非命題的意味解析のための日本語文末表現意味体系

本田　聖晃† 　田辺　利文† 　吉村　賢治† 　首藤　公昭†
†福岡大学工学部

１　はじめに

近年,Web の普及による文書情報の爆発的な増加に伴い,より精度の高い自然言語処理の必要性が認識されている.特に,意味を考慮した処理システムの需要が増加していることは Semantic Web 研究等の発展を見ても明らかである.筆者らは,意味を考慮した言語データ処理においては,複単語表現（Multiword Expression: MWE）を組み込んだシステム構築が不可欠であると考える.MWE に関しては,2002 年の論文「Multiword Expressions: A Pain in the Neck for NLP」（Sag et al. 2002）を皮切りに 2007 年夏までに ACL において MWE のワークショップが計 4 回実施されるなど,重要性が世界的に認識されつつある.

これまで筆者らは,熟語性,語彙的一体性,確率的束縛性のうち少なくとも１つの性質を持つ単語列を MWE として収集してきた.熟語性とは,構成している単語の通常の意味から全体の意味を構成するのが難しいことを意味する.また,語彙的一体性とは分離しにくさを,確率的束縛性とは要素単語相互の確率的な縛りの強さを意味する.（Sag et al. 2002）では,WordNet1.7 での見出しの約 41%が MWE であること,また（田辺ら 2006）では,日本語の述部における文末表現を構成する助動詞,終助詞相当の MWE の出現比率が約 42%であることが報告されており,MWE の適切な処理が自然言語処理の質を向上する上で必要不可欠であることを示唆している.発話者の主観を表す日本語の述部の文末表現には,必要性を表す「なければならない」,欲求を表す「たい」,様態を表す「ようだ」のようにバラエティに富んでおり,かつ「なければならない/ようだっ/た」のように複数個連接して用いられることも多く,発話者の主観に基づく処理を行うためには機能語性 MWE を意識した文末表現の取り扱いが必要となる.

本論文では,第２章で非命題的意味構造についての紹介を行い,第３章で機能語性 MWE を組み込んだ体系としての日本語文末表現意味体系について述べる.第４章で非命題的意味構造間の関係に基づく日本語文末表現意味体系の拡張について述べた後,第５章でまとめ,今後の課題について考察する.

２　非命題的意味構造

一般的に文の意味は,命題的意味と非命題的意味からなっていると考えられ,日本語においては,非命題的意味を表す部分は述語に後接した,時制,判断,否定,話し手の態度など,広義の様相情報を与える助動詞,終助詞およびそれらに相当する機能語性 MWEにより構成される場合が多い.筆者らはそれらの表現を助述表現と呼び,これまで約 1,450 個の機能語性MWEと約 50 個の助動詞,終助詞を収録した辞書を作成している[1].機能語性MWEとしては,例えば,「て」と「いる」をまとめた「ている」,「かも」と「しれ」と「ない」をまとめた「かもしれない」などを収録している.助述表現は,発話者の主観に基づく表現であるとみなすことができる.筆者らは網羅性の高さを目標の一つと考えており,例えば,（森田ら 1989)の収録表現のうち該当表現はすべてカバーしている.機能語性MWEを適切に設定することで,日本語文の構造の大枠は次の生成規則で表すことができる.

(1)　$S_0 \rightarrow BP^* \cdot PRED$

(2)　$S_i \rightarrow S_{i-1} \cdot e_i \quad (1 \le i \le n)$

S_0，BP，‘*’，$PRED$，e_i はそれぞれ骨格文,文節,閉包演算子,述語,助述表現を表す.(2)は,S_{i-1} と e_i とで文 S_i が構成されること,及び,構造が左分岐であることを表している.例えば,日本語文「彼は動き始めていないかもしれない」では,日本語文の構造の概形および対応する非命題的意味構造は図 1 のようになる[2].

助述表現に対応する意味を意味関数として捉え,並びを逆順にすることで,非命題的意味構造を作ることができる.一般的な非命題的意味構造は,次のような入れ子型表現で表すことができる.

(3)　$M_m[M_{m-1} \cdots [M_2[M_1[S_0]]] \cdots]$

但し,$M_i(1 \le i \le m)$ は意味関数である.(1),(2)に示される構文構造と,(3)に示される意味構造とには一種の同型性があると言える.助述表現が文の述部にいくつも並んだ複雑な文末表現の場合でも,意味関数との対応をとることにより非命題的意味構造を求めることが可能である.このように(3)は,構造のシンプルさと同時に対応可能な表現の多様さから工学的に重要な性質を示していると考えられ,言語依存性も無いとされるため[3],言い換えや機械翻訳を行う際の中間言語として有効であると考えられる.

---

[1] これらの数値は，漢字-かな などの表現のゆれをまとめて１見出しとした値である.

[2] 図１の記号‘･’は通常の単語境界を表し，‘/’はMWEによる単語境界を表す.また,動詞に後接する「はじめる」は動詞ととらえる考え方もあるが,アスペクト情報を含むと考えられるため本論文では助述表現としている.

[3] (Cinque 1999; Cinque 2006)は，非命題的意味構造における意味関数の生起順序には，言語に依存しない規則性が存在するのではと報告している.

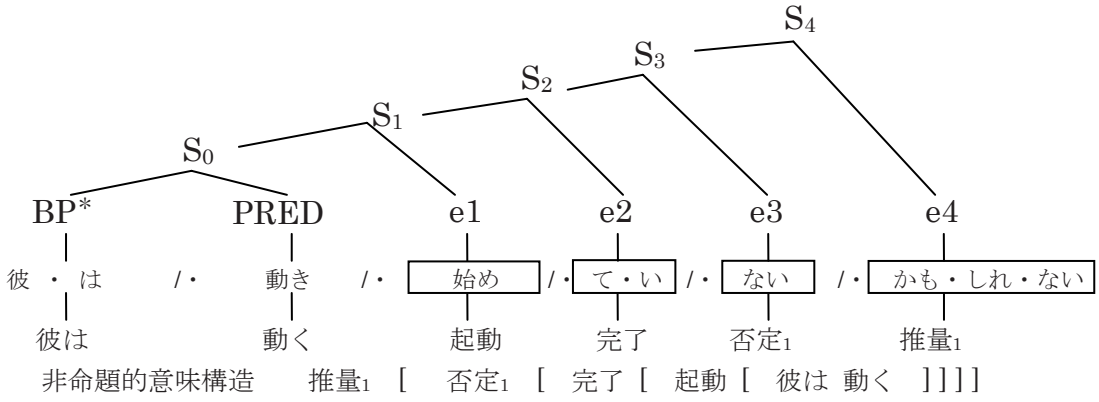

図１　日本語文の構造の概形及び非命題的意味構造

３　日本語文末表現意味体系

３.１　意味関数の体系

助述表現に対応した意味関数は,ツリー状に体系化しており,意味関数間の類似度が定義できる.意味関数は

・最上位ノードを４種類設定

・より細分化された意味を下位ノードとした階層化[4]

・葉ノードを 139 種類設定[5]

して詳細に体系化した.詳細な体系化により,発話者の主観情報の詳細な取り扱いが可能になる.各ノード（意味）には助述表現の集合が対応づけられる.そのため,意味関数間で定義される類似度から,助述表現間の類似度を与えることができる.

３.２　助述表現間の意味的類似度

助述表現,例えば,「なければならない」と「必要がある」の間の類似度の算出について考える.ここで,意味の集合を S_M ,助述表現の集合を S_E とおくと,関数 f を,

(4)　$f : S_M \rightarrow 2^{S_E}$

と形式的に表すことができる.(4)は意味の集合から助述表現の集合のべき集合へ写像する関数となる.ここで, S_M 上の is-a 関係に基づいた意味 $m_1, m_2 \in S_M$ の類似度 $sim(m_1, m_2)$ を次のように定義する.

(5)　$$sim(m_1, m_2) = \frac{2 \cdot d_c}{d_i + d_j}$$

ただし, d_c は, m_1, m_2 の共通上位ノードのルートからの距離, d_i, d_j はそれぞれ, m_1, m_2 のルートからの距離である.また, $sim(m_1, m_2)$ が α であるとき,

(6)　$f(m_1) = \{x_1, x_2, \cdots, x_m\}$

(7)　$f(m_2) = \{y_1, y_2, \cdots, y_n\}$

である x_i と y_j の間は, S_E 上で類似度 α の関係にあると定義する.ここで, $\alpha = 1$ の場合, x_i と y_j の間は同義関係にあると定義する.これらの定義は,任意の助述表現間の意味的類似度を算出することができること,および,同義関係にある助述表現間は基本的に言い換え可能であることを意味している.

３.３　助述表現列間の意味的類似度

実際の日本語文末では,「なければならない/ようだっ/た」などのように,助述表現が連続して助述表現列として出現する場合も多い.前節では助述表現間の関係として類似度を定義したが,助述表現列間でも類似度を定義できれば望ましい.そこで本節では,助述表現間から助述表現列間の類似度の拡張を考える.ここで, S_M の閉包を S_M^* （意味列の集合）, S_E の閉包を S_E^* （助述表現列の集合）とおく.ここで,関数 g を,

(8)　$g : S_M^* \rightarrow 2^{S_E^*}$

と形式的に表すことができる.(8)は意味列の集合から助述表現列の集合のべき集合へ写像する関数となる.（但し, $S_M \subset S_M^*$, $S_E \subset S_E^*$ である.）(8)は, $m \in S_M, \alpha \in S_M^*$ において

(9)　$g(m\alpha) = f(m) \cdot g(\alpha)$

(10)　$g(\lambda) = \lambda$

と定義できる.但し, λ は空文字列,'・'は結合演算を表す.ここで,意味列 $M_1, M_2 \in S_M^*$ の類似度 $sim(M_1, M_2)$ を定義する.意味列 M_1, M_2 間の類似度は,DP（Dynamic Programming）を用いた意味列間の最短距離計算（DP マッチング）に基づき算出することができる.ここで, $d(i, j)$ は意味列

---

[4] 例えば意味「必要性」を,「必要性$_1$」「必要性$_2$」・・・「必要性$_8$」と細分化し，細分化されたそれぞれのノードを「必要性」の下位ノードとして設定している.

[5] 図１における意味関数は全て葉ノードを用いている.

begin

$d(0,0) := 0;$

for $i := 1$ to m do

$d(i,0) := d(i\text{-}1,0) + w(M_1,i);$

for $j := 1$ to n do

$d(0,j) := d(0,j\text{-}1) + w(M_2,j);$

for $i := 1$ to m do

for $j := 1$ to n do

$d(i,j) := \min\{d(i\text{-}1,j\text{-}1) + 2 * (1\text{-}sim(a_i,b_j)), d(i\text{-}1,j) + w(M_1,i), d(i,j\text{-}1) + w(M_2,j)\};$

$$sim(M_1,M_2) := 1\text{-}d(m,n) \Big/ \left(\sum_{k=1}^{m} w(M_1,k) + \sum_{l=1}^{n} w(M_2,l) \right);$$

end

$w(M_1,i)$:意味列 M_1 における意味 a_i の重要度, $0 \le w(A,i) \le 1$

$w(M_2,j)$:意味列 M_2 における意味 b_j の重要度, $0 \le w(B,j) \le 1$

図２　意味列間類似度判定処理

$M_1 = a_1, a_2, \cdots, a_i$ と意味列 $M_2 = b_1, b_2, \cdots, b_j$ の意味列間距離を表すものとする.このプロセスは(ⅰ)意味列 $a_1, a_2, \cdots, a_{i-1}$ と $b_1, b_2, \cdots, b_{j-1}$ の最短距離 $d(i-1, j-1)$ に a_i と b_j の置き換えによる距離の増加分を加算したもの,(ⅱ)意味列 $a_1, a_2, \cdots, a_{i-1}$ と $b_1, b_2, \cdots, b_j$ の最短距離 $d(i-1, j)$ に意味の重要度 $w(M_1,i)$ を加算したもの,および(ⅲ)意味列 $a_1, a_2, \cdots, a_i$ と $b_1, b_2, \cdots, b_{j-1}$ の最短距離 $d(i, j-1)$ に意味の重要度 $w(M_2,j)$ を加算したもの,の３個の数値のうち最小値を意味列 $a_1, a_2, \cdots, a_i$ と $b_1, b_2, \cdots, b_j$ の最短距離 $d(i,j)$ として求める部分である.重要度 $w(M_1,i)$ は,意味列間距離を算出する場合の寄与の度合であり,例えば,全ての意味に対して均等にする,特定の意味を重視する,また,最後尾の助述表現の意味のみを考慮するなど,アプリケーションに応じた柔軟な類似度算出が可能となる.i, j の値を,それぞれ1から m ,1から n まで１ずつ増やしながら処理を繰り返し実行することによって,最終的に意味列 $M_1 = a_1, a_2, \cdots, a_m$ と $M_2 = b_1, b_2, \cdots, b_n$ の最短距離として意味列間距離 $d(m,n)$ が求められる.最短距離計算の詳細は(安武ら 1999)を参照されたい.意味列間距離 $d(m,n)$ が求まると,意味列間類似度 $sim(M_1,M_2)$ が求まる.

同様に, $sim(M_1,M_2)$ が α であるとき,

(8)　$g(M_1) = \{X_1, X_2, \cdots, X_m\}$

(9)　$g(M_2) = \{Y_1, Y_2, \cdots, Y_n\}$

である X_i と Y_j の間は, S_E^* 上で類似度 α の関係にあると定義する.同様に $\alpha = 1$ の場合, X_i と Y_j の間は同義関係にあると定義する.これらの定義は前節と同様,任意の助述表現列間の意味的類似度を算出することができること,及び,同義関係にある助述表現列間は基本的に言い換えが可能であることを意味している. S_E^* による体系は意味的類似度を定義できる体系であり,この点でシソーラスなどの意味体系と同一であるとみなすことができる. S_E^* の元は助述表現列であり,助述表現列は日本語文末に存在しうることから,本論文では S_E^* による体系を日本語文末表現意味体系と呼ぶことにする.日本語文末表現意味体系は1500種の助述表現が複数個連接した表現による意味体系であり, n 個の助述表現が連接した場合,最大 1500^n 種の表現を網羅する大規模な意味体系であると考えることができる[6].

４　日本語文末表現意味体系の拡張

前章までで任意の助述表現列間に,意味的類似度を定義できることを示した.しかし,例えば「行かなければならないことはない」と「行かなくてもよい」のような助述表現列間の意味的類似度は文脈等に応じて高く見積もることが出来れば望ましい.そこで第４章では,非命題的意味構造間に定義できる関係と,日本語文末表現意味体系の拡張について考察する.

４.１　非命題的意味構造間の関係

４.１.１　類似関係

非命題的意味構造間の関係の中でも特に類似関係として,類似性規則が提案されている（田辺ら

[6] 助述表現間の連接には制約があり,存在しうる助述表現列の種類の概算は現在行っている.

2001）.類似性規則は,論理的規則および語用論的規則に大別される.論理的規則には

(13)　否定$_1$[否定$_1$[S]] ≒[S]
「行か/ない/こと・は・ない」 ≒「行く」

(14)　否定$_1$[必要性$_1$[S]] ≒ 許容$_2$[否定$_1$[S]]
「食べ/な・けれ・ば・ならない/こと・は・ない」
≒「食べ/なく/て・も・よい」

など,語用論的規則には

(15)　願望[受動態$_3$[S]] ≒ 依頼$_1$[S]
「見て/いただき/たい」
≒「見て/ください」

(16)　疑問$_1$[否定$_1$[可能性$_{21}$[S]]] ≒ 依頼$_2$[S]
「行く/こと・が・出来/ない/か」
≒「行って/ください」

などがある.

４.１.２　類似以外の関係

非命題的意味構造間の関係として,類似以外の関係を挙げることもできる.次のような例を考える.

(17)　過去[否定[必要性$_3$[XがYする]]]
→過去[X が Y する]
「X が Y する/べき・で・なかっ/た」
→「X が Y し/た」

(18)　願望[X が Y する]
→否定[完了[X が Y する]]
「X が Y し/たい」
→「X が Y し/てい/ない」

(17)は,「XがYするべきでなかった」ならば「XがYした」,(18)は,「XがYしたい」ならば「XがYしていない」はず,を表し,話し手の発話を受けた聞き手が推論しうる関係であると考えられる.

４.２　日本語文末表現意味体系の拡張

非命題的意味構造間の関係を規則として適用することで,日本語文末表現意味体系の拡張を考える.非命題的意味構造間の関係は S_M^* から S_M^* への写像

(19)　$h : S_M^* \rightarrow S_M^*$

として形式的に表すことができる.ここで,(5)より,合成関数 $g \circ h$ は

(20)　$g \circ h : S_M^* \rightarrow 2^{S_E^*}$

となる.そのため, $M_1 \in S_M^*$ に対して規則を適用した $h(M_1)$ と, $M_2 \in S_M^*$ の間での意味的類似度 $sim(h(M_1), M_2)$ を定義できる.ここで, $sim(h(M_1), M_2) > sim(M_1, M_2)$ であり, $sim(h(M_1), M_2)$ が α であるとき,

(21)　$g(M_1) = \{X_1, X_2, \cdots, X_m\}$

(22)　$g(M_2) = \{Y_1, Y_2, \cdots, Y_n\}$

である X_i と Y_j の間は, S_E^* 上で類似度 α の関係にあると定義できる[7].つまり,(19)を規則として適用することで助述表現列間の意味的類似度が高くなった場合,文脈などに応じて高いほうを優先することができる.このように,非命題的意味構造間の関係を規則として適用することで,日本語文末表現意味体系の拡張ができると考えている.

５　おわりに

自然語文の非命題的意味は,対話理解,文脈モデルや話者の態度の推定などの自然言語処理で重要な役割を果たす.本論文では,日本語文において発話者の主観を表すと考えられる助述表現列について整理した日本語文末表現意味体系について報告した.日本語文末表現意味体系は大規模であり,任意の助述表現列間に対し意味的類似度を算出できる.また,日本語文末表現意味体系と,自立語シソーラスをうまく組み合わせることにより,例えば「教授が述べるに違いないだろう」と「先生が話すかもしれないらしい」の間の意味的類似度の算出も可能になり,文書分類や用例に基づく機械翻訳などの応用技術への貢献が期待できる.また,本論文では,助述表現列による非命題的意味について述べたが,「たぶん」「きっと」などのような副詞も非命題的意味を持ちうる.そのため,副詞をどのように非命題的意味構造に組み込むかも検討する必要がある.今後の課題として,非命題的意味構造間の関係の収集や,実際のテキストデータへの適用などが挙げられる.

参考文献

[1] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. “Multiword Expressions: A Pain in the Neck for NLP.” The Proc. of the 3rd CICLING, pp. 1-15. 2002.

[2] 田辺利文, 本田聖晃, 高橋雅仁, 小山泰男, 吉村賢治, 首藤公昭. “日本語文末表現の取り扱いについて.” FIT2006,pp. 241-244. 2006.

[3] 森田良行, 松木正恵. “日本語表現文型用例中心・複合辞の意味と用法.” アルク. 1989.

[4] Guglielmo Cinque. “Adverbs and Functional Heads.” OXFORD UNIVERSITY PRESS. 1999.

[5] Guglielmo Cinque. “Restructuring and Functional Heads.” OXFORD UNIVERSITY PRESS. 2006.

[6] 安武満佐子, 小山泰男, 吉村賢治, 首藤公昭. “関係表現,助述表現の類似度を考慮した言語表現間の意味的類似度判定.” 福岡大学工学集報 第 63 号,pp.171-177. 9 月. 1999.

[7] 田辺利文, 吉村賢治, 首藤公昭. “日本語モダリティ表現とその言い換え.” 言語処理学会第 7 回年次大会ワークショップ論文集,pp. 47-50. 2001.

[7] ここでの h としては,類似関係のみ（類似性規則）を対象としており,類似以外の関係については今後の課題としている.