応用

3.16 英文校正

水本 智也（東北大学）

英文校正タスクとは？

英文校正は，人が書いた英語の文章を自動で校正するタスクである．英文校正と一口に言っても，文章を書いた人のレベルなどによって行う処理が異なる．たとえば，母語話者が書いた文章であれば，校正する個所は細かいところ（言い回し等）であるが，非母語話者の書いた文章であれば，校正はスペル誤りや文法誤りの修正が多くなる．

本稿では，自然言語処理の分野で多く研究されている非母語話者の英文中に出現する誤りを訂正するタスク（以後，英文誤り訂正）に焦点を絞る．**図 -1** に英文誤り訂正の例を示す．1 は冠詞誤りを含む文であり，不必要な“a”を削除している．2 は前置詞誤りを含む文であり，間違った前置詞“to”を“for”に訂正している．3 は名詞の語彙選択，主語と動詞の一致，動詞の語彙選択の 3 つの誤りを含む文をそれぞれ正しい単語に訂正している．

自動英文校正ってどうやるの？

古くはルールベースによる文法／スペルチェックツールがあり，近年では統計的手法を用いた誤り訂正が提案されている．統計的手法として，分類器や統計的機械翻訳の手法を応用した英文誤り訂正が研究されている．本稿では 2014 年に開かれた英文誤り訂正のコンペティションで 1 位であった統計的機械翻訳の手法を使った訂正手法を取り上げる．

図 -1　英文誤り訂正の例

通常の統計的機械翻訳は，日本語から英語（例：私は英語が好き→ I like English）に翻訳を行うものである．この際の翻訳ルールは日本語とそれに対応する英語が対になったデータを使って自動で構築する．一方，統計的機械翻訳を使った訂正では，誤った英語から正しい英語への翻訳（訂正）を行う（例：I likes English → I like English）．統計的機械翻訳を使った誤り訂正では，誤りを含む文と正しい文が対になった大規模なデータから訂正ルールを構築する．

英文校正ってどこが難しい？

英文校正の難しさの 1 つとして，学習者の犯す間違え方が多様であるという問題がある．間違え方が少ない誤りは比較的訂正が簡単ではあるが，間違え方の種類が多くなると訂正も難しくなる．

学習者のよく犯す誤りに冠詞誤りがあるが，冠詞の場合は“a”，“the”，“冠詞なし”の 3 種類である．前置詞であれば前置詞の個数分[☆1]である．このような誤りは決まった答えの中から正しいものを選ぶことで訂正ができる．

名詞や動詞は冠詞や前置詞に比べ単語数自体が多い．名詞の単数複数，動詞の三人称単数（主語と動詞の一致）のような誤りは冠詞や前置詞と同様で正解候補が少ないため訂正が簡単である．しかしながら，名詞や動詞の誤りでも異なる単語に書き間違え

☆1　訂正を行う際は出現数の多い前置詞のみを対象にすることが多い．

るような誤りも多くある．これはイディオムの覚え間違いで一部の名詞，動詞を違う単語に変えているものから母語の影響を受けているものなどがある．名詞や動詞自体の数が多いため，正解の候補となる単語をみつけることは難しい．

英文校正において訂正が簡単な／難しい誤りは？

統計的機械翻訳を使った英文校正の結果を実際に分析することで，英文校正で簡単な誤りと難しい誤りを見ていく．ここでは，システムがどれくらい正解の単語を当てることができるか（再現率）で訂正が簡単か難しいかを判断する．再現率はシステムが正解した個数を文中に含まれる誤りの数で割ったものである．統計的機械翻訳システムが1番良いと判断した訂正文だけで計算した再現率と，上位100個の訂正の中からスコアが最も高くなる訂正文を選んだときの再現率を比較することで訂正の難しさを判断する．

図-2に各誤りタイプの再現率を示す．冠詞，前置詞や主語と動詞の一致のような誤りは上位100個出力した場合，1つのみを出力した場合に比べて高い再現率である．正解候補が比較的少ないため，上位100個出力した中に正解の単語が入りやすいためである．一方，名詞／動詞の語彙選択においては，上位100個出力した際でも1つのみを出力した場合に比べて再現率が上がっているとは言いにくい．これは前章で説明したように名詞／動詞の語彙選択は間違え方が多様で正解の単語をみつけることが難しいためである．

誤りのタイプ	1番良い訂正のみ出力したときの再現率		上位100個出力中1番良い訂正を選んだときの再現率
冠詞	0.456	➡	0.824
前置詞	0.341	➡	0.608
主語と動詞の一致	0.299	➡	0.764
名詞の語彙選択	0.124	⇨	0.153
動詞の語彙選択	0.183	⇨	0.337
すべての誤り	0.270		0.502

図-2　誤りタイプごとの再現率

自動英文校正のこれから

これまで英文誤り訂正では分類器を使い，冠詞や前置詞の訂正を行っていた．学習者の文と正解が対になった大規模なデータが手に入るようになり，統計的機械翻訳の手法を応用して名詞／動詞の語彙選択のような誤りも扱われ始めたが，前章で示したように性能的にはまだまだである．この問題を本当に解決するには，学習者がなぜ間違えるかをもっと突き詰めて考える必要がある．母語の影響，母語自体の習熟度，学習方法による影響など個々人によって異なる部分をモデル化していくと面白くなりそうである．

難しさとしては言及していないが，冠詞や前置詞の訂正において，候補中から正解を探し出す部分も改善の余地がある．上位100個の訂正文を見たときに，システムが1番良いと判断した訂正文よりも良い訂正文があるということは，候補中から正解を選べてないということである．この問題についてはリランキングと呼ばれる訂正文を並び変える手法で改善が期待できる．また，実際にユーザにアプリケーションを使用してもらう際，1番目だけを提示するのではなく，2，3番目の訂正文も提示してユーザに選んでもらう方法で解決できる．

（2015年10月1日受付）

水本 智也（正会員）tomoya-m@ecei.tohoku.ac.jp
東北大学情報科学研究科特任助教．博士（工学）．2015年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．専門は自然言語処理．