構文構造に基づく英語表現の自動獲得とその評価

葛原 和也 † 加藤 芳秀 ‡ 松原 茂樹 †

† 名古屋大学大学院情報科学研究科 ‡ 名古屋大学情報基盤センター

1 はじめに

英語を母語としない研究者にとって，正しく英文を作成することには困難が伴うが，その困難さを軽減する一つの方法として，英語表現集などを活用して，書きたい内容に近い英語表現を参照することが考えられる．英語表現集はいくつか出版されているものの（例えば文献 [4]），それらに記載されている表現や用例の数は十分に多いとは言い難い．

この問題を解決するために，酒井らは，大量に電子化された英語論文から論文作成に有用な表現を自動的に獲得する手法を提案している [3]．酒井らの手法では，英語論文中に出現する単語列の中から，その出現頻度などを用いて，表現として有用な単語列を選別する．しかし，この手法では，英文中の単語の表層的な順序関係しか考慮しないため,「文中の離れた場所に単語が出現するような表現を獲得できない」,「単語間に関係性が存在しないような単語列を表現として誤って獲得してしまう」といった問題が生じる．

これらの問題を解決するために，本稿では，構文構造を利用した英語表現獲得手法を提案する．本手法では，構文的関係の一つである依存関係を利用して，英語論文から英語表現を抽出する．依存関係は，単語間の修飾・被修飾の関係を表す．依存関係は，英文上で離れて出現する単語の間にも存在するため，依存関係で連結された単語列を抽出することにより，文中の離れた場所に単語が出現するような表現を獲得できる．また，獲得する単語列を，依存関係で連結された単語列に制限することにより，得られる表現が構文的まとまりを有することを保証できる．

本手法の有効性を確認するために，評価実験を行った．実験には，ACL の 8 年分の論文を使用した．人手により有用か否かが判断されている単語列 500 個を評価用データとし，提案手法の精度と再現率を評価したところ，精度が 38.0%，再現率 81.5%であった．

2 英語表現とその獲得

本研究の目的は，英語表現集に記載できるような英語表現を自動的に獲得することにある．本節では，まず，英語表現の獲得に関するイメージを掴むために，具体例を交えながら英語表現について説明する．例として，以下の単語列について考える．

(2-1) In this paper, we describe ...
(2-2) The reason why ... is that ...
(2-3) For instance, it ...

単語列 (2-1) は，論文での目的を述べる場面で利用できる表現である．(2-2) は理由や根拠を述べるときに使用できる．これらの表現は，英文を作成するときに参考となる表現であると考えられる．一方，(2-3) については，例を示すときに利用できないわけではないが，主語として “it” を使用する必然性はなく，表現としては “For instance, ...” の方が好ましいと考えられる．本研究が目指すのは，(2-1)，(2-2) のような表現のみを英語論文から自動的に獲得することである．

2.1 英語表現獲得の関連研究

酒井らは，大量の英語論文から論文執筆に有用な英語表現を自動的に抽出する手法を提案している [3]．酒井らの手法では，英語論文に出現する連続する単語列の中から，その出現頻度などを考慮し，英語論文作成に有用な単語列を選別する．この手法では，単語の表層的な順序関係のみを利用して表現を抽出する．そのため，論文作成に有用であるにも関わらず獲得できないような表現が存在する．例えば，酒井らの手法では表現 (2-1) は獲得できるが，表現 (2-2) は原理的に獲得できない．なぜならば，表現 (2-2) を構成する “The reason why” と “is that” は一般に，英文において離れて出現するが，単語が連接しないこのような表現を酒井らの手法では獲得できないからである．また，論文中に頻出すれば，(2-3) のような単語列を誤って獲得してしまう場合がある．

2.2 英語表現にみられる構文的特徴

酒井らの手法における問題は，英文中の単語の表層的な順序関係しか考慮していないために発生していると考えられる．そこで，順序関係とは異なる観点として構文的関係の一種である依存関係に注目し，有用な表現にみられる特徴を考える．

前節の例を考えると，(2-1)，(2-2) に関しては，表現を構成する各単語が別の単語との間に依存関係を有している（図 1 参照）．一方，(2-3) において，“it” は “For” とも “instance” とも依存関係は持たない．これは一例にすぎないが，一般に，英文作成に有用な表現を構成する単語列の単語間には，依存関係が存在すると考えられる．したがって，獲得する表現を，表層的に連結した単語列ではなく依存関係により連結された単語列とすることにより，英文上で離れた位置に出

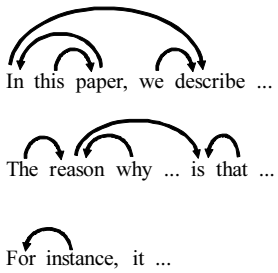

図 1: 英語表現中の依存関係

現する単語から構成される表現を獲得できるようになると同時に，有用でない表現の獲得を抑制できると期待できる．

3 依存関係に基づく英語表現獲得

前節で述べたように，英語表現の自動獲得の問題において，依存関係は一つのキーとなると考えられる．そこで本節では，依存関係を利用した英語表現の獲得手法を提案する．本手法の概略を以下に示す．まず，英文集合中の英文に対して依存関係を与える．この依存関係は文ごとに一つの木構造を構成する．次に，この木構造に出現する木構造のパターンを抽出する．この木構造パターンは，依存関係で連結された単語列と対応している．最後に，木構造パターンに関する統計情報を利用して，抽出された木構造パターンの中から表現として有用なものを選別し，木構造パターンから単語列を復元して英語表現を得る．

3.1 依存関係に基づく木構造の構築

本稿で提案する手法のポイントは，依存関係を利用することにある．一般に，英文に対する依存関係は一つの木構造を構成するので，まず，その点から説明を始める．

英文中の各単語をノードとし，依存される単語を親ノード，依存する単語を子ノードと定めると，文に対して一つの木構造が与えられる．依存関係で連結された単語列を取り出すことは，この木構造に含まれる任意の木構造パターンを抽出することに相当する．本手法は，依存関係が構成する木構造から木構造パターンを抽出し，抽出されたパターンの中から，表現として有用な単語列が得られるようなものを求める手法と位置付けることができる．

ただし，上記の木構造をそのまま使用すると，表現を獲得する際に問題が生じる．というのも，この木構造には次のような情報が含まれないためである．

1. 単語の順序関係を復元するための情報が含まれていない．w_c が w_p の子ノードであるような木構造パターンに対応する単語列が，$\cdots w_c \cdots w_p \cdots$ なのか，$\cdots w_p \cdots w_c \cdots$ なのかを決定できない．
2. 表現に使用される省略記号 “...” に対応する情報が含まれていない．英語表現集には，省略を表す記号がしばしば使用されるが，このような情報は上述の木構造上には存在しない．

これらの問題を回避するために，上記の木構造を次のように変更する．

1. 各単語に対応するノードは，2 つの子ノードを持つ．それらのラベルは，LEFT と RIGHT である．単語 w_d が w_h に左から依存するとき，w_d に対応するノードを，LEFT の子ノードとする．右から依存するときは，RIGHT の子ノードとする．依存関係の方向に示すノード LEFT と RIGHT を加えることにより，単語間の順序関係を決定できる．
2. 各ノードに対して，その単語が構成する句や節などの構成素の情報を付加する．木構造パターンに句や節を表わすラベルが含まれる場合，これを省略表現に置き換えて表現を復元する．

3.2 木構造パターンの抽出

依存関係で連結された単語列を得るため，本手法では，大量の英語論文中の英文を 3.1 節で述べた木構造で記述し，この木構造から木構造パターンを抽出する．木構造集合に含まれるすべての木構造パターンを抽出するのは，データサイズの面からみて現実的ではない．そこで本手法では，木構造パターンの抽出において閾値を設定し，出現頻度が閾値以下のパターンを抽出しない．これにより，効率的にパターンを抽出する．

木構造パターンの抽出は，木構造マイニングアルゴリズム FREQT[1] をベースとしている．FREQT は，木構造集合から，ある閾値以上出現する木構造パターンを抽出する手法である．FREQT では，初めに，木構造集合から，サイズが 1 つまり単一のノードからなるパターンを列挙する．これらの中から，出現頻度があらかじめ定められた閾値以上のものを，サイズが 1 の頻出パターンとする．次に，サイズが 1 の頻出パターンに新たにノードを一つ追加することにより，サイズが 2 の頻出パターン候補を列挙する．この候補に対して，閾値以上出現するものをサイズが 2 の頻出パターンとする．以降，サイズを 1 ずつ増やしながら同様の操作を行い，頻出パターンが得られなくなるまで繰り返す．これにより，出現頻度が閾値以上の木構造パターンを効率的に得ることができる．

木構造パターンの抽出は，FREQT をベースとしているが，いくつかの点を変更している．これは，表現に含まれる省略記号に関わるものである．本手法では，省略記号を使用した表現を抽出するために，木構造のノードラベルとして構成素を与えている．例えば，“the reason why ... is that ...” といった表現は，図 2 に示す木構造パターンから得られるが，このようなパターンを抽出するために，頻出パターン候補の列挙方法を次のように変更する．

- 頻出木構造パターン候補を列挙するとき，構成素をラベルにもつノードについては，単語をラベル

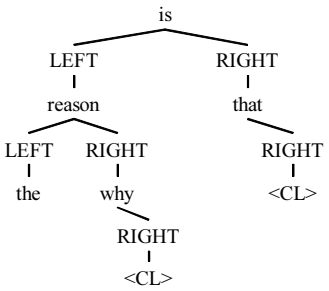

図 2: “the reason why ... is that ...” に対応する木構造パターン

にもつノードと，構成素をラベルに持つノードをそれぞれ別の木構造パターンとして列挙する．

また，頻出木構造パターン候補を列挙する際，以下のケースについては，ノードの追加を行わない．

- パターン中の構成素をラベルに持つノードに対しては，その構成素を構成する単語ノードを追加しない．

3.3 表現の有用性の判定

3.2 節の方法により，木構造集合において閾値以上の頻度で出現するパターンを得ることができる．このパターンは，依存関係で連結された単語列に対応しているため，閾値以上の頻度で出現する依存関係で連結された単語列を得ることができる．しかし，これらの単語列がすべて有用というわけではない．そこで本手法では，有用な表現のみが得られるように，木構造パターンに関する統計情報を利用して，表現として有用な木構造パターンを選別する．

3.3.1 統計情報を利用した判定

3.2 節の方法により得られた木構造パターンの中には，他のパターンと比較すると不要になるものが存在する．例えば，以下の単語列に対応するパターンを考える．

(3-1)the reason why ... is

(3-2)the reason why ... is that

(3-1) は (3-2) に含まれている．単語列 (3-1) に対して，“that” が高い頻度で共起する場合，(3-2) の方がより有用であると考えられるため，表現としては，(3-2) のみを採用し，(3-1) は除去した方がよい．

そこで本手法では，ある表現の一部を構成するような単語列を，統計情報を活用して除去する．本手法では，木構造パターン中の各ノードに対して，そのノードに対応する木構造集合中のノードに連結されるノードのラベルが，特定の単語，あるいは特定の構成素であるかどうかの程度を，エントロピーを用いて評価する．エントロピーが小さいことは，その木構造パターンに特定の単語や構成素が連結しやすいことを表わす，すなわち，そのパターンがある表現の一部を構成する

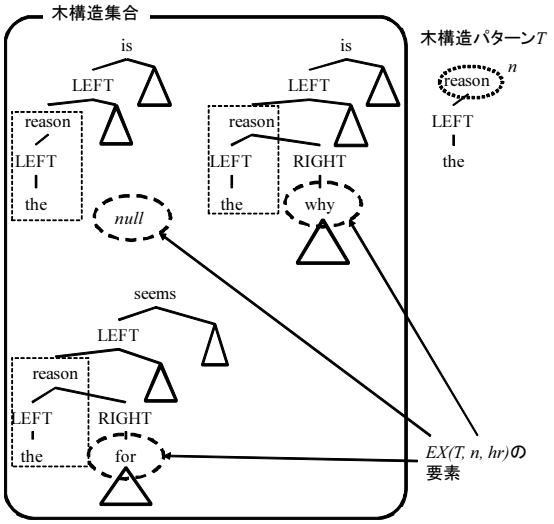

図 3: $EX(T, n, d)$ の例

可能性が高いことを意味する．本手法ではこのような木構造パターンを除去する．

以下では，エントロピーの計算方法を導くが，そのためにいくつかの記法を導入する．まず，木構造パターン T のノード n に連結しうるノードについて定義する．ここで，ノードの連結の仕方にはいくつかの場合がある点に注意する．n が T の根ノードのとき，n に連結するノードとしては，n が依存するノードと n に依存するノードが存在する．根ノードでないときは，n に連結するノードは，n に依存するノードのみである．また，ノードが右から依存するのか，左から依存するのかの違いもある．つまり，ノードの連結の仕方には $2 \times 2 = 4$ 通りの場合がある．以下では，それぞれの場合を，dr, dl, hr, hl と書くことにし，木構造パターン T のノード n に連結するノードのラベルの集合を $EX(T, n, dr)$ のように書く．ただし，n に連結する各ノードは，単語と構成素の 2 つのラベルを持つ場合があるが，この場合，構成素ラベルのみを，$EX(T, n, dr)$ の要素に加えるものとする．また，n に対応する木構造集合中のあるノードが連結するノードを持たないとき，特別な要素 $null$ を $EX(T, n, dr)$ に加える．$EX(T, n, dr)$ は，T のノード n が右から依存するようなノードのラベルの集合である（図 3 参照）．より正確には，n に対応する木構造集合中のノードが右から依存するようなノードに付与されたラベルの集合である．dl, hr, hl についてもほぼ同様に定義され，それぞれ，「n が左から依存するノードのラベルの集合」,「n に右から依存するノードのラベルの集合」,「v に左から依存するノードのラベルの集合」である．T 中のノード n が右から依存するノードのラベルの分布に関するエントロピーは次の式で定義する．

$$H_{T,n,dr}(L \mid T) = -\sum_{l \in EX(T,n,dr)} P(l \mid T, n, dr) \log P(l \mid T, n, dr)$$

dl, hr, hl についても同様に定義する．$P_{T,n,dr}(l|T)$ は木構造パターンの出現頻度に基づき計算する．すなわ

表 1: 実験結果

	精度 (%)	再現率 (%)	F 値
提案手法	37.8%	81.5%	51.7
酒井ら	23.5%	72.8%	35.5

ち，次の式を用いる．

$$P(l \mid T, n, dr) = \frac{C(expand(T, n, dr, w))}{\sum_{l \in EX(T,n,dr)} C(expand(T, n, dr, l))}$$

ここで，$C(\cdot)$ は木構造パターンの出現頻度を表わす．$expand(T, n, dr, l)$ は，木構造パターンのノード n の親ノードとして RIGHT ノードを追加，さらにその RIGHT ノードの親として l をラベルに持つノードを追加して得られる木である．

上記のように定義したエントロピーを利用して，木構造パターンに対応する単語列が，ある表現の部分となっているか否かを判定する．以下の条件を満たす n 及び $d \in \{dr, dl, hr, hl\}$ が存在するとき，T を除去する．

1. $P(null|T, n, d) < \alpha$
2. $H_{T,n,d}(L|T) < \beta$

α，及び β は，事前に定めた閾値である．1. が成り立つことは，n に対応する木構造集合中の多くのノードにおいて，連結するノードが存在することを意味する．2. が成り立つことは，その連結するノードには，特定の単語あるいは構成素がラベル付けされていることを意味する．

4 評価実験

提案手法の有効性を確認するために，評価実験を行った．実験には，ACL の 2001 年から 2008 年までの論文中の英文 165,116 文を使用し，英語表現を提案手法により抽出した．評価用データとして，人手により有用か否かが判断された単語列 500 個を使用した．

提案手法による英語表現の獲得においては，英文に対して依存関係を与えなければならないが，本実験では，各英文に対して構文解析器 Enju[2] により句構造を付与し，Pennconverter[5] を用いて付与された句構造を依存構造に変換した．

有用な表現を選別する際に使用する閾値は，それぞれ，$\alpha = 0.5$, $\beta = 1.3$ とした．頻出パターンとしては，1,925,449 個の木構造パターンが抽出されたが，そのうち提案手法により有用と判定されたものは，127,059 個であった．最終的に得られた xx 個の表現について，評価用データである 500 個の単語列と比較し，提案手法の精度と再現率を求めた．結果を表 1 に示す．酒井らの手法では，統計的特徴に加えて，人手により作成したアドホックなルールを用いて有用でない単語列の除去を行っているが，表中の精度・再現率は，統計的特徴のみを使用した場合の値である．精度，再現率ともに，酒井らの手法よりも上回っており，提案手法の有効性を確認できた．

提案手法により獲得できた表現の例を示す．

(4-1) the fact that ... suggests that ...

(4-2) since ... we can conclude that ...

この例が示すように，英文において単語が離れて出現するような表現を，提案手法は獲得することができる．

5 おわりに

本稿では，構文構造を利用した英語表現の獲得手法を提案した．本手法では，英文を依存関係に基づく木構造で記述し，その木構造から木構造パターンを抽出することにより，依存関係で連結された単語列を獲得する．獲得された単語列に対して，統計情報を利用して，単語列の表現としての有用性を判定し，英文作成に有用な表現を獲得する．提案手法の有効性を確認するために，評価実験を行った結果，表層的な順序関係のみを考慮した従来手法よりも高い性能を示し，依存関係の利用が，英語表現獲得に有用であることを確認した．

今後の課題として，論文中での出現頻度が低い英語表現を獲得する方法について検討する必要がある．本手法では，木構造パターン抽出における効率を優先した結果，低頻度の英語表現は獲得できないという問題がある．対象とする論文データの規模をより大規模にすることにより，この問題はある程度解消されると考えられるため，そのような評価実験を今後行いたい．

謝辞

本研究の一部は，公益財団法人 栢森情報科学振興財団の助成を受けて遂行された．

参考文献

[1] Asai et al.: Efficient Substructure Discovery from Large Semi-Structured Data, *Proc. of 2nd SIAM Inter. Conf. on Data Mining*, pp.158–174, 2002.

[2] Miyao and Tsujii: Feature Forest Models for Probabilistic HPSG parsing, *Computational Linguistics*, 34(1), pp.35–80, 2008.

[3] 酒井ら：英語論文からの表現集の自動生成，言語処理学会第 16 回年次大会発表論文集，pp.375–378, 2010.

[4] 崎村：英語論文によく使う表現，創元社，1991.

[5] http://fileadmin.cs.lth.se/nlp/software/pennconverter/pennconverter.jar