事例の自動抽象化に基づくルールを用いた英語冠詞の自動付与手法の提案

*乙武 北斗　*吉村 賢治　**竹内 裕己　***河合 敦夫
*福岡大学工学部
**三重大学大学院工学研究科
***三重大学工学部
{ototake, yosimura}@fukuoka-u.ac.jp
{takeuti, kawai}@ai.info.mie-u.ac.jp

1　はじめに

日本人英語学習者にとって英語冠詞の用法は非常に難しいものである．日本語には冠詞が存在しないことに加え，冠詞の用法は様々な要因から影響を受けるため一意に決定することが難しいことが原因として挙げられる．

こうした現状を受け，冠詞の誤用を自動的に指摘・校正する手法は現在までに様々なものが提案されている．しかしながら，１００%の精度で校正を行う手法は実現されておらず，システムが誤った指摘を行う事例も十分考えられる．

そこで本稿では，校正結果に加え，校正理由や例文の提示を前提とした冠詞の自動付与手法を提案する．現状の手法では冠詞の完璧な自動付与は難しいため，複数の候補を理由付きで出力することにより，ユーザーが正しい結果を選択するように促す狙いがある．また，理由や例文の提示は，学習を目的とするユーザーにとって有用だと考えられる．

以下，**2.** で関連研究との比較について述べ，**3.** でルールについて述べる．**4.** で性能評価実験とその結果について述べ，**5.** でまとめと今後の課題について述べる．

2　関連研究

英語文法誤りの校正とともに例文を提示するシステムの一つに，Gamon の手法 [1] を実装している Microsoft Research ESL Assistant[1]がある．ESL Assistant は冠詞や前置詞など複数の誤りを対象に，主に統計的手法に基づく校正を行う．その出力の際，校正前・校正後の各事例が含まれる例文を，大規模言語モデルから検索して提示する．ユーザーは例文を正解の判断材料とすることができるほか，より深い文法の理解につながりやすくなるメリットがある．

本手法においても前述したメリットをもつが，例文の提示方法が Gamon の手法と異なる．本手法では，ネイティブ文章中の事例をベースに作成するルールを用いた冠詞の自動付与・校正を行う．そのため，校正前後の事例を含む例文だけでなく，校正に用いたルールを直接ユーザーに提示することが可能である．例文から用法を推測しなければならない場面において，ルールを直接参照することによって理解が明確になると考えられる．

本手法の詳細については，次章以降で述べる．

3　冠詞付与ルール

3.1　素性

冠詞付与ルールで用いる素性を表1に示す．表1では各素性をその機能で大きく4種類（対象名詞，前置修飾，後置修飾，既出）に分けている．素性名として使用されている “Synset” は，概念辞書の一つである WordNet[2]における同義語集合を表している．これらの素性値には1つの上位語 Synset が用いられる．たとえば主名詞 Synset の値には，主名詞が属する Synset と上位語の関係にある Synset が挿入される．多義語は複数の Synset に属するため，上位語 Synset も複数

[1] http://www.eslassistant.com/
しかしながら Web インタフェースの公開は 2011 年 4 月で終了したとのこと．

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

表 1: ルールに用いる素性

分類	素性名	抽象化グループ
対象名詞	主名詞 Synset	1
	単数／複数	1
	一般／固有名詞	1
	主名詞	2
	主名詞以外の名詞	3
	句の種類	3
	前置詞	3
	前に位置する動詞	3
	前に位置する動詞 Synset	3
	後ろに位置する動詞	3
	後ろに位置する動詞 Synset	3
前置修飾	修飾詞 Synset	1
	品詞	1
	所有格をもつ	1
	修飾詞	2
後置修飾	名詞句直後の語が属する句	1
	名詞句直後の語	1
	名詞句直後の語の品詞	1
	修飾句	2
	修飾句の主名詞	3
	修飾句の主名詞 Synset	3
	修飾句の主名詞以外の名詞	3
	修飾句の前置修飾詞	3
	修飾句の前置修飾詞 Synset	3
	修飾句の前置修飾詞の品詞	3
既出	既出かどうか	3

になり得る．本手法における複数の Synset を一つに絞る方法は **4.** で述べる．これらの素性とラベルとなる冠詞を組み合わせたものがルールとなる．

ルールの適用は，以下の 2 つのうちどちらか一方の条件を満たす場合に行われる．

- 適用対象の素性値とルールの素性値が完全に一致
- 適用対象の素性値がルールの素性値をすべて含む

3.2 事例の抽象化

本手法では，ネイティブによって書かれた英文の構文解析結果から名詞句を抽出し，素性および冠詞を組み合わせてルールとする．しかしながら，実例をそのままルールとするだけでは，ユーザーの様々な入力に対応することは難しい．特に多くの素性値を含むルールは，適用可能な対象が大きく制限される．

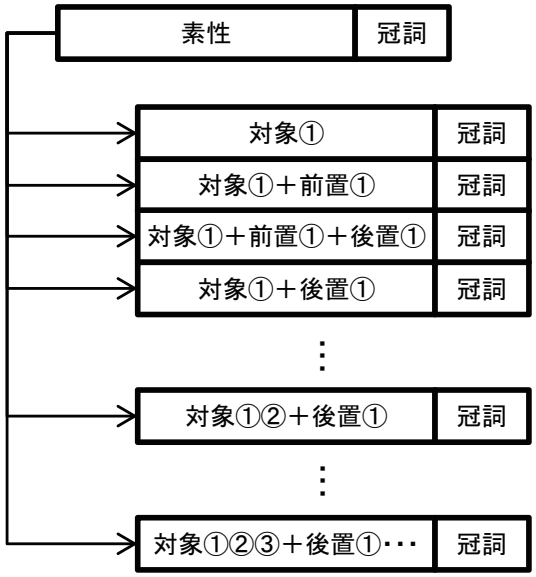

図 1: 事例の抽象化

そこで，本手法では多くの素性値を含むような事例から汎用的なルールを作成するために，事例の抽象化を行う．図 1 に事例の抽象化の概要を示す．図 1 の最上部に示す素性と冠詞の組み合わせは，抽象化の基となる事例を示している．基事例から伸びている矢印の先にある複数の事例が，抽象化された事例を表している．

図 1 の抽象事例の素性部分に示されている “対象①” は，基事例から継承した素性の分類名とグループ番号を表している．各素性の分類とグループ番号は表 1 にあるように設定した．抽象化の際には，すべての素性分類とグループ番号の組み合わせについて抽象事例を作成する．ただし，大きいグループ番号を用いる際はそれより小さなグループ番号を包含する．すなわち，“対象②” 単体ではなく，“対象①②” のように必ず自身より小さなグループ番号とセットで用いる．

4 性能評価実験

4.1 実験データ

本実験では，ルール作成用データとして Reuters Corpus[3] の英文記事約 2000 万語を用いた．素性ベクト

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

ル抽出のために構文解析を行うツールとして，Stanford Parser[2] を用いた．

また，作成したルールの信頼度スコアを計算するために，同じく Reuters Corpus の英文記事約 200 万語を用いた．信頼度スコア計算については，次節 **4.2** で述べる．

テストデータはトレーニングデータとは別の Reuters Corpus 中の 1,258 個の冠詞を含む英文を用いた．Reuters Corpus には冠詞誤りは含まれないと仮定しているため，本実験では最も高いスコアを持つルールの冠詞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．

4.2 実験手順

本実験を始める前に，まずルールの作成，および信頼度を表すスコア計算を行う．ルール作成用データから抽出・抽象化を経て得られたルールを対象に，スコア計算用データを用いてスコアを計算する．スコアは以下の式によって定義される．

$$Score = \frac{\text{正適用回数}}{\text{適用回数}} \tag{1}$$

式 (1) の適用回数は，スコア計算用データに対して **3.1** で述べた条件を満たし，ルールの適用が可能だった回数を表す．正適用回数は，ルールが適用可能なことに加え，ルールが示す冠詞とスコア計算用データの冠詞が一致した回数を表す．したがって，高いスコア値を持つルールは信頼性が高いと言える．

ルールが作成された直後は，適用回数，正適用回数ともに 1 であるため，スコアも 1 となる．この場合のスコア値は信頼できる値とは考えられないため，本実験では適用回数が 2 以上のルールのみを用いて評価を行うこととした．

また，**3.1** で述べた上位語 Synset の決定方法による性能の差異を確認するため，以下に示す 3 つの方法を用いて実験を行った．

- #0：Synset 素性を用いない
- #1：WordNet の辞書中で常に先頭に記述されているものを選択
- #2：対象文のトピック情報を用いた Synset の選択 [3]

#2 の手法では，Reuters Corpus の記事データに含まれるトピック情報を用いて，そのトピックで出現しやすい Synset を推定して選択する．

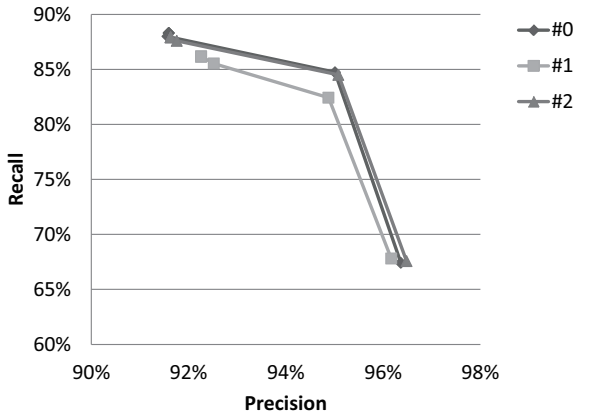

図 2: Precision-Recall グラフ

4.3 評価の指標

本実験では 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 2，3 で定義される．

$$P = \frac{\text{正しく冠詞を提示した数}}{\text{冠詞を提示した数}} \tag{2}$$

$$R = \frac{\text{正しく冠詞を提示した数}}{\text{冠詞の総数}} \tag{3}$$

本手法で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，最も高いスコア値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．

4.4 結果と考察

図 2 に，用いるルールのスコア値に対して 0 から 1 まで 0.2 区切りで閾値を設定した際の，Precision-Recall グラフを示す．各グラフの最も左上の点が閾値 0 を表しており，最も右下の点が閾値 1 を表している．

表 2 に，最も Precision が高かった閾値 1 の場合の，各手法における冠詞別の評価結果をまとめる．どの手法の結果においても，無冠詞の付与性能は著しく高い．これは，Reuters Corpus 中に出現する無冠詞の比率が約 8 割であり，無冠詞を付与するルールが他と比べて著しく多いことが原因として考えられる．

また，Synset 選択の手法の違いに着目すると，トピック情報を用いた Synset の選択を導入することで，

表 2: 個別の冠詞の結果

Synset の選択	冠詞	Precision	Recall
#0	**a**	66.67%	29.41%
	the	57.78%	24.53%
	null	99.02%	72.63%
#1	**a**	62.50%	29.41%
	the	55.32%	24.53%
	null	99.15%	73.08%
#2	**a**	66.67%	29.41%
	the	60.00%	25.47%
	null	99.03%	72.72%

定冠詞の付与精度を向上できることが確認された．しかしながら本実験ではテストデータの数が約 1,200 と少なく，その約 8 割は無冠詞の名詞句であるため，特に不定冠詞・定冠詞の付与性能検証は不十分であると考えられる．

5 まとめと今後の課題

本稿では，事例の自動抽象化に基づくルールを用いた英語冠詞の自動付与手法を提案した．実験の結果，最も Precision が高い結果で，Precision=96.48%, Recall=67.57% となった．また，ルールのスコア値に対する閾値を操作することで，Precision と Recall の優先度を変化させることが可能である．

今後は，ルール作成用とスコア計算用，およびテスト用のデータを増加させたうえで，再度評価実験を行うことを予定している．そこで既存の手法と同程度以上の結果が確認されれば，本手法の特徴であるユーザーへのルール提示に関して，有効性を検証する実験を行いたいと考えている．

参考文献

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

[2] D. Klein and C. D. Manning, “Accurate Unlexicalized Parsing,” Proc. of ACL 2003, pp.423–430 (2003)

[3] H. Takeuchi, H. Miyake, A. Kawai, R. Nagata and H. Ototake, “Extension of Phrases for Article Determination using WordNet Thesaurus,” Proc. of 6th International Global WordNet Conference, pp.349–356, (2012)