

ネットワークの活性値に基づく深層格解析手法 A Method for Deep Case Analysis Based on Activity in a Network

渋谷 英潔^{†*}
Hideyuki Shibuki

荒木 健治[‡]
Kenji Araki

桃内 佳雄^{††*}
Yoshio Momouchi

栃内 香次^{‡‡}
Koji Tochinai

1. まえがき

我々は、要約や対話システムなどへの応用を目的として、文章全体を一つのネットワークで表現する文脈解析器の開発を試みている。文脈解析に関する従来研究として、文献[1]などが挙げられる。しかしながら、従来研究では、表層的な文の構造や単語の距離などを重視するため、本来意味的に同一である文が、表現上の違いにより異なる結果となりやすいと考えられる。我々は、文献[2]などにおいて、単語の概念ごとに特定の深層格に解釈される傾向（深層格選好）が存在するという考え方に基づく深層格推測手法を既に提案している。この考え方に基づくならば、文章中の単語に着目することで、表現上の違いによる影響を抑えた解析が可能になると考えられる。そのために、本稿では、文章中の単語を入力とし、ネットワークの活性値に基づいて現在の主題を判断することで、文の単位にとらわれずに深層格を解析する手法を提案する。

本手法のネットワークは、名詞と動詞をノードとし、各ノードは深層格でラベル付けされたリンクにより関連付けられる。ネットワークの活性値に基づいて格解析を行う手法として竹内ら[3]などが挙げられるが、竹内らの手法では一文単位での処理を行うのに対し、本手法は文脈を考慮した処理を行っている。また、解析の時点までに入力された全ての単語をネットワーク上のノードとして等しく表現するため、以前に入力された文の単語に対しても、現在入力された文の単語と同一の機構で処理することができ、照応に関する処理を特別に行う必要がない。本稿では、手法の提案と、小規模な実験を行った結果について報告する。

2. 基本的な考え方

本手法では、現在の話題が何であるかを表すために、ネットワークによる活性値を用いる。例えば、「明日、14時から会議があります。議題は来月の学園祭に関してです。場所は6階の62番教室です。」という文章を人間が読んだ場合を仮定する。このとき、3文目の「場所」が「会議の場所」を指していると理解されるのは、「会議」という単語が「2つ前の文に含まれているから」というような明確な理由によるものではなく、もっと漠然とした「現在の話題が会議に関することだから」というような理由によるものと思われる。仮に、2文目が「先週頼ま

れた機材が届いています。」という文であった場合には、「場所」が「機材の場所」を指す可能性が高くなる。これは、現在の話題が、2文目で「機材」に関する話題に移ったからだと考えられる。

本手法では、このような各文を読んだ時点での話題の移り変わりを活性値の変化により表現し、「話題は常に移り変わるものであり、現在の話題を継続させるためには、現在の話題に関連した事柄を常に述べる必要がある」という捉え方をする。すなわち、最初の例で「会議の場所」と理解できたのは、2文目に「議題は～」という「会議」に関連した文を続けたことにより、話題を「会議」から変化させなかったからだと考え、一方、2文目が「機材」に関する文となった場合には、話題を「会議」に留めようとしなかったために、次の話題である「機材」に移ることとなったと考える。以上の考え方を実装するために、本手法では、全ての活性値を常に減少させることで話題を変化させようとする力と、入力された単語のノード及びそれに関連するノードの活性値を増加させることで話題を留めようとする力の、二つの力を拮抗させることで現在の話題を求めるというアプローチをとる。ここでの関連するノードとはネットワーク上でリンクが張られているノードのことである。

3. 処理

本手法では、日本語文章中の名詞と動詞を入力とし、逐次的に解析を行う。名詞を入力する場合には、助詞も共に入力する。ネットワーク中のノードは、名詞に関するノード（名詞ノード）と動詞に関するノード（動詞ノード）の2種類である。動詞ノードからは、下位範疇化する名詞ノードに対して agent や object などの深層格のリンクが張られている。それぞれの名詞ノードには活性値が存在し、活性値の高いノードほど話題性が高いことを意味する。

以下の手順で解析を行う。まず、入力単語の表記を基に、対応するノードを検索する。名詞の場合には、同一表記の名詞ノードがネットワーク上に既に存在しているか検索し、存在する場合には、そのノードと同一であるとし、存在しない場合には、新たにノードを作成する。検索された名詞ノードに対し、共に入力された助詞の情報を付加する。動詞の場合には、本稿では動詞への照応を対象としていないため、常に新たなノードを作成することとした。一般に既知の単語への言及よりも新規な単語への言及の方が話題性が高いと考えられるため、新たに作成されたノードの活性値を高め設定することとした。本稿では、便宜的に活性値増加量が2倍となるようにした。

次に、入力単語が動詞であった場合、その動詞が下位範疇化する名詞ノードを探索する。名詞ノードの探索は、動詞ノードの深層格選好と選択制限、名詞ノードの深層

[†]北海学園大学大学院経営学研究科, Graduate School of Business Administration, Hokkai-Gakuen University

[‡]北海道大学大学院情報科学研究科, Graduate School of Information Science and Technology, Hokkaido University

^{††}北海学園大学工学部, Faculty of Engineering, Hokkai-Gakuen University

^{‡‡}北海学園大学経営学部, Faculty of Business Administration, Hokkai-Gakuen University

*北海学園大学ハイテク・リサーチ・センター, High-Tech Research Center, Hokkai-Gakuen University

表 1: 実験結果

全体			
評価対象数	130	被覆率	74.6%
出力数	97	精度	90.7%
正解出力数	88	正解率	67.7%
ゼロ代名詞			
評価対象数	49	正解率	63.3%
正解出力数	31		

格選好と意味素性、助詞の深層格推測規則、名詞ノードの活性値に基づいて行われる。深層格選好及び深層格推測規則に関しては文献 [2]、選択制限と意味素性に関しては文献 [1] に述べられているため、紙面の都合により本稿では省略する。深層格選好と深層格選好規則は、本来コーパスから学習される知識であるが、本稿では、活性値による話題変化の有効性を検討することと、最初の段階として小規模な実験システムを作成するという理由から、人手で作成した知識を与えた。また、本来の確率値ではなく、ある深層格と解釈される可能性がある場合には 1、ない場合には 0 とする二値で表現した¹。意味素性制約を満たす名詞ノードに対し、以下の式に基づいて尤度 $pl(n, v, g, d)$ を計算し、閾値以上の尤度となる名詞ノードを解析結果として出力する²。

$$val(n, v, g, d) = dp_N(n, d) \times \sum_{i \in D} dp_V(v, d, i) \times gr(g, d) \times act(n)$$

$$pl(n, v, g, d) = \frac{val(n, v, g, d)}{\sum_{i \in D} val(n, v, g, i)^2} \quad (1)$$

n, v, g, d, D は名詞、動詞、助詞、深層格、深層格の集合をそれぞれ表す。 $dp_N(n, d)$ は名詞の深層格選好、 $dp_V(v, d, i)$ は動詞の深層格選好、 $gr(g, d)$ は助詞の深層格推測規則、 $act(n)$ は名詞ノードの活性値である。式 (1) は、文献 [2] の尤度計算式に活性値を掛け合わせたものである。

入力された単語に対応するノードを中心とし、以下の手順でノードの活性値を増加する³。名詞の場合には、対応するノードの活性値を 4 増加させ、そのノードを下位範疇化している動詞ノードが下位範疇化する名詞ノードの活性値を 1 増加させる。動詞の場合には、下位範疇化している名詞ノードの活性値を 1 増加させる。

動詞が入力された場合、節の区切りと考えられるため、全ての名詞ノードに付加された助詞の情報を消去する。また、話題を変化させるために、活性値の減少を行う。突出した活性値を平均化するため、現在の活性値の平方根となるように活性値を減少させる。

4. 実験

話題の移り変わりを活性値として捉えることの有効性を検討するため、小規模なシステムを作成し実験を行う

¹また、同様の理由により、助詞は、「に」、「と」、「が」、「は」、「へ」、「で」、「を」、「から」、「も」、「の」の 10 種類を設定し、意味素性は、「時間 (むかし)」、「場所 (山)」、「人間 (おじいさん)」、「体部分 (手)」、「物 (桃)」、「液体 (水)」、「容器 (たらい)」、「行為 (洗濯)」、「名称 (桃太郎)」の 9 種類を設定した。これらの知識に関しては、今後、大規模な実験を行う際に見直す予定である。

²閾値は文献 [2] と同じく 0.7 とした。

³以下の増加量は便宜的な値であり、対応ノードとの距離の 2 乗に反比例するように設定した。

た。実験で対象とする文章として、青空文庫 [4] の楠山正雄「桃太郎」を用い、桃太郎が生まれるまでのシーンに含まれる、延べ 156 単語 (名詞 90 語、動詞 66 語) を入力とした。また、異なり語数は名詞 37 語、動詞 59 語であった。調査対象とした深層格は、agent, object, source, goal, purpose, place, time の 7 種類であり、入力中に 130 の解析箇所を設定した。130 の解析箇所のうち、動詞と同一文中に含まれる名詞が正解となるのが 81 箇所、含まれない名詞 (ゼロ代名詞) が 49 箇所であった。評価基準を被覆率、精度、正解率とし、それぞれ以下の式により計算した。

$$\text{被覆率} = \frac{\text{出力数}}{\text{評価対象数}}, \quad \text{精度} = \frac{\text{正解出力数}}{\text{出力数}}$$

$$\text{正解率} = \frac{\text{正解出力数}}{\text{評価対象数}} \quad (2)$$

正解かどうかの判断は第一著者の判断によった。

解析結果を表 1 に示す。被覆率 74.6%、精度 90.7% となり良好な結果が得られた。また、ゼロ代名詞である 49 箇所における正解数は 31 (正解率 63.3%) であった。今回の実験は予備実験的なものであり従来手法との比較を行える段階ではないが、省略解析を扱った文献 [1] では適合率 53.9%、再現率 46.3% という結果であり、表 1 に示す結果は本手法の今後を期待できる数値であると考えられる。

5. おわりに

本稿では、現在の話題をネットワーク上の活性値で表現し、文の単位にとらわれずに深層格を解析する手法を提案した。本手法は、全ての活性値を常に減少させることで話題を変化させようとする力と、入力された単語のノード及びそれに関連するノードの活性値を増加させることで現在の話題を留めようとする力を拮抗させるという考え方に基づいている。また、入力された全ての単語をネットワーク上のノードとして表現することにより、照応処理を特別に行うことなく解析を行う。小規模な実験を行った結果、被覆率 74.6%、精度 90.7% となり、良好な結果が得られた。今後、コーパスからの学習を行い、大規模な実験を行っていきたいと考えている。

謝辞

本研究の一部は、北海学園大学ハイテク・リサーチ・センター研究費、及び、北海学園大学学術研究助成金による補助のもとに行なわれた。

参考文献

- [1] 河原大輔, 黒橋禎夫: 自動構築した格フレーム辞書と先行詞の位置選好順序を用いた省略解析, 自然言語処理, Vol.11, No.3, pp.3-19 (2004).
- [2] 渋木英潔, 荒木健治, 桃内佳雄, 柄内香次: 深層格選好に基づく深層格推測手法の英文への適用, 言語処理学会第 11 回年次大会発表論文集, pp.1056-1059 (2005).
- [3] 竹内俊行, 荒木健治, 柄内香次: Marker Passing 形式のパーザにおける力学的制約を用いた意味つり上げモデルによる格構造解析手法, 電子情報通信学会論文誌 D-II, Vol.J83-D-II, No.12, pp.2755-2766 (2000).
- [4] 青空文庫: <http://www.aozora.gr.jp/>.