誤りに関する説明を提示可能な前置詞誤り訂正手法

永田　亮†　MikkoVilenius††　EdwardWhittaker†††
† 甲南大学知能情報学部　†† 教育測定研究所　††† Inferret Limited
E-mail: †rnagata@ konan-u.ac.jp., ††vilenius@jiem.co.jp, †††ed@inferret.co.uk

1. はじめに

機械学習の導入により，英文誤り訂正の性能は大きく向上したが，一方で，従来手法には，誤りに関する説明（本稿ではフィードバックメッセージと呼ぶ）ができないという大きな問題が残されている．機械学習に基づく手法では，なぜその訂正候補が選択されたのかを人間が直感的に解釈できる形で表現することは難しい．そのため，フィードバックメッセージを生成することも難しい．このことは，語学学習支援を目的とした誤り訂正では特に問題となる．なぜなら，語学学習支援では，適切なフィードバックメッセージを学習者に提示することが重要となるからである．

このような背景を受けて，我々は，前置詞の誤りに対するフィードバックメッセージを生成するための枠組み，誤り格フレーム [3] を提案した．誤り格フレームは，図 1 のような情報を含む．詳細な説明は 2. に譲るが，図 1 の誤り格フレームは “*John often goes shopping to the market with his family. ” などの誤りに対応している．誤り格フレームの特徴として，(1) 人間が解釈できる，(2) 人手で情報を追加，修正できるという 2 点を挙げることができる．この 2 点の特徴により，フィードバックメッセージの提示が可能となる．

一方で，誤り格フレームにも課題が残されている．その一つとして，文献 [3] の誤り格フレーム生成手法は任意格が認識できず，柔軟性が低いことが挙げられる．多様な前置詞誤りに対応した柔軟な誤り訂正を実現するためには，任意格の同定が必要不可欠である．また，評価に関する課題も残る．文献 [3] では，誤り格フレームの生成性能のみを評価しており，前置詞誤り訂正の性能とフィードバックメッセージの有効性については評価していない．語学学習支援への応用を考えた場合，これらの評価も重要である．

そこで，本稿では，これら二つの課題を解決する．具体的には，任意格を自動的に同定し，誤り格フレームを生成する手法を提案する．この手法では，ヒューリスティクスを利用して任意格を同定する．また，学習者コーパスを用いた評価実験により，提案手法は，従来手法と同程度以上の訂正性能を有し，かつ，前置詞の学習に有益なフィードバックメッセージを提示できることを示す．

2. 誤り格フレーム

誤り格フレームは，動詞，格リスト，フィードバックメッセージの三つの部分からなる．図 1 では，破線で区切られた部分がこの三つに対応する．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:” や “Prep_dobj:” 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．ただし，“PERSON” のように，大文字のアルファベットのみからなる語は，複数の語を表すための特別な語とする．例えば，“PERSON” は “John” や “he” などの人を表す特別な語とする．また，格要素が複数ある場合は，各格要素をカンマで区切り記述することとする（例：*Prep_to:{store,market}）．更に，任意格は括弧 “ ( ) ” を用いて表すこととする（例：(Prep_with:{family})）．以後，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．

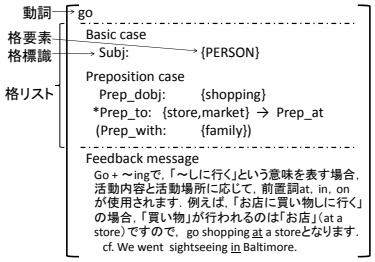

図 1：　誤り格フレームの例．

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 では “go” が動詞である．

格リストは，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．格は基本格と前置詞格の 2 種類に分類される．基本格は一つ以上の格からなると定義する．基本格に入る格標識は，“Subj:”（Subject），“Prt”（Particle）とする(注 1)．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なると定義する．動詞が取りうる前置詞を格標識 “Prep_x” を用いて記述する（x の部分には前置詞が入る：例 Prep_to）．ただし，直接目的語（Prep_dobj）と間接目的語（Prep_iobj）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の誤り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．誤りがある前置

(注 1)：誤り格フレームでは，便宜的に particle も格として扱う．

詞格に “*” 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．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．

フィードバックメッセージは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．

3. 誤り格フレームの生成手法

本稿で提案する生成手法は文献 [3] の手法を基本とするが，新たな処理として任意格の同定がある．また，その影響により，他の処理についても変更が必要となる．

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） 入力の選択
（2） 格フレームの生成
（3） 任意格の同定
（4） 格フレームの統合
（5） 誤り格フレーム候補の取得
（6） 訂正情報の決定
（7） 誤り格フレームの拡張
（8） 誤り格フレームの出力

「(1) 入力の選択」では，格フレーム生成に不適切な文を除外する．後述するように，格フレームの生成には構文解析を伴うが，長い文は構文解析に失敗することが多いため除外する．関連研究 [1] を参考にして，提案手法では，20 トークン以上の長さの文を除外する．また，カンマを含む文は構文が複雑であることが多いため除外する．なお，入力の選択を行うのは母語話者コーパスのみとする．非母語話者コーパスは母語話者コーパスに比べて利用できる量が少ないため全ての文を生成に利用する．

「(2) 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．前処理として入力文を構文解析する．図 2 に示すように，構文解析の結果から得られる動詞と格を格フレームの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）を原形にしたものを用いる．また，非母語話者コーパスの場合，スペルチェッカを用いて綴り誤りを自動的に訂正する．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．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，別々に行う．以後，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．

「(3) 任意格の同定」では，次のヒューリスティクスを用いて，格フレーム中で任意となる前置詞格を同定する：(a) 目的語は常に必須格とする；(b) 動詞より左に出現する前置詞格は常に任意格とする；(c) 動詞の右に出現する前置詞格は，動詞に一番近いものを除いて全て任意格とする．(a) は，目的語は常に必須格となることを規定する（目的語も便宜的に前置詞格として扱うことに注意）．(b) は，動詞と前置詞格の位置関係を比べ，前置詞格のほうが左に出現する場合は，常に任意格とすることを意味する．例えば，“In the morning, he went shopping.” では，下線部の前置詞格は動詞より左に出現しているため任意格とする．(c) により，動詞の右に現れる前置詞格は，動詞に一番近い前置詞格を必須格とし，残りを任意格と同定する．例えば，“He went to the market with his family” の場合，動詞に近い “to the market” を必須格，遠いほうの “with his family” を任意格とする．以上のヒューリスティクス以外にも，二つの格フレームを比較することにより，任意格が自然に同定できる場合もある．例として，次の二つの文を考えることにする：“He went shopping.”；“He went shopping at the market.” この場合，前者は，前置詞格 “at the market” がなくても後者が文として成り立つことを示している．したがって，このような場合は，二つの格フレームの比較により任意格と同定する．

「(4) 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) 必須格である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，前置詞格を統合する（具体例を図 3 に示す）．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

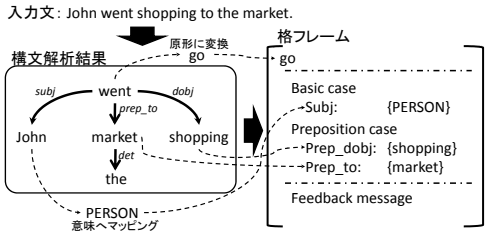

図 2： 格フレームの生成例．

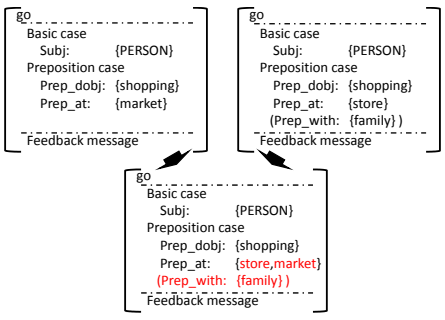

図 3： 格フレームの統合処理．

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．

「(5) 誤り格フレーム候補の取得」では，母語話者格フ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．母語話者格フレームのいずれにもマッチしない非母語話者格フレームを誤り格フレームの候補とする．ここで，マッチの条件を，(条件 1) 動詞と基本格が同一である，(条件 2) 両方の格フレームにおいて必須格である前置詞格が同じ格要素を少なくとも一つ含む，と定義する．

「(6) 訂正情報の決定 」では，誤り格フレーム候補の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．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マッチすれば，その前置詞を訂正情報と決定する（マッチの条件は前述の通り）．訂正情報が決定できた場合，誤り格フレーム候補に記述し，誤り格フレームであると確定する．

「(7) 誤り格フレームの拡張」で，格リストの拡張を行う．処理 (6) 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されている．例えば，図 4 の場合，誤り格フレームが左上の図，対応する母語話者格フレームが右上の図となる．この誤り格フレームより，誤りは “*Prep_to:{market}” であり，正しい前置詞は “at” であることがわかる．ここで，母語話者格フレームには，処理 (4) により，複数の格要素が記述されていることを利用する．この例の場合，“Prep_at:” には “market” に加えて “store” も格要素として記載されている．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．ただし，この拡張が真に誤りを表すかどう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のいずれにもマッチしない場合にのみ，対応する母語話者格フレームの格の情報を誤り格フレームに追加する．ここでのマッチは，処理 (5) で定義したマッチの 2 条件に加え，(条件 3) 誤りがある前置詞格も共通な格要素を含む，で定義する（以降でのマッチとはこの定義に従う）．拡張された誤り格フレームは図 4 の下図のようになる．

「(8) 誤り格フレームの出力」で，以上の処理で生成された誤り格フレームを出力する．図 4 に示されるように，出力される誤り格フレームは人間が解釈可能な形式をしている．そのため，誤り格フレームを解釈し，フィードバックメッセージを記述することができる．また，必要に応じて，格や格要素を追加，削除することもできる．これらの編集は人手で行うことになるが，誤り訂正のための規則とフィードバックメッセージを一から人手で作成することに比べると格段に効率が良い．なぜなら，「(7) 誤り格フレームの拡張」により，一つの誤り格フレームに非常に多くの誤りが集約されており，効率良く情報の追加，修正が行えるからである．

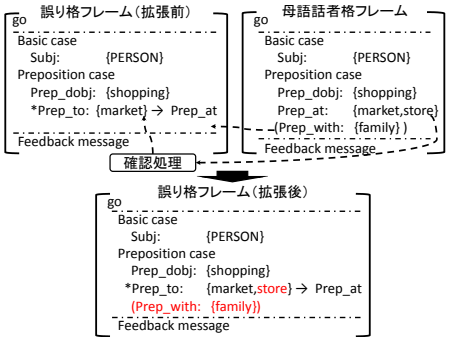

図 4： 誤り格フレームの拡張．

4. 誤り格フレームを用いた誤りの訂正

誤りの訂正は，生成された誤り格フレームを訂正対象の英文に適用することで行う．まず，3. の処理 (2) と (3) を用いて，訂正対象の英文から格フレームを抽出する．抽出された格フレームが誤り格フレームのいずれかにマッチすれば，その誤り格フレームの訂正情報に従い誤りを訂正する．

予備実験により，この手法の訂正性能を評価したところ，誤り格フレームでは原理的に訂正できない前置詞誤りがあることが明らかになった．より正確には，前置詞誤りと対応する正しい表現が非母語話者コーパスと母語話者コーパスにそれぞれ出現しても，誤り格フレームが生成されない前置詞誤りが存在することが明らかになった．具体的には，場所を表す副詞（“there” など）に，場所に関する前置詞（“at, in, on, to”）を付けた誤り（例：“*John went to there.”）と時間，頻度，期間を表す名詞に，時間，頻度，期間を表す前置詞（“at, for, in, on”）を付けた誤り（例：“*John goes shopping in every morning.”）である．

これらの誤りは動詞と基本格には独立であり，種類数が少ないことを考慮して人手で誤り格フレームを作成することとした．全ての単語を意味する特別な語 ANY を定義して，例えば，“[ANY Subj:{ANY} *Prep_to:{here,somewhere,there} → Prep_dobj]” を作成した（作成にあたっては，文献 [5] を参考とした）．また，これらの誤りはフレーズに関連するものもあるため，人手で作成する誤り格フレームのみ格要素にフレーズを許した（例：“[ANY Subj:{ANY} *Prep_in:{every morning} → Prep_dobj]”）．

5. 評 価 実 験

提案手法を訂正性能とフィードバックメッセージの有効性の二つの観点から評価した．訂正性能は，訂正率，訂正精度，F 値で評価した．フィードバックメッセージの有効性は，英語教育の専門家 3 人による人手の評価とした．各専門家は，訂正結果とフィードバックメッセージを読み，フィードバックメッセージが前置詞の学習に役に立かどうかを判定した．

訂正対象の英文として，Konan-JIEM learner corpus（KJ コーパス）[4] を用いた．同コーパスのトレーニングセット

とテストセットを用いて誤り格フレームを生成し，テストセットで訂正性能を評価した(注 2)．また，母語話者コーパスとしてはEDRコーパス(注 3)，Reuters-21578 corpus(注 4)，LOCNESS corpus(注 5)を併せたものを使用した．構文解析器として，Stanford Statistical Natural Language Parser (ver.2.0.3) [2] の lexicalized dependency parser を利用した．

これらのデータを利用して，任意格あり／なしの提案手法を実装した．比較対象としては，機械学習に基づく手法[6]と統計的機械翻訳に基づく手法[7]を選択した．前者は，KJコーパスに同梱されているベンチマークシステムの中で最も性能が高いシステムである．後者は，CoNLL Shared Taskにおいて，前置詞誤り訂正の性能が最も高いシステムである．

表 1 に訂正性能の評価結果を示す．表より，任意格の同定処理により，訂正率も訂正精度も向上し，任意格同定の処理に効果があることがわかる．任意格ありの誤り格フレームでは，現時点で最も性能が高い従来手法と同程度以上の F 値を達成することもわかる．ただし，提案手法は訂正精度重視，従来手法は訂正率重視という傾向がある．

この訂正の傾向の違いは次のように分析することができる．提案手法は，動詞と格という文の骨格となる要素に基づくため，前置詞周辺の他の誤りから受ける影響は小さい．逆に，前置詞周辺の情報を利用する従来手法では影響が大きい．その結果，提案手法では訂正精度が高くなる傾向にある．同様の理由で，提案手法と従来手法では訂正できる前置詞誤りの種類にも差異がある．構文情報を利用する提案手法では，動詞から離れた位置にある前置詞誤りも動詞とその他の前置詞格の情報を利用して誤りの訂正を行える．例えば，評価実験では，“*In the univerysity, I studied English in the moring.” に対して，構文解析により “In the universyity” が動詞 “studied” の前置詞格であることを認識し，適切に誤りを訂正した（“In” の周辺の情報だけを利用する従来手法は訂正に失敗した）．一方，誤り格フレームは，動詞にかかる前置詞のみを扱うため，名詞にかかる前置詞の誤りは原理的に訂正できない（評価実験では，この種の誤りは全体の 10%であった）．

フィードバックメッセージに関する評価実験では，20 の誤りに対してフィードバックメッセージが提示できた．そのうち，82%（評価者 3 人の平均）が前置詞の学習に役に立つと判定された．このようにフィードバックメッセージが提示可能となると，誤り訂正の方法として，次の 3 種類を考えることができる：(i) 正しい前置詞を提示する（従来の訂正方法である）；(ii) 正しい前置詞と共にフィードバックメッセージを提示する；(iii) フィードバックメッセージのみを提示する．(i) の場合，学習者は，単に正しい前置詞をコピーするだけで英文の修正が行えるため，学習効果が低くなる可能性がある．(ii) でも，フィードバックメッセージを読まずとも，英文の修正が行えるため同様の可能性がある．一方，(iii) では，学習者は実際にフィードバックメッセージを読み，理解しないと正しい前置詞を選択することができない．そのため，(iii) が一番学習効果が高いと予想できる．したがって，我々は，(iii) の誤り訂正方法を提案する．我々が知る限り，機械学習に基づく手法と同程度以上の訂正性能を達成し，(iii) の誤り訂正が実現可能であるのは誤り格フレームのみである．

表 1： 訂正性能の評価結果.

手法	訂正率	訂正精度	F 値
誤り格フレーム（任意格なし）	0.092	0.523	0.156
誤り格フレーム（任意格あり）	0.130	**0.680**	**0.218**
機械学習に基づく手法	**0.167**	0.310	0.217
統計的機械翻訳に基づく手法	0.115	0.385	0.176

6. おわりに

本稿では，誤り格フレームに基づいた前置詞誤り訂正手法を提案した．評価実験により，任意格を自動的に同定する提案手法は，従来手法と同程度以上の訂正性能を達成することが明らかになった．また， 提案手法が提示したフィードバックメッセージの 8 割以上が前置詞の学習に役に立つと評価された．従来手法と同程度以上の訂正性能を有し，フィードバックメッセージが提示可能である提案手法は，前置詞誤り訂正の有力なアプローチの一つになると期待される．

参考文献

[1] D. Kawahara et. al., “Acquiring reliable predicate-argument structures from raw corpora for case frame compilation,” Proc. of LREC, pp.1389–1393, 2010.

[2] M.C. de Marneffe et. al., “Generating typed dependency parses from phrase structure parses,” Proc. of LREC, pp.449–445, 2006.

[3] 永田他, “前置詞誤り検出／訂正のための誤り格フレーム,” 言語処理学会第 19 年次大会発表論文集, pp.616–619, 2013.

[4] R. Nagata et. al., “Creating a manually error-tagged and shallow-parsed learner corpus,” Proc. of 49th Annual Meeting of ACL, pp.1210–1219, 2011.

[5] R. Quirk et. al., “A Comprehensive Grammar of the English Language,” Longman, New York, 1985.

[6] K. Sakaguchi et. al., “NAIST at the HOO 2012 shared task,” Proc. of 7th Workshop on the Innovative Use of NLP for BEA, pp.281–288, 2012.

[7] I. Yoshimoto et. al., “NAIST at 2013 CoNLL grammatical error correction shared task,” Proc. of 17th Conference on CoNLL: Shared Task, pp.26–33, 2013.

(注 2)：訂正対象の英文を誤り格フレームの生成に用いても，評価はオープンテストとなる．なぜなら，誤り格フレームの生成には誤りの情報を必要としないため，誤り訂正時に訂正対象の英文から誤り格フレームが生成できるからである．ただし，そのようにして生成した誤り格フレームでは，人手で追加するフィードバックメッセージは利用できない．

(注 3)：http://www2.nict.go.jp/out-promotion/techtransfer/EDR/

(注 4)：http://www.research.att.com/ lewis

(注 5)：http://www.uclouvain.be/en-cecl.html