Vol.2014-ICS-176 No.4
2014/7/22

中心名詞をノードとした文法構造を反映した構文解析ツリーの可視化

相川大輔[†1]　我妻広明[†1†2]

本研究では，曖昧な文章や多義的な構造を含む文脈依存的文章理解のために，人が理解している文法や概念の包含関係を考慮したネットワーク構造について検討を行った．人が使う辞書を用い，文法構造を反映したネットワークを生成する過程で形成される際の文章特異的な分岐や循環構造など構造（トポロジー）の可視化を試みた．

A Visualization of the Parse Tree in Word-Sentence English Matching Test Through an Acquisition Process of Trees with Nodes of Principal Nouns Obtained in Each Sentence

DAISUKE AIKAWA[†1]　HIROAKI WAGATSUMA[†1†2]

We have been proposed and demonstrated a tree structure to discriminate the right answer in the word-sentence English matching test, which is built as the network of “principal nouns” obtained from the tight link referring the English dictionary and thesaurus. In this model, we hypothesized that the level of complexity can be used for judgements to determine which is the right answer, however the evaluation and accuracy could not achieve its desired level. In this report, we attempt to visualize individual differences of the tree structure depending on the depth to dig in the dictionary and step into the improvement of the proposed model.

1. はじめに

人が文章を読むときは，主語（S），動詞（V），目的語（O）などの要素を文法に従って解釈し，辞書にある限定した単語の関係性から，その意味を理解する．近年は統計手法やデータマイニングなど大規模データベース化による構文解析が主流である[1]．一方，自然言語を扱う人間の言語能力を考えてみると，母国語の場合は生まれてからその言語に長く触れてきたことから十分な量の辞書が脳の中に記憶されているとも推量できるが，第二言語は習得に同じだけの量と質を確保できないことは明らかである．そのような不十分な辞書の場合に，意味ネットワークが，構文理解・意味抽出をどれだけ補佐するのかが，本研究の中心課題である．具体的には，文法構造から推定される単語間の関係性（意味の近さ）と，人が使う辞書の小規模の関係性データベースから英文の構文解析を行う場合の有効性を検討する．特に，命令文理解に注目し，SVO 型の目的節（O）の解釈を目指す．“ take a big ball ” など，絶対的基準よりも相対性を活用した文は日常指示文に多く，状況を解決するに十分な構造が名詞をノードとするネットワークから得られると仮説を立て，英文-単語意味整合テストにおいて有効性を検証した．

†1 九州工業大学大学院生命体工学研究科
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology

†2 理化学研究所脳科学総合研究センター
RIKEN BSI

2. 方法

本研究では，SVO 型の命令文において目的語を正確に把握し，ネットワークを形成する方法を提案する．まずネットワーク形成に必要な「中心名詞」について定義する．

2.1 中心名詞の定義

目的節の中心となる名詞を「過去分詞，現在分詞，that 節，what 節，of などの直前の名詞」と定義し，「中心名詞」と呼ぶ．また，過去分詞，現在分詞などの中心名詞を判別するための単語を中心名詞判別単語と呼ぶ．したがって，図 1 では “ female ” が中心名詞となる．

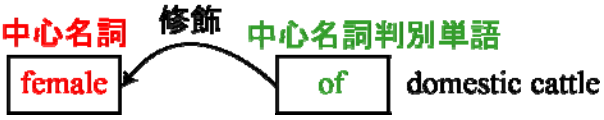


図 1　中心名詞の例

2.2 ネットワーク生成

次に中心名詞ネットワークを定義する．英英辞典（Princeton University, http://wordnet.princeton.edu/wordnet）を用い，ある名詞の意味（辞典に掲載されている説明文）の中心名詞に注目して辿ることを考える．“ cow ” から始まる中心名詞ネットワークは，図 2 のように “ cow ” から “ female ”,“ animal ”と連結される構造が得られる．ここで，赤色は中心名詞，緑色は中心名詞判別単語を表している．この手続きを繰り返すことで，中心名詞ネットワークを生成することができる．また，英英辞典で単語を調べていく回数を，深度（*DEPTH*）とする．図 2 では，“ cow ” から 2 回英英辞典で調べている．よって，*DEPTH* = 2 （以後，

DEPTH 2 のように記述する）のネットワークである．

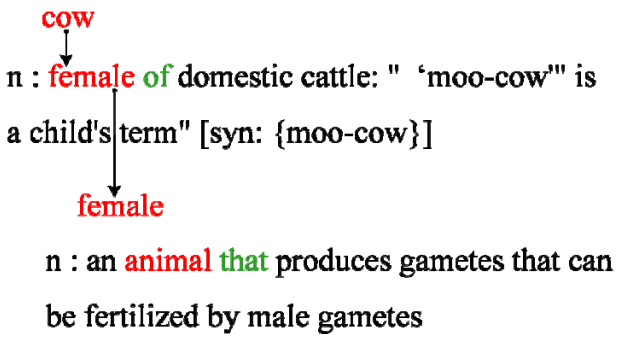


図 2　cow のネットワーク生成

2.3 類義語辞典の活用

中心名詞ネットワークの有効性検証の課題として，図 3 のような英文テストを用意した．この例題の解答は“ sheep ”だが，英英辞典の説明文“ woolly usually horned ruminant mammal related to the goat ”と例文は単純に合致せず，また文に含まれる単語も異なる．また，英英辞典のみを辿る中心名詞ネットワークでは，図 2 に示したように調べれば調べるほど抽象的な説明となり，他の単語から辿った場合の類似性が高まり，単語の意味を区別する目的には適当でない．また同ネットワークはその特性から循環構造になる可能性もある．そこで，深度調整と類義語辞典（ARTFL Project, http://artfl-project.uchicago.edu/）の相補的活用によって考慮する．図 4 に，類義語辞典を併用した“ cow ”の中心名詞ネットワークを示す．ネットワーク生成過程に現れた名詞は，問題文中の名詞と類義語辞典で同じグループに属した場合「関連名詞」として，見出し語を経由してノードとなる．

問題番号100: a farm animal used for its meat and hair
選択肢　(a) cow (b) sheep (c) plant (d) pig

図 3　英語テスト例

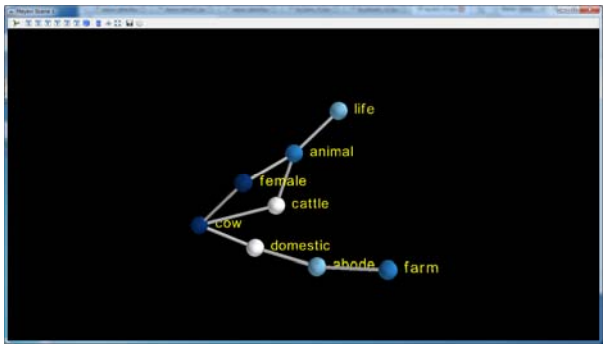


図 4　類義語辞典を併用した中心名詞ネットワーク
（*DEPTH* 2）

2.4 構文解析のための評価値 E

図 3 のような英文テストの解の自動検出のために，中心名詞ネットワークを用いるため，課題解決の評価値を定義する．評価値として，ネットワーク内のノード分岐数の総和 B と，ネットワークの基点となる名詞（“ cow ”；問題文の「選択肢」）からネットワークに含まれる問題文中に現れた名詞（“ farm ”, “ animal ”）までの距離（ノード間ステップ数）の総和 D に注目した．この 2 つを m_b および m_d の重み付けによって，評価値 E を式(1)に定義する．読解対象文をコンテキストとする，つまり問題文に含まれる名詞を用いたネットワーク形成を問題文の「選択肢」毎に生成し，各ネットワークを評価値 E で比較し，E が高ければ高いほど，対象とする問題文と当該「選択肢」の関係性が高いと考える．よって，評価値 E によって順位付けされた「選択肢」の 1 位が英文テストの正答と一致することかどうかを検証する．

$$E = m_b B + m_d D \tag{1}$$

(1) ネットワーク内のノード分岐数の総和 B

式(1)の B について定義する．あるノードの分岐数とは，そのノードから他のノードに連結されている数である．図 4 において矢印（矢じりを太線にした）はネットワーク生成時に辿った方向に向いている．ここでは，その向きは問わない．したがって，ネットワークのノード数を N とすると総和 B は以下のように定義できる．

$$B = \sum_{i=1}^{N} (b_i - 1)(i = 1,2,\cdots,N) \tag{2}$$

ここで，b_i は i 番目ノードにおける分岐数である．分岐が多い程，基点名詞（選択肢）から英英辞典を辿って現れた名詞と，問題文内名詞は整合性（類義語特性）が高かったということになる．

(2) 選択肢（基点）から各問題文中名詞への距離の総和 D

式(1)の D を定義する．図 4 のように選択肢 “ cow ”から関連名詞“ animal ”, “ domestic ”などまでのネットワーク内距離が近い程，その選択肢が問題文の意味に関連が高く，正答である可能性が高いと考える．そこで，*DEPTH* に中心名詞–関連名詞，関連名詞–類義語間の冗長距離（ここでは $\eta = 3$）を加えることで最長距離 が得られる．そこで，この関連度は正の数の総和となり，以下が与えられる．

$$D = \sum_{j \in M} (DEPTH + \eta - d_j) \tag{3}$$

ここで，M はネットワークに含まれる類義語の見出し語の集合で，d_j は j 番目見出し語ノードまでのネットワーク距離である．

3. 実験

英文-単語意味整合テストを 101 問用意し（図 3 に例），評価値 E の値で順位付けを行った．E が同じ場合は同ランクとする．

Vol.2014-ICS-176 No.4
2014/7/22

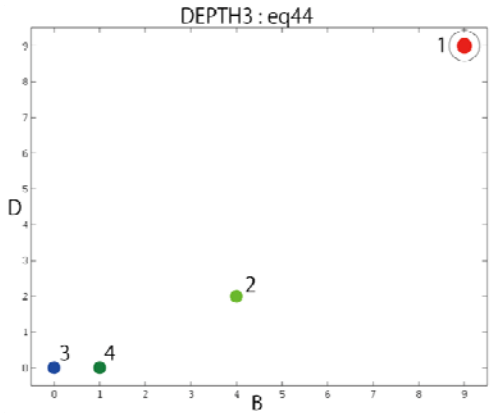


図 5 英語問題 44 (*DEPTH*3)

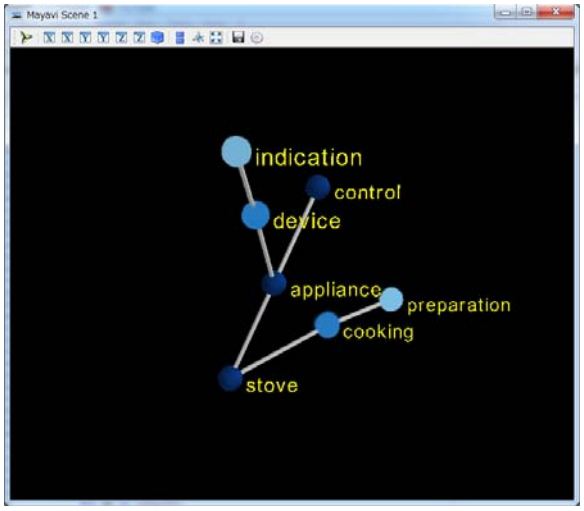


図 6 英語問題 44 選択肢 stove

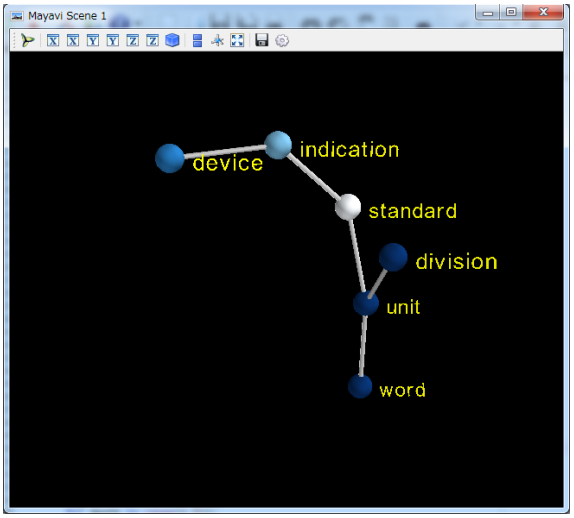


図 7 類英語問題 44 選択肢 word

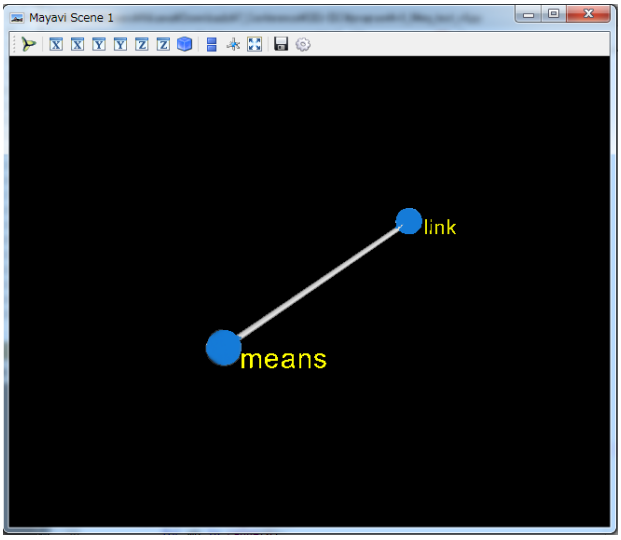


図 8 英語問題 44 選択肢 link

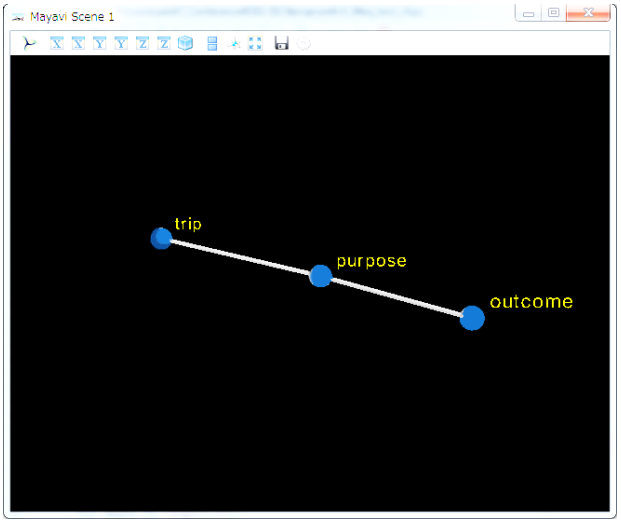


図 9 英語問題 44 選択肢 trip

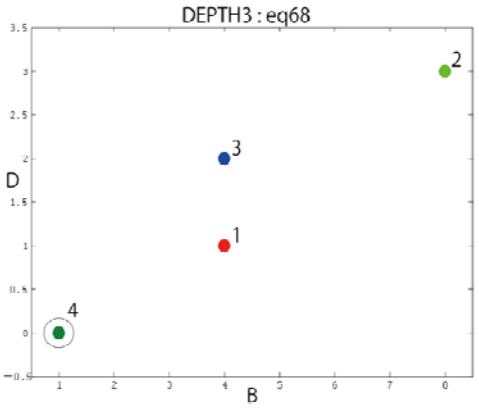


図 10 英語問題 68 (*DEPTH*3)

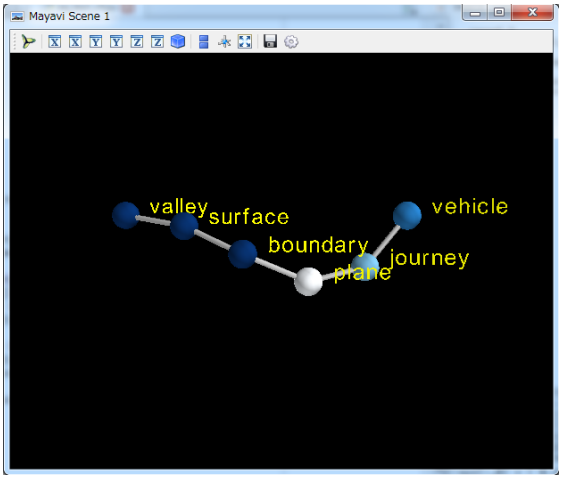


図 11　英語問題 68 選択肢 valley

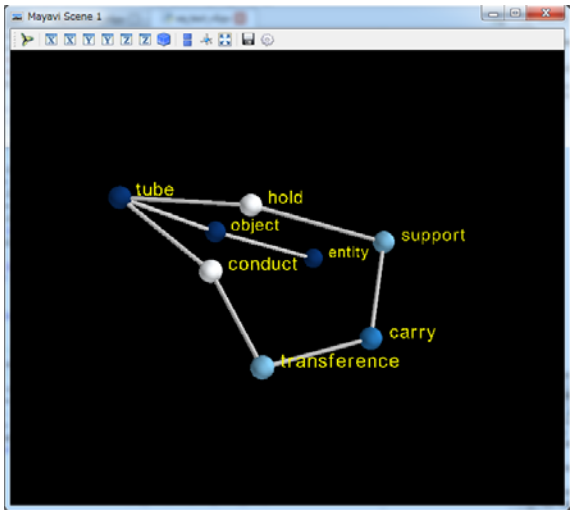


図 12　英語問題 68 選択肢 tube

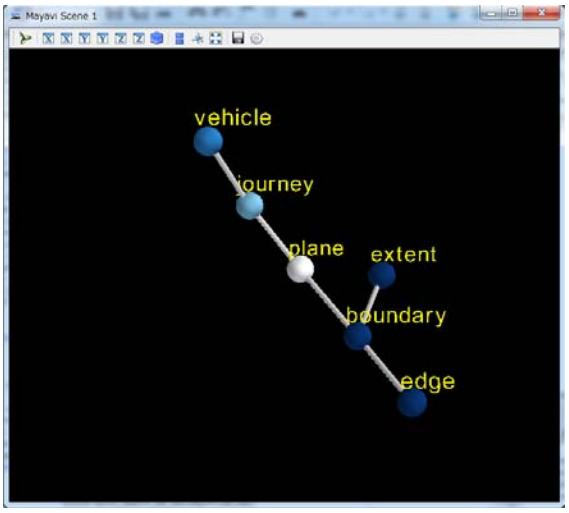
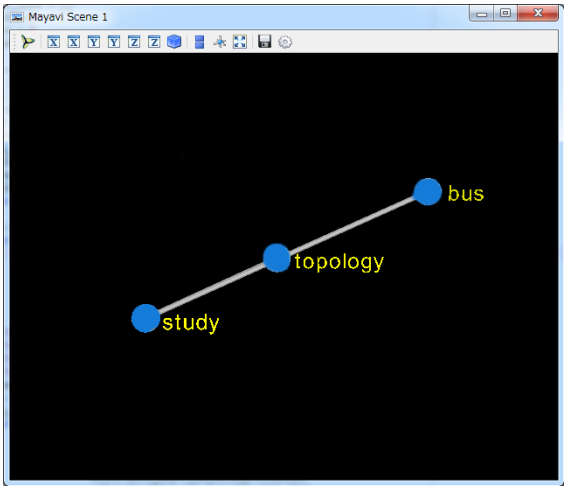


図 13　英語問題 68 選択肢 edge



英語問題 68 選択肢 bus

可視化においては，各ネットワーク形状については 3 次元描画を行い，問題選択肢の評価値の分布については B（式 2）および D（式 3）の二つを軸とする散布図によって可視化した．尚，正解の選択肢については，丸で囲んだ．つまり◯で囲まれた選択肢が右上にある場合が，本提案法が有効に働く場合の問題構造であり，逆に左下にある場合は，パラメータ m_b および m_d を変化させても正解に至ることは困難である．

4. まとめ

可視化によって，ここで取り扱っている英語テストの構文には，本提案法に対して有効なタイプと，逆に無効となるタイプが二分することがわかった．今後，分類による最適化も含めて検討して行きたい．

謝辞　本研究は一部 JSPS 科研費 26240032 の助成を受けた．

参考文献

1) 北研二: 言語と計算−確率的言語モデル，東京大学出版会(1999)
2) 我妻広明，相川大輔: 言語指示−ロボット遂行課題で実世界照会を用いる構文解析ツリーにおける伸縮リンクの提案，日本神経回路学会全国大会(JNNS 2012)予稿集, P3-23.
3) 相川大輔，我妻広明: 英文解析のための中心名詞をノードとするネットワーク生成手法の提案，第 3 回日本知能情報ファジィ学会九州支部学生部会研究発表会予稿集，P11-12
4) 相川大輔，我妻広明: ロボット言語指示課題に向けた文法構造を反映した構文解析ツリーの検討，ネットワークが創発する知能研究会(JWEIN 2013)論文集，P78-83
5) 相川大輔，我妻広明: 脳内言語処理から発想を得た人-ロボット指示理解:読解対象文をコンテキストとするフィルタリング・ネットワーク形成の提案，人工知能学会全国大会(JSAI 2014)論文集，1F2-5