冠詞誤り訂正時における訂正根拠の提示

梅澤次郎[†]　水野淳太[†]　岡崎直観[†‡]　乾健太郎[†]
東北大学[†]　科学技術振興機構　さきがけ[‡]
{umezawa, junta-m, okazaki, inui}@ecei.tohoku.ac.jp

1 はじめに

近年，日本人が英文を書く機会が増加しており，英作文を支援することへのニーズが高まっている．英語を母語としない日本人は，スペル，文法，冠詞，前置詞などで間違いを犯すことが多い．自然言語処理の分野では，これらの誤りを自動的に訂正する技術が研究されてきた [1, 2, 3, 4]．その中でも，冠詞や前置詞の選択誤りは，Helping Our Own (HOO) [5] といった評価型ワークショップが開催されるなど，近年注目を集めている．

冠詞や前置詞を自動的に訂正する研究は，着実に進歩を遂げているが，十分に実用的な段階には到達していない．このため，冠詞誤り訂正のユーザは，システムに提示された訂正結果を鵜呑みに出来ず，最終的には正しいと思う冠詞を各自の判断で選ぶことになる．つまり，学習者の文章を改善することを目的として考えた場合，システムの冠詞誤り訂正の性能を改善するだけでなく，最終的に学習者が冠詞を選ぶまでのトータルでの支援が重要である．そこで，本研究では，冠詞の誤り訂正の精度向上とは異なるアプローチとして，訂正の根拠を提示することで，学習者が最終的に正しい冠詞を選ぶことに寄与するかどうか探求する．

既存の冠詞誤り訂正システムとして，ESL Assistant [6] がある．このシステムは自動訂正と同時に根拠を提示するため，本研究と関連が深いが，ESL Assistant は誤りの訂正の際に，ウェブ検索エンジンから実際に使われている用例を提示するだけで，訂正の根拠としての良さを考慮している訳ではない．

冠詞は文脈から概ね決定することが可能である．理想的には，冠詞誤り訂正のルールを作り込み，そのルールを訂正の根拠として提示したいが，冠詞決定のルールを人手で書き尽くすことは困難である．近年の冠詞誤り自動訂正の研究も，ルールベースから学習ベースのものに移行し，訂正の精度を大幅に改善している．特に，Knight ら [7] や Han ら [2] は，名詞句に対する冠詞の分類問題に帰着させることで，誤用コーパスを必要とすることなく，高い精度で冠詞を訂正出来ることを示した．

我々もこれらの先行研究のアプローチを踏襲し，名詞句に対する冠詞の分類問題を考える．訂正の根拠としては，分類器が算出した冠詞の事後確率，分類に貢献した（冠詞を決定づけた）素性，入力文と類似した例文の 3 種類を提示する．評価実験では，冠詞の分類性能の評価と，根拠の提示の有効性を検証するための評価を行った．実験では，英文中の名詞句に対して，自動訂正結果のみを参照する場合，自動訂正結果と一緒に根拠情報を参照する場合を比較し，英語学習者の冠詞選択の正解率が向上することを確認した．この結果から，自動訂正時にその根拠も提示することの有用性が示された．

2 課題設計

本研究では，英語の学習者が入力した文を自動的に解析し，冠詞の用法に関する間違いを検出したうえで，その訂正候補と根拠となる情報（冠詞の用法に関する知識や例文）を提示する．例として以下の文を考える．

A remainder of this section describes ...

下線で示した名詞句 *a remainder* の冠詞は誤りであるから，提案手法では *the memainder* を訂正候補として提示することが期待される．この箇所の冠詞が *the* になる理由としては，*of this section* という限定が付くことにより「何の残りなのか」が明確であること，*remainder* という名詞は実体・概念との（残余という）関係を記述するものであるから，そもそも指し示すものが限定される傾向にあることが挙げられる．このように，冠詞が決定される根拠を学習者に提示することで，提案手法が提示した訂正候補の信頼性を検証できるようにするとともに，学習者の今後の誤用を防止する効果も期待できる．

さらに，*a remainder* の冠詞が *the remainder* になりやすいことは，次のような用例からも判断できる．

The remainder of this paper is organized
Formulas for the remainder term of Taylor

万が一，システムが *remainder* の冠詞を正しく訂正出来なかったとしても，これらの訂正根拠を閲覧する

ことで，ユーザーが適切な冠詞を選べることが期待できる．

以上のように，本研究は「冠詞の誤り訂正」「冠詞の訂正根拠の提示」「冠詞訂正に関する例文の提示」の３つのタスクをセットで扱い，英作文の支援システムを構築する．これらの３つのタスクは互いに関連があり，矛盾する情報を学習者に提示するのは好ましくない．また，英語の冠詞は複合的な要因で決定されるので，冠詞を決定するためのルールを書き尽くすことが難しい．冠詞の自動訂正の研究でも，近年ではルールベースの訂正手法と比較して，教師あり学習に基づく手法の方が性能が高い．

そこで，本研究では教師あり学習手法で冠詞の訂正モデルを構築するとともに，訂正に寄与した素性を見つけ出すことで，誤り訂正の根拠として提示することを考える．さらに，冠詞の訂正に関連する例文を提示するときも，訂正モデルで学習した素性の情報を利用することで，訂正根拠の理解を助けるような例文を提示する．

3 手法説明

3.1 冠詞の誤り訂正モデル

冠詞誤り訂正モデルを学習するには，誤りを含み，かつその訂正情報が付与されたコーパスが必要である．そのようなコーパスとして，Camblidge Lerener Corpus (CLC)[1], NICE[2] が挙げられるが，機械学習に十分な量のデータが入手できないという問題がある．そこで，冠詞誤り訂正モデルを学習するのではなく，正しく書かれた大量の文書に含まれる名詞句の冠詞の用法を，名詞句の文脈から予測するという分類問題に帰着するアプローチが取られてきた [7, 2]．

本研究でもこのアプローチを採用する．名詞句の周辺文脈から作成された素性ベクトルを $\boldsymbol{x}$，その名詞句に対応する冠詞を $y \in \{\mathrm{a}, \mathrm{the}, \phi\}$ とする．ϕ は冠詞無しを示す．最大エントロピー法を用いると，冠詞 y の条件付き確率は次式で求められる．

$$P(y|\boldsymbol{x}) = \frac{\exp\left(\sum_i w_{y,i} x_i\right)}{\sum_{y \in \{\mathrm{a},\mathrm{the},\phi\}} \exp\left(\sum_i w_{y,i} x_i\right)} \quad (1)$$

ここで，$w_{y,i}$ は素性 x_i に対して冠詞 y を予測するときの重みである．素性の重み $\boldsymbol{w}$ の学習には，Classias[3]を用いた．

[1] http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus)

[2] http://sugiura5.gsid.nagoya-u.ac.jp/~sakaue/nice/

[3] http://www.chokkan.org/software/classias/

表 1: 素性の一覧

素性タイプ
名詞句全体の単語
名詞句全体の品詞
名詞句末の単語
名詞句末の単語と品詞
名詞句の 1 語前の単語と品詞
名詞句の 1 語後の単語と品詞
名詞句の 2 語前の単語と 1 語前の単語
名詞句の 2 語前の単語と 1 語前の品詞
名詞句の 1 語前の単語と名詞句末の単語
名詞句の 1 語前の品詞と名詞句末の単語
名詞句の 1 語前の単語と 1 語後の単語
名詞句の 1 語前の単語と 1 語後の単語
名詞句末の単語と名詞句の 1 語後の単語
名詞句末の単語と名詞句の 1 語後の品詞
SPECIALIST Lexicon の variant 属性
名詞句末の語の可算性
名詞句の 2 語前と 1 語前が (type of, sort of, kind of...)
名詞句の最初が数字
名詞句の最初の語が大文字始まり (文頭の場合は考慮しない)
名詞句全体の語が大文字始まり (文頭の場合は考慮しない)
to+不定詞が続く
名詞句内に the を伴いやすくする語がある (usual, only, following...)
名詞句末の語が自己限定名詞 (back, balance, center...)
that 節を伴って冠詞が決まりやすい語

3.2 節で詳しく述べるが，素性は訂正の根拠としてユーザーに提示することを想定しているため，分類精度の向上と同時に，人間にとっての可読性の良さを確保したい．機械学習を用いたアプローチでは，素性の組み合わせを導入することで分類精度を改善させるが，本研究では冠詞の推定に効果がありそうな素性を作り込むことによって，分類精度と可読性を両立させることを考えた．

表 1 は，今回の研究で用いた素性の一覧である．単語，品詞などの一般的に用いられる素性の他に，ヘッドの名詞の加算性を SPECIALIST Lexicon[4] で調べたものと，文献 [8] から設計した素性を追加した．

冠詞誤り訂正モデルを学習するときは，正しい英語が書かれている文書を Natural Language Toolkit (NLTK) [9] を用いて文単位に分割し，さらに GENIA tagger [10] を用いて，品詞のタグ付けとチャンキングを行うことにより，名詞句を認識する．認識されたそれぞれの名詞句に対して，付与されていた冠詞を正解のラベルとし，その名詞句の文脈から素性ベクトルを作成する．

3.2 冠詞訂正根拠の提示

本研究では，学習者が入力した文中の名詞句に対して，その冠詞が *a/an*, *the*, ϕ である場合の根拠を提示する．冠詞の誤り訂正モデルでは，素性 x_i はその名詞句の冠詞が *a/an*, *the*, ϕ である重みとして，それぞれ $w_{\mathrm{a},i}$, $w_{\mathrm{the},i}$, $w_{\phi,i}$ が求められている．そこで，素性 x_i が冠詞 y を推す強さ $d(i, y)$ を，次式で求める．

$$d(i, y) = \frac{\exp(w_{\mathrm{label},i})}{\exp(w_{\mathrm{a},i}) + \exp(w_{\mathrm{the},i}) + \exp(w_{\phi,i})} \quad (2)$$

[4] http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/2011/web/index.html

本研究では，それぞれの冠詞 y に対して，$d(i, y)$ の値が高い 3 つの素性を選び，冠詞訂正の根拠とした．実際にユーザーに根拠を提示するときは，素性の名前を英語学習者にもわかりやすい表現に言い換えて表示している．

3.3 冠詞訂正に関する例文の提示

冠詞訂正の根拠として例文を提示するときは，分類器が予測した冠詞 $\hat{y}$ の決定に最も貢献した 2 つの素性を選び，その素性を含む名詞句を候補としてコーパス中から検索する．そして，訂正対象の名詞句の素性ベクトルと，候補の名詞句の素性ベクトルのコサイン類似度を計算し，類似度の高い 3 つの名詞句を含む文を用例として提示する．

4 評価実験

本手法を評価するために，2 つの評価実験を行った．1 つは冠詞予測の分類性能を評価する実験．もう 1 つは被験者により本手法の訂正性能の評価を行う実験である．

4.1 実験データ

2 つの評価実験のいずれにも，実験データは，ACL Anthology Reference Corpus (ACL ARC) [11] を用いた．ACL ARC の XML から文部分を取り出した結果，約 400 万文となった．さらに，著者や図のキャプションといった，不要な部分を取り除くフィルタリングを行った結果，893,201 文，4,766,272 個の名詞句が得られた．このうち，4,760,588 名詞句を学習用データとして用い，残りの 5,684 名詞句を評価用データとして用いた．

4.2 分類性能の評価

名詞句に対する冠詞の分類性能について評価実験を行った．全ての事例を ϕ に分類するベースラインの正解率は 60.31%である．分類器の学習には学習用データの全てを利用し，評価用データに対する冠詞分類の正解率を計測した．3 種類の冠詞の分類の正解率は 85.61%で，ベースラインを大きく上回る性能であった．

4.3 被験者による評価

被験者による評価実験では，実際に被験者に対して冠詞を選択してもらい，その正答率を評価した．被験者は英語を第二言語とする日本人大学生 1 名で，本システムの開発には携わっていない．

被験者には，冠詞を選択すべき名詞句が含まれた文が与えられる．まず，文のみを閲覧して，名詞句にふさわしいと思われる冠詞を選択してもらう．次に，システムが提示した情報を閲覧して，一度のみ選択を変更することを許す．システムが提示する情報は，(1) 分類器の冠詞予測結果のみ，(2) 冠詞ごとの予測確率，重みの高かった素性，用例，の 2 つのケースを用意した．文のみを閲覧したときの正答率に対し，(1) の情報を閲覧したときと，(2) の情報を閲覧したときとで，正答率がそれぞれどの程度向上するかを調べた．(2) の情報を閲覧したときの向上率の方が大きければ，訂正根拠を見せるという提案手法の有効性を確認できる．

図 1 にシステムの出力例を示す．この例において，被験者は，まず文脈から *roles* の冠詞は *the* であると判断したが，同時に自信がないことも明記していた．次に，システムが提示した根拠を見ることで，最終的に *the* を選択した．その理由として，提示された根拠の素性から，名詞句末の単語と名詞句の 1 語後の単語が *role of* のとき，冠詞が *the* になる傾向があることを理解したと考えられる．また，用例には修飾語がないことから，*the* を選択したことも考えられる．

実験の結果，文のみを閲覧したときの学習者自身の正答率は 72.59%であった．(1) の情報を閲覧することで正答率は 84.17%に上昇し，(2) の情報を閲覧した場合は正答率は 87.98%となった．(1) と (2) での正答率の向上幅は，ぞれぞれ 11.58%と 14.34% であり，訂正根拠が正答率の向上に貢献することが確認された．また，(2) の場合の正答率 87.98%は，システムの分類性能である 85.61%を上回った．

5 おわりに

本稿では，冠詞の誤り訂正という課題に対して，システムの訂正性能を改善するだけでなく，最終的に学習者が冠詞を選ぶまでのトータルの支援として，訂正の根拠を提示することを提案した．評価実験では，冠詞の分類性能の評価と，根拠の提示の有効性を検証するための評価を行い，自動訂正時に訂正根拠も提示することの有用性を示した．

今後の課題として，まず，大規模な評価実験があげられる．次に，より精緻な評価として，今回は 3 種類の根拠を同時に提示したが，今後はそれぞれの情報を個別に提示することで，根拠ごとの有効性を評価する．最後に，訂正の根拠を提示するという本研究のアプローチを，冠詞訂正だけでなく，前置詞の訂正や，他の誤りの訂正時にも適用していく．

謝辞 本研究は，文部科学省科研費（23240018），文部科学省科研費（23700159），および JST 戦略的創造研究推進事業さきがけの一環として行われた．

参考文献

[1] 永田亮, 井口達也, 脇寺健太, 桝井文人, 河合敦夫, 井須尚紀．前置詞情報を利用した冠詞誤り検出．電子情報通信学会論文誌．D-I, 情報・システム, I-情報処理, Vol. 88, No. 4, pp. 873–881, 2005.

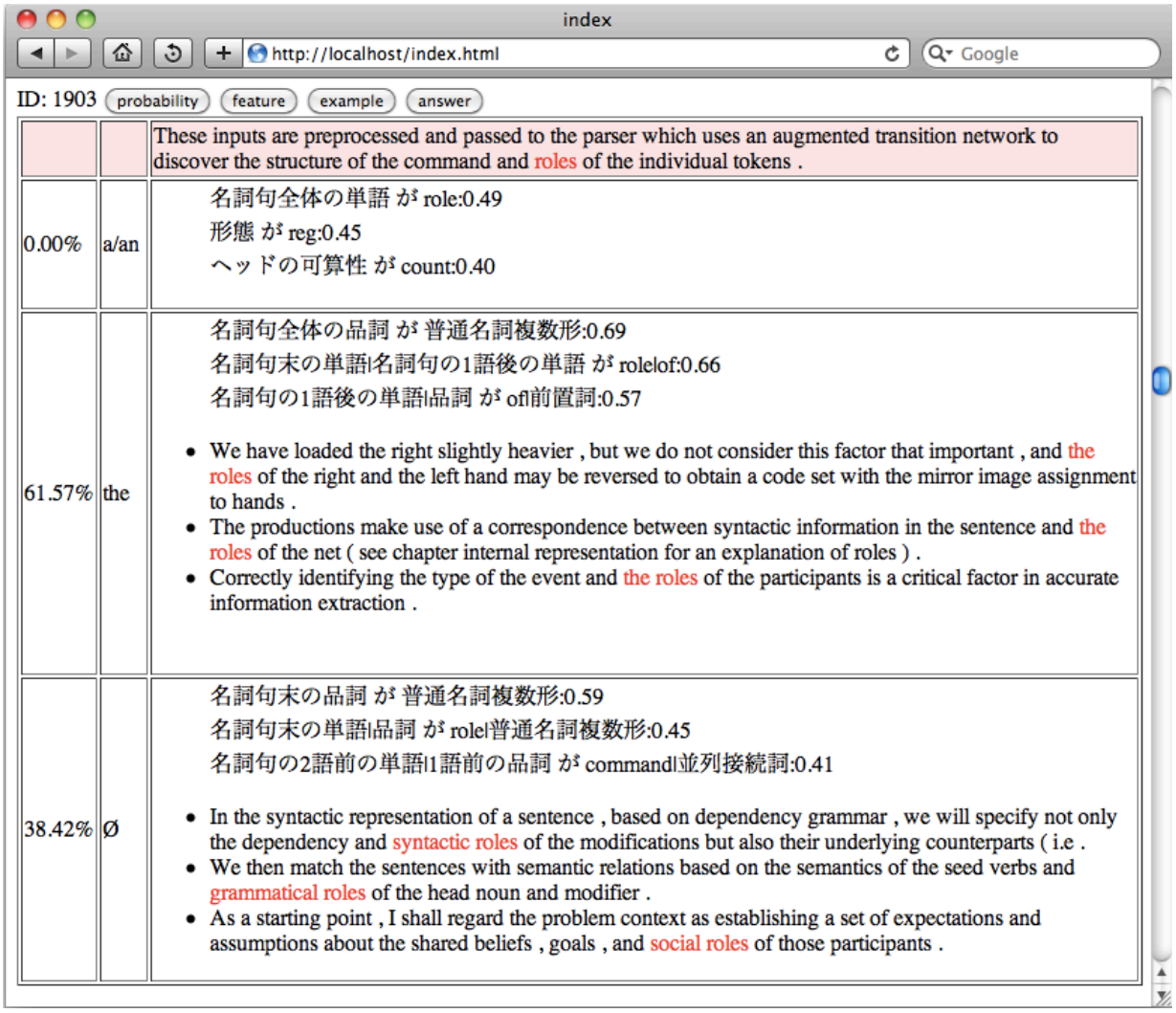

図 1: システム出力例

[2] Na-Rae Han, Martin Chodorow, and Claudia Leacock. Detecting errors in english article usage by non-native speakers. *Natural Language Engineering*, Vol. 12, pp. 115–129, 2006.

[3] 永田亮, 若菜崇宏, 河合敦夫, 森広浩一郎, 桝井文人, 井須尚紀. 可算/不可算の判定に基づいた英文の誤り検出. 電子情報通信学会論文誌. D, 情報・システム, Vol. 89, No. 8, pp. 1777–1790, 2006.

[4] 平野孝佳, 平手勇宇, 山名早人. 検索エンジンを用いた英文冠詞誤りの検出. 情報処理学会研究報告. データベース・システム研究会報告, Vol. 2007, No. 65, pp. 139–144, 2007.

[5] Robert Dale and Adam Kilgarriff. Helping our own: Text massaging for computational linguistics as a new shared task. In *INLG*, 2010.

[6] Martin Chodorow, Michael Gamon, and Joel Tetreault. The utility of article and preposition error correction systems for English language learners: Feedback and assessment. In *Language Testing 2010*, 2010.

[7] Kevin Knight and Ishwar Ch. Automated postediting article selection. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp. 779–784, 1994.

[8] 例文詳解技術英語の冠詞活用入門. 日刊工業新聞社, 2000.

[9] Edward Loper and Steven Bird. NLTK: the Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1*, pp. 63–70, 2002.

[10] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. Developing a robust Part-of-Speech tagger for biomedical text. In *Advances in Informatics*, Vol. 3746, pp. 382–392, 2005.

[11] Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.