The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

主語と補語の意味的関係を考慮したbe動詞の過剰一般化誤り検出

A Method for Detecting Overgeneralized *Be*-Verb based on Subject-compliment Identification

永田 亮*1*2 西手 翔矢*1 乙武 北斗*3
Ryo Nagata Shoya Nishite Hokuto Ototake

*1甲南大学 Konan University
*2国立研究開発法人科学技術振興機構，さきがけ JST, PRESTO
*3福岡大学 Fukuoka University

This paper addresses the detection of overgeneralization of *be*-verb found in learner English. It is an error where the subject and complement are not semantically equivalent in a *be*-verb sentence as in **Paris is rain.* This paper presents a method for detecting it by predicting through word embeddings whether a given subject and complement pair is semantically equivalent or not. This paper also presents a method for determining the hyperparameters in the method efficiently and effectively. Experiments show that the present method outperforms two baseline methods based on corpus statistics and WordNet ontology. Looking into the detection results brings out a way of generating feedback messages for learners that explain why the detected error is not a valid English expression.

1. はじめに

英語を対象にした文法誤り検出の性能は著しく向上しているが, 依然として研究例の少ない誤りも存在する. その代表例に，be 動詞の過剰一般化誤りがある. be 動詞の過剰一般化誤りとは, 例えば，“*Paris is rain.（パリは雨だ。）” のように，意味的な関係が同一でない主語と補語を be 動詞でつなげる誤りを指す*1.

be 動詞の過剰一般化誤りは日本人英語学習者によくみられる誤り [小島 88] であり，検出対象として重要である．この種の誤りが頻出する主な要因に母語からの影響がある [小島 88]. したがって，語学学習支援という観点からは，母語における be 動詞相当表現と英語の be 動詞の用法の違いを学習者に説明できることが望ましい．そのためには，まずは，be 動詞の過剰一般化誤りを他の誤りと区別して検出できる必要がある．

従来の文法誤り検出／訂正手法では，be 動詞の過剰一般化誤りを検出することは難しい．冠詞や前置詞など代表的な文法誤りの検出では，対象語（冠詞や前置詞）の周辺情報に基づいて正しい語を選択するという分類問題を解くことが多い．しかしながら，この方法は be 動詞の過剰一般化誤りに適用することは困難である．そもそも be 動詞の過剰一般化誤りの場合，周辺情報から正しい be 動詞を選択するという問題ではない．上の例でみたように，be 動詞の過剰一般化誤りの検出では，主語と補語の組み合わせが意味的に同一かどうかを判別しなければならない．be 動詞の過剰一般化誤りの情報が付与された学習者コーパスは皆無であるため，正誤を判定するという分類問題 [Izumi 03] としても解くことも困難である．翻訳問題として文法誤り訂正を解く手法（例えば文献 [Chollampatt 16]）であれば，一部の誤りについては訂正できる可能性はある．しかしながら，誤りの種類を区別せず全ての誤りを同時に訂正するため，誤りの理由は説明できない．また，be 動詞の過剰一般化誤りには，翻訳問題として解くことが難しい，文全体を書き換える必要のある誤りが少なくない（例：“*Paris is rain.” → “It is raining in Paris.”）.

そこで，本稿では be 動詞の過剰一般化誤りを検出する手法を提案する. 上で述べた通り be 動詞の過剰一般化誤りの検出には, 主語と補語の意味的な同一関係の判定が重要となる．提案手法では，単語の分散表現を利用して，この問題を解決する．単語の分散表現は，単語の通時的な意味変化を捉える研究 [Hamilton 16] や英語と日本語の意味のずれを検出する研究 [Takamura 17] などに応用されており，意味的な同一関係の判定にも適していると予想される．提案手法では，単語の分散表現を利用して意味的に同一でないと判定された主語と補語の組を be 動詞の過剰一般化誤りとして検出する.

本研究の貢献は次の通りである. 我々が知る限り，提案手法は，世界初の be 動詞の過剰一般化誤り検出手法である．その性能は，二種類のベースライン（主語と補語の共起に基づいた手法と WordNet [Miller 95] のオントロジーに基づいた手法）よりも有意に高い．また，検出結果を利用して，誤りの理由を自動的に説明する方法も示す.

2. be 動詞文と be 動詞の過剰一般化誤り

本稿では **be 動詞文**を次のように定義する：

be 動詞文：第 2 文型の文のうち， 動詞が be 動詞である，かつ主語，補語どちらとも名詞である文.

この定義の下の be 動詞文には，次の二つの用法がある [Quirk 85].

同一性の表示： 主語と補語が意味的に同一であることの表示.
例) Kevin is my brother.

特徴づけ： 主語の名詞を補語により特徴づける.
例) Dwight is an honest man.

以降, 本稿では，この二つの用法をまとめて同一関係と表記することにする.

上述の通り，be 動詞文中の主語と補語は同一関係になければならないが, 学習者の英文では，そうでない場合がある*2. 本稿では，これを be 動詞の過剰一般化誤りと呼ぶ. 以降では，特に断らない限り，be 動詞の過剰一般化誤りを，単に，過剰一般化誤りと表記する.

連絡先: 永田亮，甲南大学，nagata-ai @ hyogo-u.ac.jp.

*1 厳密な定義は 2. 節で述べる.

*2 実際には，英語でも “I'm coffee.” という表現が可能である．しかしながら，限られた場面でのみ用いられる特殊な用法である．学習者が取り組むライティングで用いられることは稀である.

過剰一般化誤りが起こる要因の一つに母語干渉がある．例えば，日本語には，be 動詞文に（見かけ上の）用法が類似する「〜は〜だ。」という表現がある．この表現では，主語と補語が意味的に同一関係にない場合でも使える．例えば,「パリは雨だ。」,「今日は学校だ。」,「飛行機は危険だ。」のように表現できる．日本語のこの用法が英語に適用され過剰一般化誤りが起こると分析できる．このように，母語の be 動詞相当表現に，主語と補語が同一関係でないものを結びつける用法があると，過剰一般化誤りが起こりやすい．

3. 提案手法

3.1 検出手順

提案手法の誤り検出手順は次の通りである：(1) 検出対象文書の入力；(2) 主語と補語の抽出；(3) 主語と補語の同一関係の判定；(4) 誤り検出；(5) 後処理；(6) 検出結果の出力．

手順 (1) で，検出対象となる文書から各文を読み込む．手順 (2) で，入力文を依存構造解析し, 主語および補語となる主名詞を抽出する．抽出対象とするのは，be 動詞文のみである．手順 (3) では，主語と補語の同一関係の判定を行う．この処理の詳細については，3.2 項で述べる．手順 (4) で，判定結果に基づいて誤り検出を行う．同一関係にないと判定された主語と補語の組み合わせを誤りとする．手順（5）で，後処理として，検出結果に対する例外処理を行う．具体的には，主語または補語が次の単語の場合は，手順（4）の結果にかかわらず常に正しいと判定する：it，they，this，these，that，those，thing，things．これらの語は，指し示す意味範囲が非常に広いため誤り検出対象外とする．最後に，手順（6）で，検出結果を出力する．

3.2 主語と補語の意味的同一関係の判定

主語と補語の意味的な同一関係の判定は単語の分散表現を利用して行う．そのため，単語の分散表現を大規模な母語話者コーパスから，あらかじめ学習しておく．

主語と補語の意味的な同一関係の判定を定式化するため，次の記号を導入する．いま，主語と補語を，それぞれ s と c で表す．また，それぞれに対する分散表現（ベクトル）を $\boldsymbol{v}_s$ と $\boldsymbol{v}_c$ で表す．

このとき，主語 s と補語 c の同一関係の判定を $\boldsymbol{v}_s$ と $\boldsymbol{v}_c$ の余弦類似度に基づいて行う．具体的には，θ を閾値として，

$$\cos(\boldsymbol{v}_s, \boldsymbol{v}_c) > \theta \tag{1}$$

を満たす主語 s と補語 c との組み合わせを同一関係にあると判定する（閾値 θ の決定方法は，3.3 項で説明する）．すなわち，主語と補語が同一関係にあるかどうかを，余弦類似度で測定された主語と補語の類似度に基づいて近似的に判定していることになる．これは，一見荒い近似にみえるが，5. 節で示すように，その判定精度は高い．なお，低頻度な単語に対しては，分散表現がうまく学習できない可能性が高いため，訓練データ中での頻度が一定以上の主語と補語に対してのみ，式 (1) を用いて同一関係の判定を行う．それ以外の場合は，常に正しいと判定する．

3.3 ハイパパラメータの推定方法

提案手法で，高い検出性能を達成するためには，ハイパパラメータを適切に設定することが重要となる．具体的には，式 (1) の閾値 θ，分散表現の次元数や文脈窓幅などのハイパパラメータである．通常は，開発データを用いて決定することが多い．しかしながら，過剰一般化誤りの情報が付与された学習者コーパスが乏しい現状では，この方法の利用は現実的でない．

解決策として，母語話者コーパスから，ハイパパラメータ決定用の疑似的な開発データを生成する．その手順の概要を図 1 に示す．まず，母語話者コーパスから，3.1 項の手順 (1) と (2) を用いて，主語と補語の組を抽出する（図 1（1））．次に，得られたペアのうち，主語または補語が学習者コーパスに一度も出現しないものを除外する（図 1（2））[*3]．これは，学習者の英文に極力近い単語からなる開発データを生成する目的で行う．この結果を，同一関係にある正しい組み合わせとみなす（以降，この組を集めたものを正例データと呼ぶ）．次に，正例データから，主語と補語をそれぞれ独立にランダムに抽出して新たな組を作る（図 1（3））．ただし，新たに作られた組が正例データに存在する場合は除外する．得られた組を，同一関係にない主語と補語（すなわち，過剰一般化誤り）とみなす（以降，この組を集めたものを負例データと呼ぶ）．最後に，正例データと負例データを統合して疑似的な開発データとする（図 1（4））．この疑似的な開発データを対象にして，様々なハイパパラメータで誤り検出を行い，その検出性能（例えば，F 値）を見積もる．性能値が最大となる組み合わせを選択する．

以上の手順を実施するためには，正例データと負例データを混合する割合が問題となる．本稿では，事前分布に基づいて決定する．言い換えれば，be 動詞文のうち何割が過剰一般化誤りになるかの事前知識に基づいて，負例データを生成する割合を決定する．厳密には，事前分布の推定のためには訓練データが必要であるが，訓練データの希少性を考慮し，本稿では事前分布は与えられているとする．

更に，頑健なハイパパラメータ推定とするため，提案手法では，複数の分散表現を利用する．まず，ハイパパラメータ（例えば，文脈窓幅や次元）が異なる複数の分散表現を得る．次に，それぞれの分散表現について，上述の方法で，式 (1) の閾値 θ を決定する．最後に，各分散表現を用いて誤り検出を行い，その多数決を最終的な検出結果とする．多数決をとることにより，より頑健な誤り検出となることが期待される．

4. 評価実験

検出対象として，Konan-JIEM Learner Corpus 第 5 版 [Nagata 16]（KJ）を使用した．KJ には，過剰一般化誤りの情報は付与されていないため，人手により新たに誤りの情報を付与した．まず，Stanford Parser Ver.3.5.0[*4]の LexicalizedParser を用いて KJ を解析し，be 動詞文および主語と補語の組を抽出した．その結果，294 の主語と補語の組が得られた．この 294 組に対して，原文を見ながら，第一著者と第二著者が独立に正誤のタグを付与した．その後，不一致する箇所について両者で相談し，最終的な正誤を決定した．その結果，83 の過剰一般化誤りを同定した．

分散表現の訓練については，文分割とトークン同定済みの One Billion Word Model Benchmark[*5]中の news.en-00001-of-00100 から news.en-00099-of-00100 を用いた[*6]．次元数を

*3 学習コーパス中で出現するかどうかの判定は，主語，補語以外の使用も考慮して行う．

*4 `https://nlp.stanford.edu/software/lex-parser.shtml`

*5 `http://www.statmt.org/lm-benchmark/`

*6 訓練には，word2vec（`https://github.com/tmikolov/word2vec`）を用いた．使用したハイパパラメータの値は次のとおりである：-negative 25 -sample 1e-4 -iter 15 -cbow 1 -min-count 5．なお，全てのトークンを小文字に変換した．

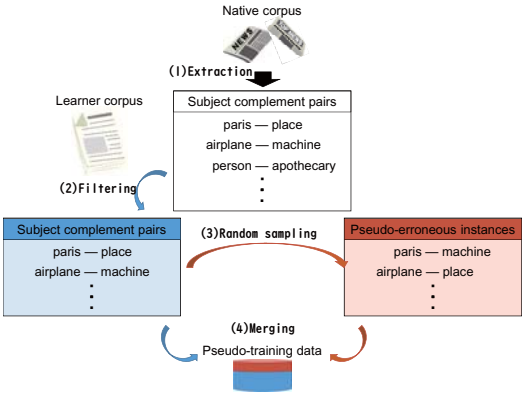

図 1: 擬似的な開発データの生成の流れ.

200，400，800，文脈窓幅を 5，10，15，20，25 とした計 15 種類の分散表現を用いた．ただし，訓練コーパス中での出現頻度が 5 回以上の単語のみを対象とした（それ以外のパラメータは脚注のように固定した）．

閾値の決定にも同じコーパスを用いた．負例データの割合は，KJ での過剰一般化誤りの割合に合わせて 28%とした．このデータを対象にして，閾値を 0〜1 の範囲で 0.01 刻みで変化させたときの F 値を算出し，F 値が最大となる閾値を選択した．なお，本評価は，過剰一般化誤りの割合に関する事前知識を用いているため，厳密にはブラインドテストではないことに注意が必要である．

比較のため，2 種類のベースライン手法を実装した．一つ目は，共起に基づいた手法である．すなわち，母語話者コーパス内で一度も出現しない主語と補語の組み合わせを誤りとする手法である．二つ目は，Lin の類似度 [Lin 98] に基づいた手法とした．Lin の類似度は，WordNet のオントロジーに基づいて計算される，二単語間の意味的な類似性を表する指標である．主語と補語の組に対する Lin の類似度を求め，閾値以下となる場合を誤りと判定した．閾値については，このベースラインの性能限界を明らかにするため，検出対象である KJ 上で F 値が最大になる値を設定した（したがって，このベースラインについても完全なブラインドテストではない）．両ベースラインとも，提案手法の場合と同一のコーパスと解析器を利用した．また，提案手法と条件を合わせるため，主語，補語共に母語話者コーパスでの頻度が 5 以上のみの場合を検出対象とした（これ以外の場合は，常に正しいと判定した）．

評価尺度は，判定精度（正誤の判定に成功した主語，補語の組の数を主語，補語の組の総数で除した値）を用いた．また，再現率，適合率，F 値も用いた．

表 1 に，結果を示す．表 1 には，参考として，KJ に対する F 値が最大となるハイパパラメータを用いた提案手法の性能「提案手法（最適ハイパパラメータ）」も記載している．なお，全て正しいと判定した場合の判定精度は 0.717（=211/294）である．表 1 から，提案手法の方が，ベースライン手法よりも判定精度，F 値共に高いことがわかる；提案手法（推定ハイパパラメータ，多数決あり）と最も性能がよいベースライン（「ベースライン（共起）」）における，判定精度の差は有意である（MacNemar 検定，$p = 0.015$）．コーパス内での単純な共起に基づくと，スパースネスの問題により誤判定が多くなることがわかる．一方，提案手法では，より高い精度で，意味的な同一関係の判定に成功している．提案手法は，高い適合率が求められる語学学習支援への応用に特に適しているといえる．

提案手法（推定ハイパパラメータ，多数決あり）と提案手法（推定ハイパパラメータ，多数決なし）を比較すると，多数決を取ることの効果がうかがえる．複数の分散表現に基づいた検出結果の多数決を取ることで，最適なハイパパラメータを設定したときに近い性能を達成している．

このことをより詳細にみるために，図 2 に，ハイパパラメータと判定精度の関係を示す．図の横軸は文脈窓幅に対応する．また，縦軸は，判定精度に対応する．各線は，分散表現の次元数に対応する．なお，比較のため，共起に基づいたベースライン手法の判定精度も示している．図 2 より，文脈窓幅を 10 単語以上にとると，大抵の場合，最適なハイパパラメータを設定したときの判定精度に近くなるが，中にはベースラインの性能を下回る場合もあることがわかる．このことは，ハイパパラメータを最適に設定することの重要性を示唆する．多数決を取ることで，ハイパパラメータを唯一の組み合わせに設定する必要がなくなり，よりロバストな検出手法となることもわかる．以上より，開発データが利用可能でない場合は，文脈窓幅も分散表現の次元数もある程度の大きさの範囲に設定し，複数の分散表現による検出結果の多数決をとればよいことがわかる．

5. 考察

提案手法の検出の傾向を明らかにするため，提案手法（最適ハイパパラメータ）による検出結果を分析した．同手法で，余弦類似度が負の値になった主語と補語の組は 20 件であった（全て誤りと判定した）．そのうち，実際に誤りであったのは 12 件（判定精度 60%）であった．この判定精度は一見低いようにみえる．しかしながら，誤検出 8 件中 6 件は，主語または補語が固有名詞であり，それ以外の 12 件の一般名詞における判定精度は 83%と非常に高いことが明らかとなった．また，余弦類似度が 0.3 より大きいの主語と補語の組は，48 件存在した（全て正しいと判定した）．そのうち，実際に正しい組み合わせは，41 件であり，こちらの方も高い判定精度であることが確認できた（判定精度 85%）．これらの結果は，少なくとも一般名詞については，余弦類似度の大小により，主語と補語の意味的な同一関係がある程度判定できることを示す．

全体的に，固有名詞に対して検出に失敗する傾向が見られた．日本人名や日本の地名などの固有名詞は，母語話者コーパスでは出現回数が少なく，信頼性の高い分散表現が得られなかった可能性がある．更に，一部の名詞については，偶然に英語の固有名詞と同一表記になることがあり，判定失敗の大きな要因となっていた．例えば，“Kobe”（神戸）や “Himeji Oden”（姫路おでん）は，今回用いた母語話者コーパスでは，“Kobe (Bryant)” と “(Greg) Oden”（両者ともバスケットボール選手）としての使用が散見された．そのため，“Kobe is a nice place.” などの表現での誤検出につながる（“Kobe”，“Oden” は全部で 7 回出現し，その内 4 回について検出に失敗していた）．このことから，固有名詞に対しては検出を行わない，別の方法で検出を行うなどの処置が必要なことがわかる．

検出結果をより深く理解するために，検出対象に含まれる 83 の誤りを訂正方法により細分類した．その結果 7 種類を同定した．表 2 に分類結果と対応する再現率を示す．

表 2 より，最も多いのは「文全体を書き換える必要がある誤り」であることがわかる（例：“*Japan was a winter.”）．以降,「主語または補語を別の名詞にすべき誤り（例： *My job is an acceptance.）」,「動詞を変更すべき誤り（例： *I can be fun.）」,「補語を現在／過去分詞にすべき誤り（例： *I was worry.）」,「補語を形容詞にすべき誤り（例： *Airplanes is danger.）」と続く．どの誤りについても，ある程度，検出に成

表 1: 評価結果.

手法	判定精度	再現率	適合率	F 値
提案手法（最適ハイパパラメータ）	**0.786**	0.614	**0.622**	**0.618**
提案手法（推定ハイパパラメータ，多数決あり）	0.762	0.578	0.578	0.578
提案手法（推定ハイパパラメータ，多数決なし）	0.643	0.639	0.414	0.502
ベースライン（共起）	0.673	0.530	0.436	0.478
ベースライン（Lin の類似度，最適ハイパパラメータ）	0.599	**0.735**	0.389	0.509

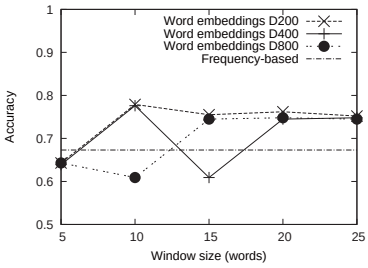

図 2: ハイパパラメータと判定精度の関係.

表 2: 誤りの細分類と再現率.

誤りの種類	再現率	
文全体を書き換える必要がある誤り	0.47	(9/19)
主語または補語を別の名詞にすべき誤り	0.50	(9/18)
動詞を変更すべき誤り	0.50	(5/10)
補語を現在／過去分詞にすべき誤り	0.88	(7/8)
補語を形容詞にすべき誤り	1.0	(7/7)
補語の前に前置詞を挿入すべき誤り	0.50	(3/6)
主語と補語の帰属関係が逆である誤り	0.00	(0/1)
その他	0.57	(8/14)
計	0.58	(48/83)

功しており，分散表現に基づいた同一関係の判定が有効であることがわかる．例外としては，主語と補語の帰属関係が逆である誤りがある（例： *Fruits are bananas.）．ここまで，棚上げにしてきたが，提案手法では主語と補語の帰属関係の判定は行わない．したがって，この種の誤りは検出できない．しかしながら，表 2 からわかるように，この種の誤りの頻度は非常に低い．過剰一般化誤りが頻出するのは，学習者が，主語と補語の帰属関係を判断できないためではなく，母語の be 動詞相当表現に備わった他の用法を英語に適用するためであると考えると，この現象をうまく説明できる．これが正しいとすると，主語と補語の帰属関係を判定しなくても，実用上の問題は少ないと予想される．

主語と補語の帰属関係が逆である誤り以外については，母語話者コーパスから得られる統計量に基づいて誤りの種類を識別できる可能性がある．例えば,「補語を形容詞にすべき誤り」では，補語の名詞に対応する形容詞があるかを辞書で確認し，存在すれば，主語，be 動詞，形容詞の三つ組の頻度をコーパスから得る．そのほかの誤りでも同様に（例えば,「動詞を変更すべき誤り」であれば，主語，動詞，目的語（= 補語）の三つ組）の頻度を得る．それらの頻度を適切に正規化して比較することで誤りの種類を特定できる可能性がある．その結果，例えば,「be 動詞文では，主語と補語は，意味的に同一なものを用いなければなりません．補語 “danger” を形容詞に代えられませんか？」のような説明が可能となる．仮に，有効な統計量が得られなければ,「文全体を書き換える必要がある誤り」であるとし,「文全体を書き換えて別の表現にしてみましょう.」などの説明を与えることが考えられる．

6. まとめ

本稿では，主語と補語の意味的な同一関係の判定に基づいて，be 動詞の過剰一般化誤りを検出する手法を提案した．評価実験により，提案手法の性能は，共起に基づいた手法および WordNet のオントロジーに基づいた手法よりも有意に高いことを示した．また，検出結果を利用して，be 動詞の過剰一般化誤りに対する解説を自動生成する方法も示した．

謝辞

本研究は，JST，さきがけ，JPMJPR1758 の支援をうけたものである．

参考文献

[Chollampatt 16] Chollampatt, S., Taghipour, K., and Ng, H. T.: Neural Network Translation Models for Grammatical Error Correction, in *Proc. of 25th IJCAI* (2016)

[Hamilton 16] Hamilton, W. L., Leskovec, J., and Jurafsky, D.: Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change, in *Proc. of 54th ACL* (2016)

[Izumi 03] Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., and Isahara, H.: Automatic error detection in the Japanese learners' English spoken data, in *Proc. of 41st ACL* (2003)

[Lin 98] Lin, D.: An Information-Theoretic Definition of Similarity, in *Proc. of 15th International Conference on Machine Learning* (1998)

[Miller 95] Miller, G. A.: WordNet: A Lexical Database for English, *Communications of the ACM*, Vol. 38, No. 11 (1995)

[Nagata 16] Nagata, R. and Sakaguchi, K.: Phrase Structure Annotation and Parsing for Learner English, in *Proc. of 54th ACL* (2016)

[Quirk 85] Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J.: *A Comprehensive Grammar of the English Language*, Longman, New York (1985)

[Takamura 17] Takamura, H., Nagata, R., and Kawasaki, Y.: Analyzing Semantic Change in Japanese Loanwords, in *Proc. of 15th EACL* (2017)

[小島 88] 小島 義郎, 日本語の意味 英語の意味, 南雲堂 (1988)