応用

3.15 日本語校正

山本 和英（長岡技術科学大学）　鄭 育昌（富士通研究所）

日本語校正タスクの定義・特徴

日本語校正タスクは任意の日本語文章を入力して，誤りや不自然な部分（以後は不自然さも含めて単に「誤り」と呼ぶ）を指摘もしくは訂正するタスクである．これは検出のみの場合と訂正候補を提示する場合に分かれ，さらに候補提示の場合は単一候補と複数候補の提示に分かれる．入力が手書きの場合は仮名や漢字の表記も対象であり，電子化文書に限定しても，表記，語彙の選択，助詞，文法，意味的な整合性や文体，2 文以上が関係する場合など，日本語に関するありとあらゆる項目が誤りの候補（＝校正の対象）となり得る．

下記にいくつかの日本語誤りの例を挙げる．例文と訂正結果は文献 1）のものである．

（助詞の誤り）

日本の中でいろいろ<u>の</u>（⇒な）場所に行った．

（文法の誤り）

いつもお金を使いすぎ<u>ました</u>（⇒ます）．

（語彙選択の誤り）

バイクは<u>全然</u>（⇒完全に）壊れました．

（複合的な誤り）

私は<u>多いお金の使わないを知って</u>（⇒多くお金を使ってはいけないと知っていながら），買います．

日本語校正の現状

日本語校正に関する研究としては，日本語学習者（外国人）が執筆した文章が主な処理対象となっている．ここでは助詞の誤り検出・訂正のみに限定した研究が多く，産業日本語向けの研究や日本語教育向けの研究も行われている．一方，日本語話者向けの校正では保険関連文書向けの校正支援システムの研究がある．また，ある市販ソフト☆1 では語彙的な校正項目に加えて，助詞関連（助詞抜け，助詞の誤り，同一助詞の連続），修飾関係，並列関係，呼応表現なども対象としている．

日本語校正がほかの自然言語処理タスクと大きく異なるのは，処理の前提として形態素解析ができない可能性があるという点である．さらに，対象とするテキストを母語話者が執筆したかどうか，執筆者は子供か一般成人か，内容は専門的かどうかなど，文書の性質によって課題が大きく異なることも問題を困難にしている．

日本語校正システムの出力評価

我々は，2 つの日本語校正システムを用いて日本語学習者テキストに対して校正処理を行い，その結果に基づき日本語校正技術の現状を議論した．ここではその内容について紹介する．

校正システム A（市販製品）は，誤字・誤用，不適切な表現や，表記ゆれなどをチェックすることが目的であり，表現の洗練を求めるユーザを対象として開発された．校正システム B（開発企業内のみで使用）は，仕様書などの技術文書の品質を向上させることを目的とし，企業内の文書品質を高めるため開発された．ここで注目すべきは，実用化された両システムには想定ユーザが日本語話者であるという共通点があることである．システムの校正項目（機能）の考察から，両システムは「間違った日本語を正す」ことが目的ではなく，「より良い日本語を追求する」ことが目的である．

本タスクの評価に使用したテキストは，公開されている「オンライン日本語誤用辞典」[1] の事例を元に収集した日本語学習者テキストである．ここから抽出した 491 文（1,023 件の誤り）を処理対象として前記の校正システムに入力した．

☆1　http://www.justsystems.com/jp/products/justright/

誤用分類（大分類）	説明	件数	システムAの一致件数	システムBの一致件数
文法	助詞，複合辞，文型，テンス・アスペクトなどの誤用例	652	10	3
語彙	動詞，形容詞，名詞，副詞，連体詞，接辞，連語などの誤用例	334	97	65
句・文全体	文（句）の意味が不適切のため，全体的に書きなおす修正	37	0	0
合計		1,023	107	68

表 -1　日本語学習者テキストの誤用分類（大分類）と校正システムの一致件数

処理結果に対する考察

表 -1 に日本語学習者テキストの誤用分類と件数，および両校正システムが正確に指摘した件数を示す．表より，両システムは対象文書の誤用を指摘する能力は十分ではなく，特に大分類「文法」の指摘がほとんどできていないことが分かる．指摘できなかった事例で最も多い種類は，基本的な文法力と語彙力がある日本語学習者が「不自然な日本語」を作文したものである．たとえば，誤り原文「*言語大学で日本語を**勉強します***」に対し，下線部は「*勉強しています*」に修正すべきである．誤り原文自体は，形態素・構文解析の処理などが成功し，校正システムが搭載する文法誤りのパターンに合致しないため，検出できていない．このような作文は意思疎通に大きな支障は出ず，かつすべての不自然な言葉使いを網羅することができないため，両校正システムによる対応が困難であると考える．

一方，日本語表記の誤りによる形態素解析の失敗は，逆に誤り個所の検出に有効な情報になる．たとえば，誤り原文「*…先進国は**コーペンヘーガン**で国際的な**説金**に賛成すれば…*」の誤り個所（下線部分正解：コペンハーゲン，税金）の形態素解析が失敗，両システムにおいて未登録語と認識され，言葉の誤用であることを正確に指摘できた．特に校正システムAは未登録語の検出成功件数が多かった．ただし，未登録語としてユーザに修正を促す機能を有するが，未登録語に対する正確な言葉を提示することはできていない．

また，助詞の誤用のみの場合，助詞以外の単語が正確に解析されることで，助詞の指摘と校正が可能である．たとえば，誤り原文「*…毎回，**何を**する前に，…*」の誤り個所（下線部分正解：何か）は助詞扱いのみのため，ほかの形態素解析結果が正確であり，正しい校正ができた．校正システムBは特にこのような事例に長ける傾向が見える．

近い将来の達成可能性

前述したように，日本語校正といっても取り扱うべき現象は多岐にわたり，それらを網羅的に実現するのは当面不可能である．ただ，すでに一部は実用化されていることから分かるように，現状の技術で対処可能な誤りもある．今後は，自然言語処理全般の技術進展に同調して日本語校正の技術も徐々に高性能化していくであろう．

ただし，自然言語処理のほとんどの技術は形態素解析が正しく行われていることを前提に設計されているので，下記のような文字レベルで誤った日本語入力に対する校正の大部分は当面困難と予想する．残念ながら誤った日本語入力に対する形態素解析の研究はきわめて少ないのが現状で，実現のためには今後の研究活性化が望まれる．

（形態素解析できない誤り例）

学校でどんな大学を選ぶについてながい時間が考いました．

あとで日本語を勉強して初めると，この専問はひじょうに気にした．

でもコの語はいっしゃけんめいべんきょうする不可欠です．

いちばんいいつやくしゃはキイワげんご国立大学でそっぎょうします．

なお，本稿の詳細な内容については文献 2）を参照されたい．

参考文献

1）オンライン日本語誤用辞典（公開版 Ver.1.1），東京外国語大学望月圭子研究室，http://cblle.tufs.ac.jp/llc/ja_wrong/

2）山本和英，鄭　育昌：Project Next 日本語校正タスク，言語処理学会第 21 回年次大会併設ワークショップ (2015).

(2015 年 9 月 30 日受付)

山本 和英（正会員）yamamoto@jnlp.org

1996 年豊橋技術科学大学博士課程修了．博士（工学）．ATR 研究所を経て 2002 年から長岡技術科学大学，現在准教授．自然言語処理の研究に従事．

鄭 育昌（正会員）cheng.yuchang@jp.fujitsu.com

2008 年奈良先端科学技術大学院大学博士課程修了．博士（工学）．（株）ジャストシステムを経て 2011 年から富士通研究所．自然言語処理の研究開発に従事．