前置詞誤り検出／訂正のための誤り格フレームの生成

永田　亮†　水本　智也††　EdwardWhittaker†††
† 甲南大学知能情報学部　†† 奈良先端科学技術大学院大学　††† Inferret Limited
E-mail: †rnagata@ konan-u.ac.jp., ††tomoya-m@is.naist.jp, †††ed@inferret.co.uk

1. はじめに

非母語話者にとって英語の前置詞を正しく使用することは難しい．前置詞の用法は複雑であり，文脈に応じた適切な前置詞を選択することは困難を伴うことが多い．実際，非母語話者の書く英文には前置詞の誤りが多いことが報告されている．例えば，日本人英語学習者の英文を収録した Konan-JIEM Learner Corpus(注 1) では，前置詞誤りは 3 番目に多い誤りであり，文法誤り全体の 14%を占める．

このような背景を受けて，前置詞誤りを検出／訂正する手法が数多く提案されている（文献 [1], [2], [5] など）．現在主流であるのは，前置詞誤りの検出／訂正を分類問題として解く手法である．すなわち，与えられた前置詞の正誤を 2 値分類問題として解く手法，または，正しい前置詞を選択する n 値分類問題として解く手法である．どちらの場合も，前置詞周辺の情報（単語や品詞など）を素性として用いることが一般的である．

従来手法の大きな問題として，分類器の検出／訂正規則を解釈することが難しいという点を挙げることができる．そのため，専門家の言語知識を手法に直接反映することは困難である．また，この問題に関連して，検出／訂正に対する説明を与えることも難しい．このことは，学習支援を目的とした誤り検出／訂正では特に問題となる．従来手法では，正しい前置詞を選択できたとしても，なぜその前置詞が正しいのかを説明することは難しい．例えば，英文 “I walked with my dog in the morning. ” の下線部の前置詞が不要であることを特定できたとしても，その理由を説明することは難しい．一方，英語教師であれば，「“walk with a dog” では，犬と一緒になって犬のように歩く様子を想起させるので，犬を散歩させるという意味の場合は，“walk a dog” が自然である」というような説明を学習者に与えるであろう．

そこで，本稿では，これらの問題を解決する枠組みを提案する．具体的には，(i) 解釈が容易，(ii) 言語知識を直接反映可能，(iii) フィードバックメッセージが付与可能という特徴を備えた枠組み「誤り格フレーム」を提案する．誤り格フレームは，その名の通り格フレームに基づく．ただし，通常の格フレームとは異なり，前置詞誤りを記述するための格フレームである．また，前置詞誤りの検出／訂正を目的とした格フレームである．具体例を図 1 に示す．詳細は，2. に譲るが，図 1 (a) が “He will go back _ Japan.” や “They went back _ Tokyo yesterday.” などの前置詞誤りに対応していることが，直観的に解釈できるのではないかと思われる．また，図 1 (b) に，人手で情報を追加した誤り格フレームを示す（赤字が人手による追加情報を示す）．この例のように，誤り格フレームでは，人間が解釈し，情報を追加することが可能である．更に，本稿では，誤り格フレームをコーパスから自動生成する手法も提案する．生成に必要となるのは，母語話者コーパスと非母語話者コーパスのみである．ただし，パラレルコーパスである必要はなく，誤りの情報も必要としない．

(注 1)：http://www.gsk.or.jp/catalog/GSK2012-A/catalog.html

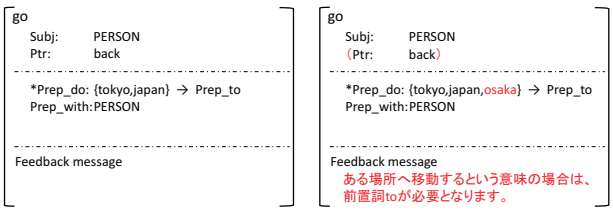
2. 誤り格フレーム

誤り格フレームの中心となるのは動詞である．したがって，誤り格フレームは必ず動詞を持つこととする．図 1 の例では，左上に記述されている “go” が動詞になる．

動詞以外に，誤り格フレームは三つの部分から成る．図 1 では破線で分けられた部分に対応する．上から順に，基本格，前置詞格，フィードバックメッセージと定義する．基本格と前置詞格は，動詞がどのような表層格を取るかを記述する部分である．特に，前置詞格は誤り情報を記述するという意味において重要である．フィードバックメッセージは前置詞誤りに関する説明を記述する部分である．

各部分の詳細な説明を行う前に，誤り格フレームで使用される用語の定義を行う．主格などの格の種類を表すためのラベルを格標識と呼ぶことにする．図 1 では，“Subj:”, “Prep_do:” などが格標識である．また，格標識が付与される語を格要素と呼ぶことにする．例えば，“Subj:” の格要素は “PERSON” である．更に，特に断らない限り，格とは，格標識と格要素を合わせたものを指すことにする．

以上の用語を用いて，基本格は一つ以上の格から成ると定



(a) 自動生成された誤り格フレーム　(b) 修正を加えた誤り格フレーム

図 1：　誤り格フレームの例.

義する．基本格に入る格標識は，“Subj:”（Subject），“Prt”（Particle），“Com”（Complement）の三種類とする(注 2)．このうち，“Subj:” は必須であるとする．

前置詞格も一つ以上の格から成ると定義する．基本的には，動詞が取りうる前置詞を記述する．具体的には，格標識 “Prep_x” を用いて記述する．ただし，x の部分には前置詞が入る（例：Prep_to）．また，動詞の直接目的語（Prep_do）と間接目的語（Prep_io）も便宜的に前置詞格に含める．これは，前置詞の抜けと余剰に対応するためである．以上に加えて，前置詞格に誤り情報を記述する．前置詞格の中で，誤りがある格に “*” を付与する．ただし，一つの誤り格フレームでは，誤りとなる格は一つとする．例えば，図 1（a）では，“*Prep_do:{tokyo,japan}” の部分が誤りとなる格である．言い換えると，直接目的語として “tokyo” や “japan” を取ることは誤りであり，何らかの前置詞が必要なことを意味する．更に，誤りである格の後ろに，訂正情報を “→” を用いて記述する．例では，“→ Prep_to” の部分である．

基本格と前置詞格で，共通して使用される記述方式を 2 種類定義する．一つは，任意格に関する記述である．任意格は，括弧 “ () ” を用いて表す（例：図 1（b）の “（Ptr:back）”）．二つ目は，格要素が複数ある場合の記述である．その場合，複数の格要素をカンマで区切り，中括弧で囲うこととする（例：*Prep_do:{tokyo,japan}）．

フィードバックメッセージは，前述の通り，前置詞誤りに関する説明を記述する部分である．誤り格フレームを解釈し，人手で記述する．この説明は，誤り検出／訂正の際に，学習者へのフィードバックなどに使用することができる．

3. 誤り格フレームの生成

誤り格フレーム生成の基本アイデアは非常にシンプルである．非母語話者コーパスに存在し，母語話者コーパスには存在しない格フレームを誤り格フレームとするものである．ただし，このシンプルなアイデアのみでは，正しい格フレームが誤り格フレームとして抽出されてしまう．そのため，誤り格フレームのみをいかに選択するかということが重要となる．

具体的な処理の流れは，次の通りである：

（1） 格フレームの生成
（2） 格フレームの統合
（3） 誤り格フレーム候補の取得
（4） 訂正情報の決定
（5） 格要素の拡張
（6） 誤り格フレームの出力

（1）と（2）については，格フレームを自動生成する手法 [3] を参考としている．ただし，文献 [3] とは研究の目的が異なるため，処理の詳細についても異なる部分がある．以下，各処理について詳しく説明する．

「（1）格フレームの生成」では，コーパスから通常の格フレームを生成する．前処理として，コーパスの各文を構文解析する．構文解析の結果から，格フレームの各スロットを埋めることで格フレームを生成する．すなわち，動詞および格を対応する箇所に配置する．その際，格要素には対応する名詞相当句の主辞（head）を小文字かつ原形にしたものを用いる．ただし，接尾辞 “-ing” は前置詞の決定に影響を与えることがあるため，語尾が “-ing” である語については原形にしない．更に，一部の語については，対応する意味を表す特別な語に置換する．例えば，“he” は人を表す “PERSON” に置換する（本稿では，意味を表す特別な語は大文字のみを用いて表記することにする）．この置換は単純な辞書引きに基づいて行う．以上の処理を母語話者コーパス，非母語話者コーパス，それぞれについて行う．以下，母語話者コーパス／非母語話者コーパスから抽出された格フレームの集合を母語話者格フレーム／非母語話者格フレームと呼ぶことにする．

以上が「格フレームの生成」であるが，次の三つの条件のいずれかに当てはまる場合には，例外として格フレームの生成を行わない．一つ目は，動詞が接続詞により並列されている場合（例：go and get it）である．これは，並列により前置詞の用法が変更されることがあるためである．二つ目は，“be”，“do”，“have” は助動詞としても使われる特殊な動詞であるため例外とする．最後に，格要素が，“it”，“this”，“that”，“one”，および，通常名詞の働きをしない単語（例：the）である場合も例外とする．“it”，“this”，“that”，“one” は，具体的に指すものにより格の用法が異なると考えられるためである．残りについては，構文解析の誤りの可能性が高いためである．

「（2）格フレームの統合」で，抽出された格フレームの統合を行う．統合処理は，二つの格フレームが次の三つの条件を満たすときに行う：(i) 動詞が同一である；(ii) 基本格が同一である；(iii) 前置詞格の格標識が同一である．この三つの条件が満たされる場合，前置詞格の格要素を格標識ごとに統合する(注 3)．例えば，[go Subj:PERSON Prep_to:tokyo] と [go Subj:PERSON Prep_to:japan] は [go Subj:PERSON Prep_to:{tokyo,japan}] と統合される．ただし，この統合処理は母語話者格フレームについてのみ行う．なぜなら，非母語話者格フレームには，正しい格フレームと誤り格フレームの両方が含まれるため両者が統合されるのは好ましくないためである．非母語話者格フレームについては，動詞，基本格，前置詞格が同一である場合のみ統合を行う．

「（3）誤り格フレーム候補の取得」では，母語話者格フ

（注 2）：誤り格フレームでは，便宜的に particle や complement も格として扱う．

（注 3）：統合の際に，各格要素の頻度を記録してもよい．本稿では特に考慮しないが，頻度情報を誤り格フレームの生成に利用しても良い．

レームと非母語話者格フレームとを比較し，誤り格フレームの候補を取得する．単純に，非母語話者格フレームにのみ存在する格フレームを誤り格フレームの候補とする．

「(4) 訂正情報の決定 」では，誤り格フレーム候補に対して訂正情報を決定する．まず，前置詞格内の格標識を別の格標識に変更する．その際，一度に変更する格標識は一つのみとする．どの格標識に変更するかは，母語からの影響を考慮した confusion set により決定する（confusion set の作成方法は 4. で述べる）．例えば，[go Subj:PERSON Prep_do:tokyo Prep_with:PERSON] の “Prep_do:” を “Prep_to:” に変更して [go Subj:PERSON Prep_to:tokyo Prep_with:PERSON] を得る．このようにして得られた格フレームが母語話者格フレームに存在すれば，その格標識を訂正情報と決定する．更に，訂正情報を誤り格フレーム候補に記述したものを誤り格フレームであると確定する．

「(5) 格要素の拡張」で，誤り格フレームのカバー率を向上させる．訂正情報により，誤り格フレームに対応する正しい格フレームが母語話者格フレームにおいて特定できる．例えば，誤り格フレーム [go Subj:PERSON *Prep_do:tokyo → Prep_to] であれば，対応する正しい格フレームは [go Subj:PERSON Prep_to:{tokyo,japan}] などになる．この例からもわかるように，統合処理により母語話者格フレームには，統合された格要素の情報も記載されている．この格要素の情報を元々の誤り格フレームの対応する格に追加することで格要素を拡張する．上述の例であれば，[go Subj:PERSON *Prep_do:{tokyo,japan} → Prep_to] のように格要素 “japan” が追加される．ただし，この拡張が真に誤りを表しているかを確認するために，新しく得られた格フレームが母語話者格フレームに存在しない場合にのみ拡張を許すこととする．

「(6) 誤り格フレームの出力」では，以上の処理で得られた誤り格フレームを出力する．2. で説明した情報を，例えば，XML 形式で出力する．

以上が，誤り格フレームの生成処理である．上述の通り，誤り格フレームは，母語話者コーパスと非母語話者コーパスさえあれば生成できる．言い換えれば，時間と労力を要する誤り情報の付与という作業を必要としない．これは，提案手法の特徴の一つである．提案手法では，誤り情報を必要としない代わりに，二つのコーパスを二度にわたり比較することで誤り格フレームの正当性をチェックする（処理 (3) と (4)）． もちろん，誤り情報が付与された非母語話者コーパスを用いて誤り格フレームを生成することも可能である．その場合には，誤り情報により誤り格フレームの選択と訂正情報の決定を行う．

4. 母語の影響を考慮した生成

母語の影響を考慮する単純な方法は，母語に応じて使用する非母語話者コーパスを変更することである．例えば，フランス語母語話者を対象とする場合には，フランス語母語話者が書いた英文を非母語話者コーパスとすることで自然に母語の影響を考慮することができる．最近では，各国語話者の書いた英文データが利用可能になりつつあるため，効率的かつ現実的な方法である．

より直接的に，母語に応じた前置詞誤りの傾向を考慮する手法も考える．前置詞の誤りはランダムに起こるのではなく，母語に応じた誤りの傾向がある [5]．例えば，フランス語の前置詞 “à” は，英語の前置詞 “at”， “in”， “to” などに対応するため，フランス語話者は，これらの前置詞を混同する傾向にあると予想できる．そこで，母語に応じて，前置詞ごとに confusion set を用意し，誤り格フレームの生成に使用する．具体的には，この confusion set に基づいて，前節「(4) 訂正情報の決定 」の際に格標識を変更する．

ここで問題となるのは，どのようにして confusion set を作成するかということである．専門家の知識により，confusion set を作成することは可能である．しかしながら，母語は多岐にわたるため，母語および前置詞に応じて confusion set を作成することはそれほど容易ではない．そこで，提案手法では，統計的機械翻訳の翻訳テーブルを利用して自動的に confusion set を作成する．直観的には，確率の値に基づいて，混同されやすい前置詞を特定する．具体的に説明するために，フランス語母語話者を想定した図 2 を考えることにする．図 2 の左の列がフランス語の単語，右の列が英語の単語である （この例では，全ての単語が前置詞であるが，必ずしも前置詞である必要はない）．図中の矢印は，フランス語の各単語が翻訳されやすい英単語を表す．すなわち，$\Pr(e|f)$ がある一定の値以上の単語の組に矢印が付与されている（ただし，e と f は英単語とフランス語の単語をそれぞれ表す）．例えば，英語の “to” は，フランス語の “à” から翻訳される確率が高いことを示す（赤い矢印）．一方で， フランス語の “à” は，“to” 以外に，“at” と “in” にも翻訳されやすいこともわかる（青い矢印）．以上をまとめると，“to” は “at” と “in” と混同されやすいことがわかる．このように，矢印を二回たどることにより confusion set を作成する．最終的に，前置詞の抜けと余剰に対応するために，“Prep_do” と “Prep_io” も confusion set に加える（i.e., {Prep_at,Prep_in,Prep_do,Prep_io}）．もし， 一回目にたどる矢印が複数ある場合（例えば，図 2 の “in”）は，それぞれ矢印をたどり，得られた前置詞の和集合を confusion set とする（したがって，“in” に対しては，{Prep_at,Prep_of,Prep_to,prep_do,prep_io} が得られる）．また，“Prep_do” と “Prep_io” に対しては，誤り検出の対象とする前置詞全てを confusion set に含める．

5. 生 成 実 験

実際に誤り格フレームを生成し，評価を行った．母語話者

コーパスは，LOCNESS(注 4)（大学生の英文のみ）と Reuters Corpus(注 5)に英語の問題集などの英文を追加したものを用いた（約 336 万トークン）．非母語話者コーパスとして，独自に収集した日本人英語学習者の英文（約 5 万トークン）と ICLE(注 6)のフランス語話者コーパス（約 18 万トークン）の 2 種類を用いた．構文解析には，Stanford Parser（ver. 2.0.3）[4] の lexical dependency parser を用いた．また，confusion set の作成には，BTEC Training Corpus(注 7)（英仏）と Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles(注 8)を用いた（条件付き確率が 5%より大きい場合に翻訳されやすいとした）．更に，対象前置詞には，文献 [5] に示される頻出前置詞 10 種類を選択した．

これらの条件のもと生成した誤り格フレームが適切に誤りを表しているかを人手により評価した．適切と判断された割合を精度とした．また，拡張された誤り格フレームについては，サンプリングを行い同様の方法で精度を推定した（元の誤り格フレームについて最大で三つサンプリングした）．

その結果，日本人英語学習者コーパスでは，抽出数が 58 で，その精度は 0.707 であった．コーパスサイズを考慮すると，この数は一見少ないように見える．しかしながら，格要素に意味を表す特殊な単語が使用されることがあるため，実際には対応する言語表現は多岐に渡る．実際に抽出された [go Subj:PERSON *Prep_do:PLACE → Prep_to] を例にすると，今回の実験で用いた辞書では，格要素 "PLACE" に対して主要な国名と都市名が 112 登録されているため，この一つで 112 の誤り格フレームに相当することになる．更に，格要素の拡張処理により，1571 の誤り格フレームが生成された（推定精度：0.810）．拡張処理により精度が向上している理由の一つとして，拡張前の誤り格フレームは適切ではないが，格要素を変更することにより適切となる場合があるということに注意する必要がある．

抽出結果を分析してみると，前置詞の抜けの誤りに関するものが全体の約 4 割を占めることがわかった（例：[belong Subj:PERSON *Prep_do:club → Prep_to]，[get Subj:PERSON prt:up *Prep_do:NUMBER → Prep_at]）．

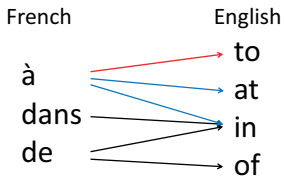


図 2： 翻訳テーブルを用いた confusion set の作成．

(注 4)：https://www.uclouvain.be/en-cecl-locness.html
(注 5)：http://about.reuters.com/researchandstandards/corpus/
(注 6)：http://www.uclouvain.be/en-cecl-icle.html
(注 7)：http://iwslt2010.fbk.eu/node/32
(注 8)：http://alaginrc.nict.go.jp/WikiCorpus/

また，[meet Subj:PERSON *Prep_with:PERSON → Prep_do]，[promise Subj:PERSON *Prep_with:PERSON → Prep_do] など，日本語からの影響と思われる誤りに関するものも抽出されていた．

一方，フランス語話者の英文においては，抽出数 31（精度 0.387）と性能が格段に悪い結果となった（拡張処理により抽出数 778（精度 0.582）まで改善）．この理由として，今回実験では，日本人英語学習者に比べフランス人英語話者のほうがライティング能力が高かったためであると推測する．具体的には，フランス人英語話者のコーパスのほうが語彙が豊富であり，母語話者コーパスに対応する正しい格フレームが存在せず，誤り格フレームが抽出できないケースがより多く起こったと分析できる．したがって，母語話者コーパスのサイズを大きくすることで抽出数は増加すると予想できる．また，構文も複雑であるように見受けられた．その結果，構文解析に失敗し，結果的に適切でない誤り格フレームを抽出しているケースも見られた．この問題に対しては，文献 [3] で提案されているように，抽出対象とする英文を予め選択するなどの対策が考えられる．

6. おわりに

本稿では，前置詞誤りの検出／訂正のための誤り格フレームを提案した．また，誤り格フレームを自動生成する手法も提案した．生成実験では，日本人英語学習者の英文を対象とした場合には精度 0.810，フランス語話者の英文を対象とした場合には精度 0.582 を達成した．誤り格フレームの特徴として，(i) 解釈が容易，(ii) 言語知識を直接反映可能，(iii) フィードバックメッセージが付与可能という点が挙げられる．

今後は，抽出した誤り格フレームを用いて，前置詞誤りを検出／訂正する手法を考案していく予定である．また，効率的にフィードバックメッセージを記述する方法も考案していく予定である．

参考文献

[1] M. Chodorow, J.R. Tetreault, and N.R. Han, "Detection of grammatical errors involving prepositions," Proc. of 4th ACL-SIGSEM Workshop on Prepositions, pp.25–30, 2007.
[2] R.D. Felice and S.G. Pulman, "A classifier-based approach to preposition and determiner error correction in L2 English," Proc. of 22nd COLING, pp.169–176, 2008.
[3] D. Kawahara and S. Kurohashi, "Acquiring reliable predicate-argument structures from raw corpora for case frame compilation," Proc. of LREC, pp.1389–1393, 2010.
[4] D. Klein and C. Manning, "Accurate unlexicalized parsing," Proc. of 41st Annual Meeting of ACL, Sapporo, Japan, pp.423–430, July 2003.
[5] A. Rozovskaya and D. Roth, "Algorithm selection and model adaptation for ESL correction tasks," Proc. of 49th Annual Meeting of ACL, pp.924–933, 2011.