定冠詞の前方照応用法を考慮した冠詞誤り訂正

吉本 一平　　小町守　　松本裕治
奈良先端科学技術大学院大学 情報科学研究科
首都大学東京システムデザイン学部
{ippei-y, matsu}@is.naist.jp
komachi@tmu.ac.jp

1　はじめに

冠詞誤りは英語学習者が犯す文法誤りの中でも高い割合を占めている．冠詞誤りを含む限定詞誤りは，HOO2012[1] で評価コーパスとして用いられた Cambridge Learner Corpus では全ての誤りの 11.7 %を占め [6]，CoNLL2013 Shared Task[2] で評価コーパスとして用いられた NUS Corpus of Learner English [2] では 12.9 %を占めている．

冠詞には不定冠詞 “a” (“an”), 定冠詞 “the”, 無冠詞[3] がある．定冠詞は聞き手，読み手が名詞の指示対象を同定可能であるときに使用され，不定冠詞は聞き手，読み手が名詞の指示対象を同定可能でないことを前提に使用される．

Halliday and Hasan [4] は指示対象を同定するための手がかりが文章外の状況であるのか，それとも文中に存在するかによって，“the” の用法を外界照応とテキスト内照応に分類した．さらに，テキスト内照応を指示対象の同定のための手がかりが先行するテキスト内にあるか，それとも後ろにあるかによって，前方照応，後方照応に分類している．

外界照応の用法においては，(１) ある特定の個体や部分集合が指示されており，その個体や部分集合が特定の状況において同定可能 (２) 指示対象が言語外の根拠によって状況によらず同定可能，という基準のいずれかによって指示対象が同定可能である．例えば下の例文は (２) の場合であり，世界 (world) は１つしか存在しないという知識により同定可能である．

> There are millions of doctors all over **the world**...

後方照応の用法[4] “the” は，名詞句の主辞の指示対象がその修飾語によって同定可能であることを示す．以下の例の “productivity” は “of” 以下の句によって限定されていることで同定が可能である．

> ... increase **the productivity** of the corporation ...

前方照応の用法においては，前に出てきたものに対してもう一度言及するときに “the” が使われる．このとき，指示対象は先行詞と同じものを指すとは限らない．同じである場合は，同じ単語や同義語，類義語によって，また違うものである場合は，常識を用いて先行詞との照応関係を推測することができるような表現によって照応がなされる．これは**橋渡し指示** (bridging reference) と呼ばれる．

(1) ... there is **a risk** of causing harm ... people still prefer to bear **the risk** ...

(2) ... his **location** ... other auto-tracking devices to find **the place** ...

(3) ... in **school** ... This can protect the **students**

(1), (2) の例では，“the” が付属する表現は前の表現と同じ対象を指示している．(1) の例では同じ単語，(2) の例では同義語によって照応がされている．また (3) は橋渡し指示の例である．

英語の冠詞訂正の先行研究では，冠詞やその主辞となる名詞の周辺の単語やそれらと依存構造木上で近くにある単語といった局所的な情報を手がかりに訂正が行われてきた．外界照応ではその名詞の主辞の性質が，また後方照応では “the” に後続する単語が手がかりとなるため，そのような局所的な情報を用いることで基本的に対処が可能だと考えられる．しかし，前方照応

[1] http://clt.mq.edu.au/research/projects/hoo/hoo2012
[2] http://www.comp.nus.edu.sg/~nlp/conll13st.html
[3] 本研究では，名詞の主辞に冠詞が付かない場合を無冠詞と呼び，便宜上冠詞に無冠詞を含める．
[4] ここでは文献 [4] にならい後方照応を構造上のものに限定した

では先行詞が先行する文や，同じ文中でも離れた場所にあることが多いため，そのような局所的な情報のみでは十分に対処できない．

そこで，本研究では共参照解析，照応解析の分野で用いられてきた手がかりを用いることで，前方文脈を考慮しつつ冠詞の訂正を行う．

2 関連研究

英語の冠詞誤り訂正の先行研究では，分類器を用いた手法 [3] ，統計的機械翻訳を用いた手法 [8] などが提案されてきた．分類器ベースの手法では，冠詞訂正を無冠詞，不定冠詞，定冠詞の 3 クラスへの分類問題として扱い，分類結果を訂正として提示する．素性には窓を限定した局所素性，依存構造木素性などが用いられる．統計的機械翻訳を用いた手法では，誤り文から正しい文へ翻訳することで訂正を行う．この際，手がかりとしては単語の表層が用いられることが多い．

本研究と類似した研究に竹内ら [12] がある．彼らは，定冠詞の前方照応用法を考慮した冠詞予測手法を提案し，ネイティブコーパスにおいて冠詞が the かそれ以外かを推定する実験を行っている．彼らの手法では，冠詞を推定する際に前方文脈に出現する名詞を手がかりとして用いる．その際に，全ての名詞を用いるのではなく，対象名詞と共起するときに対象名詞に the が付く確率が高いもののリストをあらかじめ作成しそこに含まれるもののみを用いている．

本研究においては，前方照応の用法の考慮のために共参照解析や照応解析の分野で用いられる手がかりを導入する．共参照解析は同じ指示対象を指す表現を同定するタスクである．共参照解析においては，２つの表現の主辞の表層形が同じかどうかを示す手がかりが用いられるほか，同義語，上位語といった単語間の関係に関する情報を得るために，WordNet[5]，YAGO[6] といった知識ベースが用いられている [11] ．照応解析においては，共参照関係に限らず，広く前方照応の関係にある名詞句の同定が目標とされる．部分全体関係，上位下位関係を持つような単語は照応関係を取りやすいためこうした単語同士の語彙的な関係の獲得を目的として A of B のようなパターンが用いられる [10] [5] ．本研究では上で挙げたものに類似した手がかりを用いる．

[5]http://wordnet.princeton.edu
[6]http://www.mpi-inf.mpg.de/yago-naga/yago

3 定冠詞の前方照応用法を考慮した冠詞誤り訂正

本研究では，名詞句の主辞となる名詞を対象として，それに付くべき冠詞を分類器によって決定することで訂正を行う．分類のクラスとしては，無冠詞，不定冠詞，定冠詞の３クラスを考える．提案手法では前方照応の用法に対処するために，照応関係を反映した素性（**照応素性**） を導入する．以下では本研究で新たに用いる素性について説明する．ここで用いる例文は全て NUCLE で見られた学習者の書いた文であり．また抽出する素性を例文の下に示した．[7]

3.1 単語の原形の一致

先行する文章内に，訂正対象の名詞と同じ原形を持つ単語があれば，その単語の原形を素性として用いる．これにより同じ単語による照応を考慮する．

... surveillance **technology**$_1$ violates the right since it leaks people's information without getting their permission. When ***the technology**$_2$ is not so well developed ...

$\longrightarrow$*technology*$_1$*_anaphora_lemma*

3.2 WordNet の同義語，下位語

WordNet は大規模な語彙データベースであり，各単語がそれの表す概念によって Synset というグループに分類されている．また，異なる Synset は互いに，意味的または語彙的な観点から関連付けられており，上位下位や部分全体の関係にある単語の情報を取得することができる．

ここでは，先行する文章内に，訂正対象の名詞と同じ Synset に属する単語，もしくは距離が 2 階層以内の下位語があれば，その単語の原形と対象の名詞の原形の対を素性として用いる．これにより同義語，上位語による照応を考慮する．

.. many of the cell **phone** are equipped with GPS which means that owners can find ***the device**'s location...

$\longrightarrow$*phone_device_anaphora_WordNet*

[7]訂正対象の名詞句の主辞には*で印を付けた．また，冠詞の訂正が必要な箇所には訂正前と訂正後を示した．

3.3 ConceptNet5 の関連語

ConceptNet5 [7] は人間の知識全般，またそれらの自然言語による表現を記述する大規模な意味グラフである．単語や短いフレーズが，語彙的な関係だけではなく，常識的な関係によって関連付けられており，WordNet からは取得できないような情報を利用することができる．以下に例を挙げる．

learn⟶ MotivatedByGoal ⟶ knowledge

school⟶ Related ⟶ student

ここでは，先行する文章内に，訂正対象の名詞との間に ConceptNet5 のグラフにおいて辺が存在する単語があれば，その単語の原形と対象の名詞の原形の対を素性として用いる．これにより同義語，上位語，また橋渡し指示による照応を考慮する．

.... For example, intruders can be detected if there are any in **schools** or offices. This can protect ***students** from being in a dangerous situation, ...

⟶*student_school_anaphora_conceptNet*

訂正 [**students** →**the students**]

3.4 A of B パターン

全体部分関係にあるような単語同士は照応関係を取りやすい．この関係にある単語を手がかりとして用いるために A of B のパターンを用いる．

ネイティブコーパスにおいて「名詞句 A (A) 」of「名詞句 B (B) 」の主辞のペアに対してその頻度を求める．分類の際は，先行する文章内に訂正対象の名詞句を「名詞句 A」としたとき,「名詞句 B」となるような名詞句で，上で求めた頻度が一定以上であるものがあれば，その主辞の対を素性として用いる．これにより上位語，また橋渡し指示による照応を考慮する．

Apart from preventing **crime** and terrorism, RFID can also be used as an evidence or alibi to help determine whether ***the suspect** is the criminal or not.

⟶*crime_suspect_anaphora_AofB*

4 実験設定

訂正の対象としては，主辞が所有格に修飾されるもの，冠詞以外の限定詞に修飾されるものを除く名詞句を扱った．照応素性の抽出の対象は訂正対象の名詞句と同じ文中で訂正対象より前にある名詞及びその文の前の 5 文中の全ての名詞とした．これは，文献 [5] において名詞句の先行詞のうち 81 %が前の 5 文以内に存在すると報告されていることを参考とした．ベースラインの素性としては窓を限定した局所素性と依存構造木素性を用いた．窓を限定した局所素性としては，名詞句の主辞の後ろの窓幅 3 の中の単語の表層の 1,2,3-gram，及び名詞句に冠詞が付いている場合は冠詞の前後の 2 単語，冠詞が無い場合は冠詞が挿入されうる位置を中心とした窓幅 4 の中の単語の表層形の 1,2,3-gram を用いた．依存構造木素性は名詞の主辞と依存構造木上の距離が 2 以内にある単語及びそれらの間の辺の依存構造木のラベルを用いた．また，ネイティブコーパスで訓練を行うためテストの際も元の冠詞の情報は用いなかった．提案手法ではこれらの素性に加えて，前節で説明した素性を用いた．

訓練コーパスには Wikipedia[8] を用いた．Wikipedia は 2013 年 9 月 4 日のダンプから取得した名詞句 2 千万個を用いた．評価は CoNLL2013 Shared Task の評価に用いられた NUCLE2.3.1 の 1381 文と Konan-JIEM Corpus 第 3 版 (KJ) [9] の 2414 文で行った．WordNet は version 3.0 を用いた．A of B パターンは Gigaword Corpus version 5 の AFP, APW, NYT[9] と Wikipedia の上記と同じものにおける頻度が合計で 5 以上のもの 474083 個を用いた．依存構造解析には Stanford CoreNLP v3.2.0[10] を用いた．この際依存構造木の表現としては，Basic Typed Dependency を用いた．また，分類器には最大エントロピーモデルを用い，正則化は L1 正則化で行った．実装は独自のものを用いた．評価尺度には MaxMatch [1] による Precision, Recall, F-score を用い，実装としては，CoNLL2013 Shared Task で配布された m2scorer を用いた．

5 実験結果・考察

表 1 はベースライン，ベースラインに原形の一致による照応素性を加えたもの，全ての照応素性を加えたもの，全ての照応素性から WordNet, ConceptNet を除いたもの，また，A of B パターンを除いたものの結果をそれぞれ示している．

実験の結果，原形の一致による照応素性を加えることにより，Precision，Recall とも向上することがわかった．また，全ての照応素性を用いることで更に性能の向上が見られた．ただそこから WordNet，ConceptNet

[8] http://www.wikipedia.org
[9] LDC2011T07
[10] http://nlp.stanford.edu/software/corenlp.shtml

表 1: 実験結果 （トレーニング:Wikipedia，テスト:NUCLE, KJ）

評価尺度	照応素性なし	＋原形	全て	全て −WordNet, ConceptNet	全て −A of B
Precision	34.3	34.4	34.6	34.7	33.8
Recall	52.8	53.2	53.6	53.2	52.6
F-score	41.6	41.8	42.0	42.0	41.2

表 2: 照応素性の発火比率

素性	頻度
原形の一致	0.253
WordNet	0.446
ConceptNet	0.251
A of B	1.24

表 3: 訂正のタイプごとの TP の数

訂正前＼訂正後	無冠詞	定冠詞	不定冠詞
無冠詞		88(91)	28(27)
定冠詞	211(205)		16(15)
不定冠詞	21(20)	6(6)	

を除いても精度は変わらなかった．また，全ての照応素性から A of B パターンを除いたものはベースラインのより性能が下がっている．表 2 は NUCLE 評価コーパスにおける照応素性の 1 事例あたりの発火頻度を示している．ここから A of B パターンの発火比率が他の照応素性と比べて高いことが見て取れる．表 3 は訂正のタイプごとの True Positive の数を示している．表中の左の数値が照応素性を全て用いたとき，括弧の中の数値は照応素性を用いないときの結果を示している．

今回，照応素性を加えることにより，性能の向上が見られたが，改善は小さなものであり，十分に照応関係を考慮できているとは言いがたい．

- For example , they may wait in **a public area** and wait for ... ...this kind of agency may bring the child to ***the mountain area**

上の例文では，システムは mountain area の冠詞を誤って the と判定している．この原因として public area と mountain area が照応関係にないにも関わらず，この 2 つから抽出された原形の一致素性が the に大きい値を付けたことがある．このような例を考慮するためには，名詞の主辞のみを用いるのではなく，他の素性と組み合わせて用いることが必要だと考えられる．

6 おわりに

本研究では，定冠詞の前方照応用法を考慮した冠詞誤り訂正の手法を提案した．実験の結果，照応関係を反映するための素性を加えることで，冠詞訂正の精度が Precision，Recall ともに向上することがわかった．また大規模コーパスから獲得したパターンを用いて，同じ単語以外による照応を考慮することにより，さらに誤り訂正性能が向上することがわかった．

参考文献

[1] Daniel Dahlmeier and Hwee Tou Ng. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012.

[2] Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 2013.

[3] Rachele De Felice and Stephen G. Pulman. A classifier-based approach to preposition and determiner error correction in l2 english. In *Proceedings of the 22nd International Conference on Computational Linguistics*, 2008.

[4] M.A.K Halliday and Ruqayia Hasan. *Cohesion in English*. Longman, 1976.

[5] Yufang Hou, Katja Markert, and Michael Strube. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013.

[6] Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool Publishers, 2010.

[7] Hugo Liu and Push Singh. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*, Vol. 22, pp. 211–226, 2004.

[8] Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of the 26nd International Conference on Computational Linguistics*, 2012.

[9] Ryo Nagata, Edward Whittaker, and Vera Sheinman. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 2011.

[10] Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. Learning to resolve bridging references. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics, Main Volume*, 2004.

[11] Altaf Rahman and Vincent Ng. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[12] 竹内裕己, 河合敦夫, 細田直見, 永田亮. 前方文脈を考慮した冠詞の推定. 言語処理学会 第 19 回年次大会, 2012.