情報処理学会研究報告
IPSJ SIG Technical Report

構文構造を利用した英語論文からの表現の自動獲得

葛原和也[†1]　加藤芳秀[†2]　松原茂樹[†1]

本論文では，英文を構成する単語間の構文的関係，および，統計情報を利用して英文作成に有用な表現を獲得する手法を提案する．本手法では，まず，構文的関係の一つである依存関係を利用することにより，構文的なまとまりが存在する単語列を獲得する．それらの単語列に対して，頻度などの統計情報を利用して，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．表現の獲得実験によって，提案手法の有用性を確認した．

Acquisition of English Expressions Using Syntactic Structures

Kazuya Kuzuhara ,[†1] Yoshihide Kato [†2]
and Shigeki Matsubara[†1]

In this paper, we propose a method to acquire English expressions useful for writing English sentences. The method uses syntactic relations between words and statistical information. It acquires sequences of words which are connected with syntactic relations, and selects useful expressions from these sequences using statistical information. As the result of an experiment, we confirmed the effectiveness of our method.

1. はじめに

英語を母語としない研究者にとって，正しい英文を作成することは難しい．英文作成に伴う困難さを軽減する一つの方法として，辞書や英語表現集などを利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．しかし，辞書や表現集に記載されている表現で十分であるとは言い難い．この問題を解決するために，酒井ら[1)]は大量の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．

酒井らの手法では，英文中の単語の表層的な順序関係のみを考慮する．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，酒井らの手法では構文的関係などを考慮していないため,「文中の離れた場所に出現する単語が関係を持つ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誤って獲得してしまう」という問題が生じる．

これらの問題を解決するために，本論文では，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，英単語間の構文的な関係である依存関係を利用して，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している単語間にも存在するため，依存関係が存在するような単語列を英語表現として獲得することにより，文中の離れた場所に出現する単語が関係を持つような表現も獲得できる．また，獲得する単語列を，単語間に直接の依存関係が存在するものに制限することにより，獲得される表現が構文的なまとまりを有することを保証できる．

本手法では，英文の依存関係を木構造として表現し，木構造マイニングアルゴリズム FREQT[2)] に基づき，頻出する依存関係のパターンを抽出する．これらの抽出されたパターンから，統計情報を利用して英文作成に有用なものを選別し，英文作成に有用な表現として獲得する．

本手法の有効性を確認するため，評価実験を行った．頻出する依存関係のパターンを，有用と判定されたものと，そうでないものに分け，それぞれに対して，英文作成に有用な表現であるかを人手により判定した．実験の結果，表層的な順序関係のみを考慮した手法と比較して，F 値において 7.25 ポイント高い値を示した．また，依存関係を利用することで，有用でない表現の獲得を抑制できることを確認した．

本論文の構成は以下の通りである．2 章では，本論文で扱う英文作成に有用な表現，および，その獲得に関する従来の手法について述べる．3 章では，提案する英語表現獲得手法について説明する．4 章で評価実験について報告する．5 章では関連研究について述べ，最後に 6 章で本論文をまとめる．

2. 英語表現とその獲得

本章では，英語表現の獲得に関する従来の手法について概観するが，まず英語表現の獲得

†1 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University
†2 名古屋大学情報基盤センター
Information Technology Center, Nagoya University

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[1] 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．酒井らの手法では，英文中の単語の表層的な順序関係のみを考慮する．英語論文に頻出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，基本句と呼ばれる最小の句の認識は行うものの，構文解析などの処理は行わない．その結果として，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，単語が連接しない表現を酒井らの手法では獲得できないからである．また，字面上で頻出すれば，(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 英文作成に有用な表現の特徴

前節で述べた問題は，単語の表層的な順序関係のみを考慮した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に注目し，有用な表現に観られる特徴を考える．

2.1 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何らかの構文的関係 (例えば，主語と動詞の関係など) を有している．一方，(2-3) において，“it” は “For” とも “instance” とも直接構文的な関係を持たない．このように，英文作成に有用な表現には，表現を構成する単語間に構文的なまとまりが存在していると考えられる．

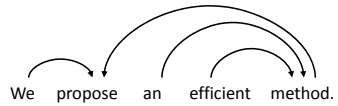

図 1 “We propose an efficient method” の構成単語間の依存関係

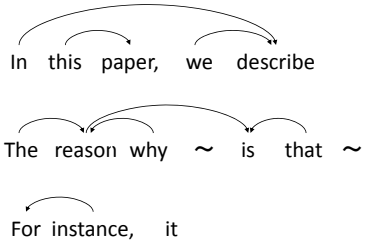

図 2 英語表現中の依存関係

3. 英文作成に有用な表現の獲得

前章で述べた特徴を持つような英語表現を獲得するため，本章では，構文的関係を利用した英語表現の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，構文的なまとまりが存在する単語列を獲得する．その単語列を，統計的特徴を利用して選別し，英文作成に有用な表現を獲得する．

3.1 依存関係の利用

依存関係は，英文を構成する単語間の修飾・被修飾関係を表す構文的関係の一つである．例えば，“We propose an efficient method.” という英文に対して，単語間の依存関係を付与すると，図 1 のようになる．

2.1 節で示した (2-1),(2-2),(2-3) を構成する単語間の依存関係は図 2 の通りである．英文作成に有用な表現である (2-1)，および，(2-2) においては，構成単語間に依存関係が存在している．(2-2) は，構成単語が文中において表層的には離れて出現するため，酒井らの手法では獲得できないが，依存関係で結合された単語列を取り出すことができれば，(2-1) のような構成単語が連接している表現だけでなく，文中の離れた場所に出現する単語が関係を持つ (2-2) のような表現を獲得することが可能となる．

一方，有用ではない (2-3) においては，“For instance” と “it” の間に依存関係は存在しない．そのため，獲得する表現を，単語間に依存関係が存在する単語列に制限することにより (2-3) のような構文的にまとまりのない単語列の獲得を避けることができる．

3.2 依存関係を利用した表現獲得手法

本節では，依存関係を利用した英語表現の獲得手法について述べる．

本手法の概略は以下の通りである．まず，英文集合中の英文の依存関係を木構造として記述する[⋆1]．次に，この木構造集合に出現する頻出パターンを抽出する．これらの抽出されたパターンに対して，統計情報を利用して表現としての有用さを判定し，有用なパターンを選別し，英文作成に有用な表現を獲得する．

3.2.1 依存関係に基づく木構造集合の構築

本手法では，木構造を利用して英語表現を獲得する．そのため，まず英文の依存関係を木構造として表現する方法について述べる．

依存関係が付与された英文に対し，英文中の各単語をノードとし，修飾される単語を親ノード，修飾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．単語が左側から依存するのか右側から依存するのかによりその単語の構文的役割は異なるため，これを区別できるようにこの木をさらに次のように拡張する．各単語に対応するノードに，二つの子ノードを与える．それらのノードのラベルはそれぞれ LEFT, RIGHT である．単語 w_d が単語 w_h に左側から依存するとき，単語 w_d に対応するノードを LEFT の子ノードとする．w_d が右側から依存するときは，RIGHT の子ノードとする．これにより，単語間の依存関係を木構造として表現できるが，さらに，ノードに対して句や節等の情報を付与する．このような情報を付与するのは，以下の理由からである．すなわち，英語表現には，英単語の他に句や節などが挿入されることを意味する記号 “〜” が含まれるが，この記号を含む表現を得るためには，そのような情報が必要となるからである．ノードが句や節等の構成素をラベルとしてもつことは，そのノードを含む子孫ノードからなる単語列がその構成素であることを意味する．

図 3 右に英文 “We say that the paths are collapsed.” に対する依存関係を表現した木構造を示す．これは図 3 左の依存関係を表現している．また，この英文において，“we” と “the paths” が名詞句 (<NP>)，“the paths are collapsed” が節 (<CL>) を構成していることを表している．

⋆1 依存関係を定める具体的な方法については 4 章で述べる

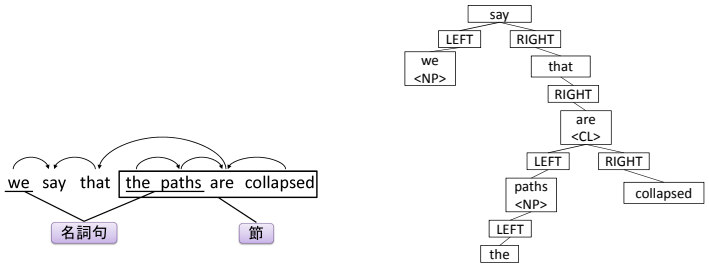

図 3 木構造表現の例

3.2.2 木構造集合からの頻出パターンの抽出

本手法では，木構造マイニングアルゴリズム FREQT[2] に基づき，木構造集合から頻出パターンを抽出することにより，依存関係で結合されている単語列を獲得する．FREQT では，サイズが 1，つまり単一のノードからなる頻出パターンをまず見つけ，パターンに新たなノードを一つずつ追加することにより，出現頻度が閾値以上のパターンを効率的に列挙できる．本手法では，FREQT に基づき，木構造集合において閾値以上出現するパターンを抽出するが，その際に以下の 2 点を拡張している．

- **LEFT ノード，RIGHT ノード**
  LEFT ノード，RIGHT ノードは依存の方向を表すノードである．そのため，これらのノードが根や葉であるようなパターンは抽出しない．例えば，図 4 のパターンは閾値以上出現していれば抽出するが，図 5 はそれぞれ，根や葉の位置に LEFT ノードや RIGHT ノードが存在しているため，たとえ閾値以上出現していても抽出しない．
- **構成素ラベル**
  パターンを列挙する際に，構成素をラベルとして含むノードに関しては，単語をラベルに持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パターンとして列挙する．例えば，“we propose” という単語列に対応するパターンに “method” と構成素 “<NP>” をラベルに持つノードを追加する場合は図 4 のように，“method” と “<NP>” をラベルにもつノードを追加したパターンをそれぞれ別のパターンとして列挙する．また，構成素をラベルにもつノードの子孫ノードは，その構成素を構成する要素であるため，パターンに新たなノードを追加する際には，構成素をラベルに持つノードには新たに子

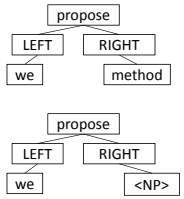

図 4 抽出パターン例

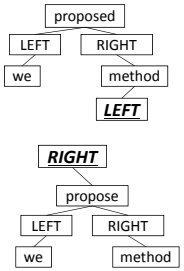

図 5 非抽出パターン例
(根や葉に LEFT ノード, RIGHT ノードが存在)

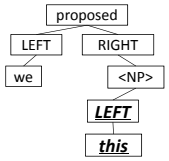

図 6 非抽出パターン例
(構成素以下にノードが存在)

ノードを追加しない．つまり，構成素 “<NP>” をラベルにもつノード以下に別のノードが存在している図 6 のようなパターンは，列挙されない．

以上のように拡張した方法により，依存関係で結合された単語列をその依存関係を表現する木構造上のパターンとして獲得する．

3.2.3 統計情報を利用した表現の選別

前節で抽出されたパターンに対して統計情報を利用して英文作成に有用な表現を選別する．英文作成に有用な表現に見られる統計的な特徴について酒井らは次の点を指摘している．

- **論文中に頻出する**
  英語には，特定の構文や言い回しといった，定型的な表現が存在する．このような定型的な表現は英文作成に有用な表現であり，多くの英文で使用されていると考えられる．そのため，統計的特徴の一つとして，頻出するということが挙げられる．
- **短すぎない**
  短すぎる表現には，単語や句動詞，複合名詞といった表現が多く，英文作成の際に参考になる表現であるとはいえない．そのため，英文作成に有用な表現は短すぎないという特徴を持つ．
- **連接する語の種類が多い**
  有用な表現は，様々な場面において使用されると考えられる．その場合，その表現と共起し，連接するような語の種類が多くなると考えられる．そのため，連接する語が多いことも統計的特徴として挙げられる．

酒井らは，これらの統計的特徴を利用して表現を選別しているが，3 つめの特徴については，単語の連接を基にしているため，そのままでは本手法で利用できない．

酒井らの手法では，表層的な関係しか考えていないため，表現に連接する単語としては，その表現の左右にどのような単語が出現しうるかを考慮するのみである．表現の左右に出現する単語の分布に関するエントロピーを求め，エントロピーが大きいほど有用な表現であるとしている．エントロピーが大きいことは，その表現に連接する語の種類が多いことを表しており，その表現が様々な文脈で使用できることを意味しているからである．

一方，本手法のように，木構造上で連接する単語を考慮する場合，考慮すべきケースは複雑になる．具体的には，以下のような単語を，連接する単語として考える必要がある．

- パターンのルートノードに対応する単語が依存しうる単語
- パターンの各ノードに関して，それに対応する単語に依存しうる単語

本手法では，パターンの各ノードに関して，そのノードに対応する単語に連接する上記のような単語の分布に関するエントロピーを求め，それにより，パターンが様々な文脈で使用できるかどうかを評価する．以下では，エントロピーの計算方法を導くが，そのために，まずいくつかの記法を導入する．G をパターン，v を G 中のノード v とするとき，$EX(G,v,d)$ を v に対応する単語に連接する単語の集合とする．ただし，$d \in \{dr, dl, hr, hl\}$ であり，$EX(G,v,dr)$ は，v に対応する単語が右から依存する単語の集合である．dl, dr, dl についてはそれぞれ,「v に対応する単語が左から依存する単語の集合」,「v に対応する単語に右から依存する単語の集合」,「v に対応する単語に左から依存する単語の集合」である．G 中のノード v が右から依存する単語の分布に関するエントロピーは次の式で定義される．

$$H_{G,v,dr}(W|G) = -\sum_{w \in EX(G,v,dr) \cup \{null\}} P_{G,v,dr}(w|G) \log P_{G,v,dr}(w|G)$$

ここで，w が $null$ であることは，パターン G のノード v に対応する単語が他の単語に依存しないことを意味する．dl, hr, hl についても同様である．$P_{G,v,dr}(w|G)$ は，パターンの出現頻度に基づき計算する．すなわち，次の式を用いる．

$$P_{G,v,dr}(w|G) = \frac{C(expand(G,v,dr,w))}{\sum_{w' \in EX(G,v,dr) \cup \{null\}} C(expand(G,v,dr,w'))}$$

ここで，$C(\cdot)$ は，パターンの出現頻度を表す $expand(G,v,dr,w)$ は，パターン G のノー

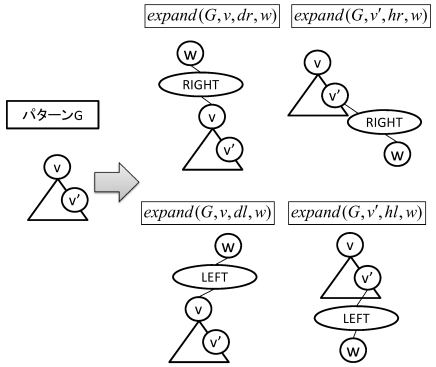

図 7 木構造の拡張

ド v 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 w をラベルに持つノードを追加して得られるパターンである（図 7 参照）．

エントロピーに加え，英文集合中での表現の出現頻度，表現の長さを考慮して，表現の有用さを測るスコア関数を以下のように定義する．

$$SCORE(G, v, d) = \log(C(G)) \times size(G) \times H_{G,v,d}(W|G)$$

$size(G)$ は G 中に出現する単語をラベルにもつノードの数である．このスコア関数は酒井らのスコア関数のエントロピー部分を我々のものに置き換えたものである．このスコア関数を利用し，有用な表現を選別する．具体的には，全ての v，x，および，$d, d' \in \{dr, dl, dr, dl\}$ に対し，以下の条件を満たすようなパターン G を有用な表現として獲得する．

$SCORE(G, v, d) > SCORE(expand(G, v, d, x), x, d')$

この式により，パターン G を包含し，かつ，G よりも有用だと判定されるようなパターンが存在した場合，G を除去する．これにより，有用な表現の選別を行う．

4. 評価実験

4.1 実験概要

提案手法の有効性を確認するため，評価実験を行った．英文データには，ACL の 2001 年から 2008 年までの論文の 165,116 文を利用した．各英文に対して，構文解析器 Enju[3] を用いて句構造を付与し，Pennconverter[4] を用いて付与された句構造を依存構造に変換した．これらのデータから，提案手法に基づいて木構造集合を作成した．出現頻度が 6 回以上のパターンを抽出したところ，518,747 個のパターンが抽出された．なお，今回の実験では単語数が 2 語以下のパターンは取り除いた．抽出されたパターンをスコア関数に基づいて選別した結果，276,380 個のパターンが有用だと判定され，242,367 個のパターンが有用ではないと判定された．

提案手法によって得られた 276,380 個の表現について，有用な表現が含まれているかを評価した．酒井らが評価に用いた有用か否かが人手により判断されている単語列 500 個を正解データとし，提案手法の，精度，再現率，F 値を算出した．なお，精度と再現率は以下のように算出される．

$$\text{精度} = \frac{\text{獲得された有用な表現の数}}{\text{獲得した表現数}}$$

$$\text{再現率} = \frac{\text{獲得された有用な表現の数}}{\text{正解データ中のすべての有用な表現の数}}$$

酒井らの手法では，統計的特徴による判定の他に，品詞や基本句の種類に基づいて除去ルールを作成し，それを利用して不要な表現を除去している．そこで本実験では，酒井らの手法のうち，統計的特徴のみを利用した手法と比較することで，本手法の英語表現抽出における依存関係の利用，および，統計的特徴による判定の有用性を評価する．

4.2 実験結果

実験結果を表 1 に示す．今回の実験で利用した表現のうち，人手によって有用であると判定された表現は 81 個であった．提案手法は，再現率において，酒井らの手法よりも低い値を示しているが，精度では上回った．F 値においては 7.25 ポイント高い値を示した．また，酒井らの手法では有用であると判定されたが，人手では有用ではないと判定された 194 表現に対して，構成単語間に依存関係が存在するかを調査した．その結果，47 個の表現において構成単語間に依存構造が存在していなかった．これにより，依存関係を利用することで，有用ではない表現の獲得を抑制できることが示され，本手法の利用可能性を確認した．

表 1 実験結果

	精度 (%)	再現率 (%)	F 値
提案手法	34.33 (44/134)	56.79 (46/81)	42.79
酒井ら	23.51 (59/251)	72.81 (59/81)	35.54

表 2 獲得された表現例

as a result,
in the sections, we describe <NP>
the fact that <CL> indicates that <CL>
we have focused on <NP>
<NP> have been proposed by <NP>

本手法によって獲得された表現の例を表 2 に示す．“in the sections, we describe <NP>” のように表層的な順序関係によって獲得できる表現だけでなく，“the fact that <CL> indicates that <CL>” のように構成単語が英文中で離れて出現するような表現も獲得できている．

4.3 考　　察

ここでは，酒井らの手法と比較して再現率が低い値を示した理由について考察する．

人手によって正解と判定されたが，本手法によって有用ではないと判定された 35 表現のうち，14 表現に関しては，木構造集合から頻出パターンとして抽出することができていなかった．これらの表現を表 3 に示す．これらの表現のうち 8 表現では，先頭の名詞句を後続する前置詞句，または，過去分詞句が修飾している．本手法が用いた依存関係を表現する木構造は，構成素に何か別の要素が依存するような関係を表現できないため，これらの表現を獲得できない．

5. 関 連 研 究

英語表現の自動獲得に関連する研究として，multi-word expression(MWE) の獲得に関する研究が挙げられる．そのような研究は数多く行われており，構文解析を利用する手法もいくつか提案されているが[5),6)]，獲得される表現は限定的である．依存関係にある 2 つの単語，あるいは動詞と名詞のペアなどから MWE を構成するものを選別する程度である．本論文で提案した手法のように，様々な長さの表現を扱うことはできない．

表 3 抽出されなかった表現

<NP> for several reasons
<NP> listed in <NP>
<NP> on the development set
<NP> of the target words
<NP> divided by the total number of <NP>
<NP> described below
<NP> with the number
<NP> along with
in the literature ,
otherwise
coreference resolution
in section <digit> ,
in addition
if any

6. ま と め

本論文では，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，表層的な順序関係のみを考慮する手法における,「単語間に関係性が存在しないような表現を誤って獲得してしまう」,「文中の離れた場所に出現する単語列が関係を持つ表現が獲得できない」という問題点を解決するために，英文に依存関係を付与し，依存関係によって結合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表現の有用性を判定し，英文作成に有用な表現を獲得する．提案手法の有用性を確認するため，評価実験を行った．実験の結果，依存関係を利用しない手法と比較して F 値において，7.25 ポイント高い値を示し，本手法の利用可能性を確認した．また，依存構造を利用することにより，有用でない表現の獲得を抑制できることや，構成単語が英文中で離れて出現するような表現を獲得できることを示し，依存関係の利用が，英文作成に有用な英語表現の獲得に寄与することを確認した．

今後の課題として，スコア関数を洗練するために，スコア関数の各項に対して重み付けを行うことや，構文的なまとまり以外の観点から英文作成に有用な表現の特徴を考え，それらを考慮したスコア関数を導入することが考えられる．

将来的には，獲得された表現の提示方法を検討し，英文作成支援のためのシステムを開発することを考えている．

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) 酒井佑太, 小澤俊介, 杉木健二, 松原茂樹: 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.
2) T. Asai, K. Abe, S. Kawasoe, H. Sakamoto, and S. Arikawa: Efficient substructure discovery from large semi-structured data, Proceedings of 2nd SIAM International conference on Data Mining (SDM'02), pp.158-174, 2002.
3) Y. Miyao, J. Tsujii: Feature forest models for probabilistic HPSG parsing. Computational Linguistics, 34(1), pp.35-80, 2008.
4) http://fileadmin.cs.lth.se/nlp/software/pennconverter/pennconverter.jar
5) Dekang Lin: Automatic Identification of Non-compositional Phrases, Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), pp.317-324, 1999.
6) Bannard, Colin: A Measure of Syntactic Flexibility for Automatically Identifying Multiword Expressions in Corpora, Proceedings of Workshop on A Broader Perspective on Multiword Expressions, pp.1-8, 2007.