依存構造を利用した英語表現の自動獲得

葛原 和也*，加藤 芳秀，松原 茂樹 (名古屋大学)

Acquisition of English Expressions Using Dependency Structures

Kazuya Kuzuhara, Yoshihide Kato, Shigeki Matsubara (Nagoya University)

1 はじめに

英語論文執筆に関する知識や慣れが十分でない研究者にとって，正確な英語論文を執筆することは難しい．そのような場合，実際に論文で使用されている表現を参照して英文を作成することが有効である．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．連続した単語の並びの中から，単語列の出現頻度などを用いて，有用な表現を選別する．構文的関係などは考慮しないため，「文中の離れた場所に出現する単語が関係をも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表現を誤って獲得する」といった問題が生じる．

この問題を解決するために，本稿では，依存関係を利用した英語表現獲得手法を提案する．本手法では，単語の表層的な順序関係ではなく，構文的な関係である依存関係を用いて英語表現を抽出する．提案手法を実装し，依存関係が英語表現の獲得に寄与することを確認した．

2 酒井らの手法とその問題点

酒井らの手法 [1] では，単語列の英語論文中で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．そのため，論文作成に有用な表現であるにもかかわらず獲得できないような英語表現が存在する．例えば，*"not only 〜 , but also 〜."* という英語表現である．この表現においては一般に，*"not only"* と *"but also"* は離れて出現する (〜の部分には動詞句や節など様々な構成素が挿入される．)．この例のような，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．また，字面上では頻出するが構文的には何のまとまりもない *"For instance, it"* といった表現を誤って獲得してしまう場合がある．

3 依存関係を利用した表現の獲得

上述の問題を解決するために，本節では，依存関係を利用した英語表現獲得手法を提案する．依存関係は単語間の構文的関係の一種であるが，依存関係が構成する木構造上での連接関係を利用することにより，英語表現中の単語が文中において離れて出現しても，その単語間に依存関係が存在すればその表現を獲得できる．また，獲得される表現において単語間には何らかの依存関係が存在することが保証されるため，意味的にまとまりのない表現を獲得してしまう問題を回避できる．

まず，依存関係について説明する．依存関係は文中における単語間の修飾・被修飾の関係を表す．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．さらに，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木を拡張する．各単語 w_h に対応するノードは，二つのノード LEFT, RIGHT を子ノードとしてもつものとする．単語 w_h に左側から依存する単語 w_d が存在するとき，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．このようにして得られる木構造を Fig.1 に示す．図左において矢印は依存関係を示している．右が対応する木構造である．

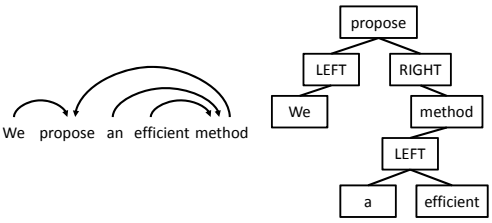

Fig. 1: An example of dependency structure and dependency tree

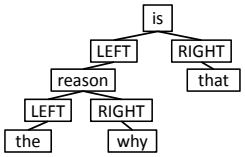

Fig. 2: An example of dependency pattern

英文に対して単語間の依存関係を定め，それを木構造として表現すれば，木構造の集合から頻出パターンを抽出する手法 (例えば [3]) を適用し，頻出する依存関係のパターンを取り出すことができる．

4 評価実験

前節で述べた手法に基づき，表現の獲得実験を行った．英文としては酒井らが利用した英語論文の 206,526 文を利用した．各英文に対して，依存構造解析器 MSTParser[2] を用いて依存構造を付与した．頻出パターンの獲得には，FREQT[3] を使用した．本手法によって獲得されたパターンを Fig.2 に示す．これは，酒井らの手法では獲得出来ない表現である．

次に，酒井らの手法が誤って獲得した不適切な英語表現について検討する．酒井らの手法により獲得された表現の一部については，その表現が有用か否かが人手により判断されている．有用でないと判断された 24 表現について，その表現を構成する単語間に依存関係が存在するかどうかを確認した．24 表現のうち，7 表現については，単語間に依存関係が存在しなかった．この結果から，依存関係を考慮することにより，有用でない英語表現の獲得を抑制できると期待できる．

5 おわりに

本稿では，依存構造を用いた英語表現の獲得手法を提案した．今後は，統計情報などを用いて，有用な表現を選別する手法を開発する予定である．

文献

(1) 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

(2) McDonald et al., Non-projective dependency parsing using spanning tree algorithms. In Proc. HLT-EMNLP, 2005

(3) T. Asai et al., Efficient substructure discovery from large semistructured data, Proc. of the 2nd SIAM International Conf. on Data Mining, pp.158-174, 2002