実験を行った。

４．１　前提条件

評価実験に際して、以下の前提条件を設けた。

・実験データに対する短単位分割、各語基への意味カテゴリ付与はすでに終了しているものとする。
・２つ以上の検定対象語を連続して含む名詞連続複合語は検定対象外とする。

４．２　実験方法

評価実験は、以下のデータを対象として行った。

・検定対象語
　３２種の異なる読みの語１００語
・意味連接辞書
　９０日分の新聞記事から検定対象語を含む名詞連続複合語を抽出し、検定対象語に隣接する単語の意味カテゴリを調査することによって作成
・実験データ
　①　教科書より抽出した検定対象語を含む名詞連続複合語（６０件）〔**正解語データ**〕
　②　①の名詞連続複合語に含まれる検定対象語をその誤り語に置き換えた名詞連続複合語（２２８件）〔**誤り語データ**〕
　③　新聞記事（意味連接辞書を作成したものとは異なる）から抽出した検定対象語を含む名詞連続複合語（３７６件）〔**正解語データ**〕

評価実験は、実験データの各名詞連続複合語に対して次の手順で行った。

（１）名詞連続複合語において検定対象語に隣接している単語の意味カテゴリを取得
（２）名詞連続複合語中の検定対象語をキーとして意味連接辞書を検索
（３）（１）で得た意味カテゴリと検定対象語との連接可否を意味連接辞書から取得
（４）検定対象語が誤りかどうか検定

（４）において誤りかどうかを検定する際には、３．２節で述べたように、連接可否には○、△、×の３つが存在するため、誤り検出の網羅性を重視して（多少正解語を誤りであると検定しても誤り語を見逃さないという立場）、△と×の両方を連接不可とする。

誤り検定の評価は、正しい語を正しいと指摘できる能力、誤り語を誤りとして検出できる能力の２つの観点から行う必要がある。本稿では、これら２つの観点を次の２つの値で評価する。

$$\text{正解指摘率} = \frac{\text{連接可と判定した語数}}{\text{実験データ中の正解語数}} \times 100$$

$$\text{誤り検出率} = \frac{\text{連接不可と判定した語数}}{\text{実験データ中の誤り語数}} \times 100$$

４．３　結果と考察

評価実験の結果を表１に示す。これは、誤り検出の網羅性を重視した結果であるので、誤り検出率は上限値を、正解指摘率は下限値をそれぞれ示していると考えられる。誤り検出率は９８．７％と非常に高い値を、正解指摘率も７２．０％と高い値を得ており、本方式が同音異義語誤り検出において有効であることがわかる。特に、正解指摘率は下限値であるにもかかわらず、従来法ではほとんど不可能であった正解語の指摘さえも７０％以上の割合で行えることを示している。

表１：　評価実験の結果

実験データNo.	誤り検出率[%]	正解指摘率[%]
実験データ①	－	６８．３
実験データ②	９８．７	－
実験データ③	－	７２．６
合　計	９８．７	７２．０

５．おわりに

本稿では名詞連続複合語に含まれる同音異義語誤りを、その前後の単語の持つ意味カテゴリを用いて自動的に検定する方式について述べた。評価実験の結果、誤り検出率＝９８．７％、正解指摘率＝７２．０％が得られ、本方式が同音異義語誤りの検出において有効であることが確認できた。

今後は、この方式に基づいて、検出だけでなく、訂正候補の提示、その順位付けの方法について検討を進める。

【参考文献】

［１］池原他：「日本文訂正支援システムREVISE」研究実用化報告、第３６巻第９号(1987)
［２］空閑　：「文書作成・校正支援システムWISE」電子通信学会、ＯＳ８６－２８(1986)
［３］福島他：「日本語文書作成支援システムCOMET」電子通信学会、ＯＳ８６－２１(1986)
［４］武田他：「日本語文書校正支援システムCRITACのユーザインタフェース」電子通信学会、ＯＳ８６－２２(1986)