# 英文冠詞誤りの自動校正手法における アプローチの違いによる傾向分析

乙武北斗　　荒木健治
Hokuto OTOTAKE　　Kenji ARAKI
{hokuto,araki}@media.eng.hokudai.ac.jp

北海道大学　大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

## 1. はじめに

英語非母語話者によって執筆された英文にはしばしば何らかの誤りが含まれる．その中でも特に冠詞の誤りの割合が多く，De Felice ら[1]が作成した英語非母語話者によって執筆された小規模コーパスに含まれる誤りのうち，17%が冠詞誤りであったことが報告されている．冠詞の用法には厳密な規則がない場合が多いため，辞書や用例から多くの事柄を調べる必要がある．このことから，冠詞誤りの校正には時間と労力，さらに専門知識も必要となる．

こうした現状を解決するために，これまで冠詞誤りを自動的に校正する手法がいくつか提案されている．それらの手法の中でも，最大エントロピー分類器を用いた手法が多く提案されている[1, 2, 3]．また，我々は意味カテゴリ情報に基づく帰納的学習（**S**emantic **C**ategory **B**ased **I**nductive **L**earning, 以降 SCB-IL と表記）によって生成されるルールを用いた冠詞誤り校正手法を提案している[4].

これら最大エントロピー分類器を用いた手法と，我々が提案した自動生成されるルールに基づく手法とでは，誤り校正のアプローチが大きく異なるため，校正結果にも違いが表れると考えられる. しかしながら，それぞれの手法において用いられている素性やトレーニングデータ，テストデータの種類および量が異なるため，各手法を同条件で比較することは困難である．

本稿では，我々が提案した SCB-IL による冠詞誤り校正手法[4]と同一の素性とトレーニングデータを用いた最大エントロピー分類器を構築し，冠詞付与実験を通して, SCB-IL と最大エントロピー分類器との比較を行う．また，実験結果からそれぞれの手法の特徴分析を行う．

以下，**2.** では用いる素性について，**3.** では性能評価実験と考察について述べる. 最後に **4.** でまとめを述べる．

## 2. 素性

SCB-IL による冠詞誤り校正手法[4]では，図 1 に示すような特徴スロットと呼ばれる素性集合を誤り校正ルールの要素として用いている．図 1 は，例文(i)における名詞 “*soccer ball*” から抽出された特徴スロットを表している．特徴スロットにおいて，名詞または動詞を表す要素は，単語と WordNet[1]から獲得されるカテゴリ情報の両方を保持する．

本稿で比較対象とする最大エントロピー分類器においても，図 1 に示す特徴スロットと同様の素性を利用する．

## 3. 性能評価実験

本章では，SCB-IL による冠詞誤り校正手法[4]と最大エントロピー分類器において，同一の素性とデータを用いた性能評価実験について述べる．

### 3.1 実験データ

本実験では，トレーニングデータとして Reuters Corpus[2]の英文記事約 2 億語を用いた．また，**2.**で述べた特徴スロットおよび素性ベクトルを抽出するために品詞タグ付けを行うツールとして，Brill’s Tagger[5]を用いた．最大エントロピー分類器は，Java 実装の一つである OpenNLP[3]の maxent パッケージを用いた．

[1] http://wordnet.princeton.edu/
[2] http://trec.nist.gov/data/reuters/reuters.html
[3] http://opennlp.sourceforge.net/

(i) This is the only *soccer ball* which I bought yesterday.

| | | |
|---|---|---|
| **Target** | Head | *ball* (noun.artifact) |
| | Preceding Noun | *soccer* |
| | Phrase | *NP* |
| | Preposition | - |
| | Preceding verb | *be* (verb.stative) |
| | Following verb | - |
| | Number | *singular* |
| | Proper noun | *no* |

| | | |
|---|---|---|
| **Preceding** | Modifier | *only* |
| | Modifier POS | *RB* |

| | | | |
|---|---|---|---|
| **Following** | **Preposition** | Preposition | - |
| | | Determiner | - |
| | | Nouns | - |
| | | Head | - |
| | | Modifier | - |
| | **Infinitive** | Verb | - |
| | | Determiner | - |
| | | Object | - |
| | | Adverb | - |
| | **Relative** | Subject | *I* |
| | | Verb | *buy* (verb.possession) |
| | | Determiner | - |
| | | Object | - |
| | | Adverb | *yesterday* |

*）要素 “–” は，該当する要素が存在しないことを表す.

図 1 特徴スロット

テストデータも同様に Reuters Corpus を用い，トレーニングデータとは別の 218,207 個の冠詞を含む英文とした. テストデータには冠詞誤りは含まれないため，本実験では各手法による出力がテストデータ中の冠詞と同じかどうかを評価した.

### 3.2 実験手順

本実験では，不定冠詞“*a*”，定冠詞“*the*”，無冠詞を表す“*none*”の 3 種類の冠詞について，それぞれ Precision (P)と Recall (R)を評価した．これら 2 つの評価尺度は以下の式(1)，(2)で定義される.

$$P = \frac{\text{正しく冠詞を提示した数}}{\text{冠詞を提示した数}} \quad (1)$$

$$R = \frac{\text{正しく冠詞を提示した数}}{\text{冠詞の総数}} \quad (2)$$

SCB-IL による冠詞誤り校正手法[4]は，対象名詞句に適用可能なルールが複数ある場合，校正候補として提示される冠詞も複数個になる場合がある．本実験では校正候補を一意に定めるため，最も高い優先度を有するルールが提示する冠詞のみを用いた．また，SCB-IL による手法においては, ルールの信頼度を表すスコアに関する閾値 $\theta$ が存在するが，本実験においては最も Precision が高い結果となった$\theta = 0.8$を用いた.

最大エントロピー分類器の出力においては，最も確率の高い冠詞を用いた.

### 3.3 結果と考察

表 1 に実験結果を冠詞の種類別に示す．また，表 2 にトレーニングデータにおける冠詞の種類別の分布状況を示す．表 1 におけるベースラインは，最も含有率の高い冠詞である無冠詞を常に提示するものである.

表 1 の結果, および表 2 の冠詞分布状況から, SCB-IL は最大エントロピー分類器と比較して，性能と冠詞の分布状況における相関が弱いことがわかる．トレーニングデータにおいて不定冠詞“*a*”は含有率が最も低い約 8%であるが，SCB-IL では定冠詞“*the*”よりも 10 ポイント以上高い Recall を達成している.

Precision においては，最大エントロピー分類器，SCB-IL ともにベースラインよりも優れた性能を達成している．また，SCB-IL と最大エントロピー分類器とを比較すると，すべての冠詞における Precision が上回っていることが確認された．一方 Recall においては，不定冠詞“*a*”を除き，SCB-IL の方が低い結果となった．SCB-IL による冠詞誤り校正手法においては，入力文に対して適用可能なルールを全く生成できなかった場合，校正候補を提示することができない．最大エン

表 1　実験結果

| | 冠詞 | Precision | Recall |
|---|---|---|---|
| ベースライン | *a* | - | 0.0% |
| | *the* | - | 0.0% |
| | *none* | 84.6% | 100.0% |
| OpenNLP maxent | *a* | 79.3% | 60.3% |
| | *the* | 77.0% | 70.4% |
| | *none* | 95.5% | 97.8% |
| SCB-IL[4] | *a* | 85.7% | 76.1% |
| | *the* | 89.1% | 65.3% |
| | *none* | 99.5% | 84.0% |

表 2　トレーニングデータの冠詞分布

| 冠詞 | 含有率（%） |
|---|---|
| *a* | 8.1% |
| *the* | 19.9% |
| *none* | 72.0% |

トロピー分類器の場合は，確率の高い候補を選択し，一意に冠詞を提示する．そのため，最大エントロピー分類器と比較して SCB-IL の Recall が低い結果となっていると考えられる．

これまで述べた各手法の特徴をまとめると，最大エントロピー分類器は，含有率の高い無冠詞に関しては非常に高い性能で提示することが可能である．一方で SCB-IL は, 全体として高い Precision を達成しており，含有率の最も低い冠詞である不定冠詞 "*a*" においては Precision，Recall ともに最大エントロピー分類器を上回る結果となった. これら 2 つの手法を組み合わせて，以下に述べるような手法を提案する．まず最大エントロピー分類器にて無冠詞かどうかの判断をし，無冠詞ではないという判断の場合には, SCB-IL による手法を用いて冠詞の最終判断を行う. このようにすることで，両手法の特徴を生かすことができ，冠詞誤り校正の性能を向上させることが可能であると考えられる．

## 4. まとめ

本稿では，我々が提案した SCB-IL による冠詞誤り校正手法と最大エントロピー分類器を用いた，英語冠詞誤り校正手法のアプローチの違いによる傾向分析を行った．実験の結果，最大エントロピー分類器の性能はトレーニングデータに含まれる冠詞の分布状況に依存することが明らかとなった．また，自動生成されるルールに基づく SCB-IL による手法は，全体的に高い Precision の結果であることを確認した．

今後は，本稿で用いた 2 種類の手法を組み合わせることで，冠詞誤り校正の性能向上を検証する．また，冠詞以外の文法誤り校正手法についても検討を行い，総合的な英文誤り校正手法の実現を目指したい．

## 参考文献


[1] R. D. Felice and S. G. Pulman, "A classifier-based approach to preposition and determiner error correction in L2 English," Proc. 22nd International Conference on Computational Linguistics (Coling 2008), pp.169-176, Manchester, UK (2008)

[2] E. Izumi, K. Uchimoto and H.Isahara, "SST speech corpus of Japanese learners' English and automatic detection of learners' errors," ICAME Journals No.28, pp.31-48 (2004)

[3] N. Han, M. Chodorow and C. Leacock, "Detecting errors in English article usage by non-native speakers," Natural Language Engineering, 12(1), pp.115-129

[4] Hokuto Ototake and Kenji Araki, "English Article Correction System Using Semantic Category Based Inductive Learning Rules," Springer-Verlag Lecture Notes in Artificial Intelligence (LNAI) Vol. 5866, pp.597-606 (2009)

[5] E. Brill, "Some Advances in Transformation-Based Part of Speech Tagging," Proc. The twelfth National Conference on Artificial Intelligence (vol. 1), pp.722-727, Seattle, Washington, USA (1994)