自動獲得されるルールに基づく英文冠詞誤り校正手法における最大エントロピー分類器の利用

*乙武 北斗　　**荒木 健治　　*吉村 賢治
*福岡大学 工学部
**北海道大学 大学院情報科学研究科
{ototake, yosimura}@.fukuoka-u.ac.jp
araki@.media.eng.hokudai.ac.jp

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誤りが含まれる．その中でも特に冠詞の誤りの割合が多いことが報告されている [1]．また，日本語のように冠詞を持たない言語を母語として持つ英語学習者は，冠詞誤りを起こす確率が高いことも報告されている [2]．

このような冠詞誤りを人手に頼らずに校正することを目的として，我々は自動獲得されるルールに基づく冠詞誤りの自動校正手法を提案した [3]．この手法では我々が独自に提案した，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）をルール生成に用いている．本手法の特徴としては，Precision が比較的高いこと，ユーザに校正理由を提示しやすいことなどが挙げられる．

一方で，最大エントロピー分類器による冠詞誤り校正手法もいくつか提案されている [1, 4, 5]．我々は同一素性を用いることで，SCB-IL と最大エントロピー分類器による冠詞誤り検出性能の比較および分析を行った [6]．その結果，最大エントロピー分類器による誤り検出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．英語の文章中には冠詞を伴わない名詞句の出現頻度が最も高いことが，様々な実験により報告されている．それゆえ，最大エントロピー分類器による冠詞誤り検出では，無冠詞の分類性能が非常に高い結果となっている．

本稿では，特徴の異なる両手法を融合することで，冠詞誤り校正における精度向上の可能性について検証を行う．冠詞を伴わない名詞句の分類精度が高い最大エントロピー分類器を用いて冠詞の有無を判断し，冠詞が必要と判断された名詞句については SCB-IL によるルールによって付与される冠詞を決定する．

以下，**2.** では冠詞の有無，および付与する冠詞の決定に用いる素性について，**3.** では性能評価実験について述べる．最後に **4.** で本稿のまとめを述べる．

2　冠詞選択の素性

我々が文献 [6] にて，SCB-IL と最大エントロピー分類器による冠詞誤り校正性能を比較した際，図 1 に示す素性を両手法で用いた．本稿においても，図 1 の素性を用いることとする．図 1 において，表の最も右の列の要素は素性値を表しており，例文 (i) の該当する値が入っている．素性値より左の要素は素性名および素性を分類するカテゴリ名を表している．

図 1 で表すように，各素性は 3 つのカテゴリに分類される．1 つ目は対象名詞句の特徴を表す “Target” カテゴリであり，主名詞，主語もしくは目的語とする動詞，単数／複数の情報などが含まれる．2 つ目は前置修飾語句を表す “Preceding” カテゴリである．3 つ目は後置修飾語句を表す “Following” カテゴリであり，対象名詞句を修飾する前置詞句，不定詞句，関係詞節の情報が含まれる．名詞や動詞については，単語そのもののほかに，WordNet[1] から獲得されるカテゴリ情報も素性として用いる．

SCB-IL による冠詞誤り校正手法 [3] では，トレーニングデータに出現する冠詞とその名詞句における素性ベクトルを組み合わせたものをルールとして用いている．ルールの素性ベクトルが，誤り校正対象の名詞句のものと一致した場合，ルールの適用が行われて校正候補が出力される．また，ルールはトレーニングデータから直接抽出されるだけでなく，SCB-IL による抽

[1] http://wordnet.princeton.edu/

(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: 素性リストと例

象化処理が行われ，より汎用性の高いルールが再帰的に生成される．この処理の詳細に関しては，文献 [3] を参照されたい．

3 性能評価実験

本章では，最大エントロピー分類器と SCB-IL の両方を段階的に利用した冠詞誤り校正の性能評価実験について述べる．実験では，比較のためにそれぞれの手法を単独適用した結果についても述べる．

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．素性ベクトル抽出のために品詞タグ付けを行うツールとして，Brill's Tagger[7] を用いた．最大エントロピー分類器は，機械学習アルゴリズムの実装の一つである Classias[8] の L1/L2 正則化ロジスティック回帰モデルを用いた．

テストデータはトレーニングデータとは別の Reuters Corpus 中の 48,325 個の冠詞を含む英文を用いた．テストデータには冠詞誤りは含まれないと仮定しているため，本実験では各手法による校正候補の出力がテストデータ中の冠詞と同一のものかどうかを評価した．

3.2 実験手順

本実験では，SCB-IL，および最大エントロピー分類器を単独で用いた冠詞誤り校正に加え，両手法を段階的に用いた誤り校正の評価を行った．両手法の段階的適用の流れを図 2 に示す．

入力として名詞句の素性ベクトルが与えられた際に，まず最大エントロピー分類器によって冠詞の有無を判別する．ここで，最大エントロピー分類器の分類結果を信用するかどうかを決定する指標として，スコアの閾値（$\theta \geq 0$）を考える．判別結果のスコア値が θ の負の値よりも小さかった場合，冠詞は必要ないと判断し，無冠詞を校正結果として出力する．スコア値が θ よりも大きかった場合，冠詞は必要であると判断し，SCB-IL による冠詞誤り校正手法にて定冠詞および不定冠詞のルールを適用し，結果を出力する．スコア値の絶対値が θ 以下だった場合は，最大エントロピー分類器による冠詞の有無の判断は考慮せずに，SCB-IL によるルールを用いて冠詞の判断を行う．

3.3 評価の指標

本実験では，名詞句全体に加え，不定冠詞 “a”，定冠詞 “the”，無冠詞の 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．

[2]http://trec.nist.gov/data/reuters/reuters.html

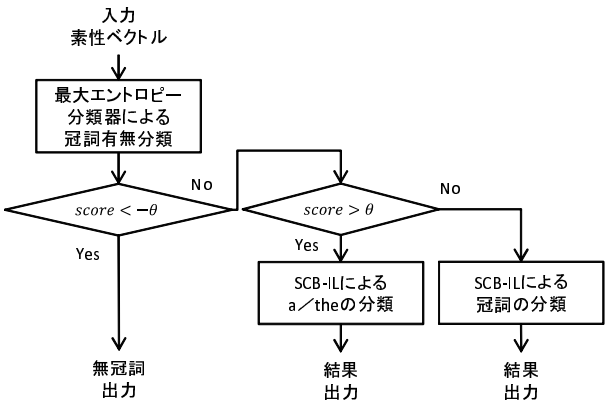

図 2: 段階的な誤り校正の流れ

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

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

SCB-IL によるルールを用いた冠詞の校正においては，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，適用可能なルールを一つも生成できなかった場合は校正候補を出力しない．

3.4 結果と考察

図 3 に，提案手法である SCB-IL と最大エントロピー分類器を段階適用した際の評価結果を示す．グラフの横軸は，最大エントロピー分類器におけるスコア値の閾値 θ を表している．図 3 より，θ の上昇とともに最大エントロピー分類器による冠詞の有無の判別を行わない事例が増加するため，SCB-IL による手法の特性である比較的高い Precision と低い Recall の傾向が強くなることが確認できる．本実験においては，$\theta = 1$ のときに最も Recall が高く，Precision と Recall の調和平均も最高値となった．また，$\theta = 2$ のときに最も Precision が高い結果となった．

表 1 に，提案手法において最も Precision が高かった $\theta = 2$ の結果，および単独の手法での結果を冠詞ごとにまとめたものを示す．表 1 より，最大エントロピー分類器単独ではトレーニングデータ中での含有率が 72% と最も高い無冠詞の Recall が最も良いことが

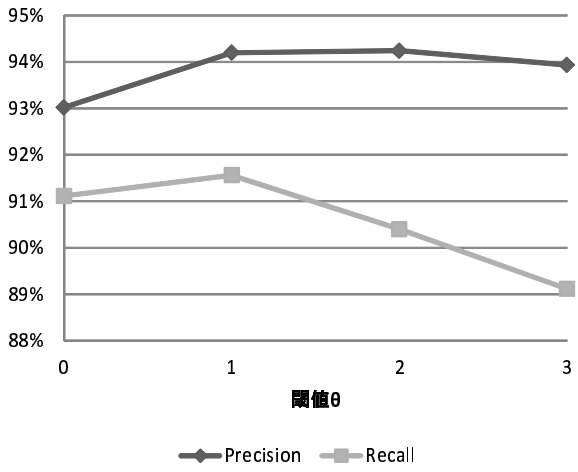

図 3: SCB-IL と最大エントロピー分類器の段階的適用による校正結果

確認できる．トレーニングデータおよびテストデータにおける冠詞の分布状況は表 2 に示すとおりである．また，SCB-IL 単独では含有率が 3 割未満と少ない定冠詞・不定冠詞の分類性能が Precision・Recall 双方において比較的高いことが確認できる．提案手法では，両手法の利点を継承できていると考えられる．SCB-IL 単独での性能と比較して，定冠詞・不定冠詞の分類性能の低下を 3 ポイント未満と小さく抑えつつ，無冠詞の分類性能を明確に向上させた．特に Recall においては 4 ポイントを超える改善が確認された．最大エントロピー分類器による冠詞の有無を判断させることは，特に無冠詞の Recall 性能向上に有効であったことが確認できた．提案手法においては，図 3 に示すように，最大エントロピー分類器のスコア値の閾値設定が性能に影響を与えるため，ユーザが期待する結果に応じた閾値設定が重要になると考えられる．

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段階的に適用させるよう融合した手法を提案し，冠詞誤り校正における精度向上の可能性について検証を行った．性能評価実験の結果，提案手法はそれぞれの手法を単独で適用するよりも高い性能（Precision= 94.24%，Recall= 90.40%）を達成することが可能となった．両手法の利点を適切に継承できたことが大きな理由として挙げられる．

今後は，実際に学習者による誤りが含まれる英文を対象に実験を行い，実用的な環境においても高い

表 1: 個別の冠詞の結果

システム	冠詞	Precision	Recall
最大エントロピー分類器	**a**	80.61%	62.89%
	the	77.87%	69.89%
	null	95.39%	97.92%
SCB-IL	**a**	90.73%	85.32%
	the	85.55%	79.57%
	null	94.96%	88.71%
提案手法 $\theta = 2$ （最大エントロピー ＋ SCB-IL）	**a**	90.71%	84.55%
	the	82.71%	77.00%
	null	96.35%	93.07%

表 2: 冠詞の分布

冠詞	トレーニングデータ	テストデータ
“a”	8.1%	6.5%
“the”	19.9%	13.2%
無冠詞	72.0%	80.3%

性能を発揮できるかを検証したいと考えている．また，このような手法の融合を冠詞だけでなく，前置詞誤りに代表されるその他の文法誤り校正手法についても適用を検討したい．現在，“http://hkt.tl.fukuoka-u.ac.jp/index.php”にて SCB-IL を用いた英文冠詞・前置詞誤りの校正手法のデモシステムを公開している．本稿で述べた提案手法も含めて，改善手法や新たな手法を継続して公開していきたいと考えている．

参考文献

[1] R. D. Felice and S. G. Pulman, “A classi er-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] C. Leacock, M. Chodorow, M. Gamon and J. Tetreault, Automated Grammatical Error Detection for Language Learners, Morgan and Claypool Publishers (2010)

[3] H. Ototake and K. Araki, “English Article Correction System Using Semantic Category Based Inductive Learning Rules,” Springer-Verlag Lecture Notes in Arti cial Intelligence (LNAI) Vol.5866, pp.597–606 (2009)

[4] N. Han, M. Chodorow and C. Leacock, “Detecting errors in English article usage by non-native speakers,” Natural Language Engineering, 12(2):115–129 (2006)

[5] M. Gamon, “Using mostly native data to correct errors in learners’ writing,” Proc. of NAACL, pp.163–171, Los Angeles, CA, USA (2010)

[6] 乙武 北斗，荒木 健治，“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”，言語処理学会第 16 回年次大会発表論文集，pp.415–417，東京 (2010)

[7] E. Brill, “Some Advances in Transformation-Based Part of Speech Tagging,” Proc. The twelfth National Conference on Arti cial Intelligence (vol.1), pp.722–727, Seattle, Washington, USA (1994)

[8] N. Okazaki, Classias: a collection of machine-learning algorithms for classi cation, http://www.chokkan.org/software/classias/ (2009)