訳語候補を手がかりとする FrameNet を用いた日本語文への意味役割付与

橋本祐樹† 鈴木基之†† 任福継††
† 徳島大学大学院 先端科学教育部
†† 徳島大学大学院 ソシオサイエンス研究部
{hashimoto, suzuki_m, ren}@.is.tokushima-u.ac.jp

1 はじめに

自然言語処理の研究分野において，意味解析の重要性が叫ばれて久しいが，今日，日本語の意味解析に関しては，利用できる資源が十分に整備されてるとは言えない．そのため，日本語の意味解析手法の研究は，英語のそれと比べて困難であると言える．これは日本語のみならず，言語資源に乏しい英語以外の言語に関しても同様の問題が生じている．

一方，英語に関する言語資源については開発の歴史が古く，実用に足る規模で実装されているものも存在する．代表的な言語資源の一つとして，フレームネット [1] が挙げられる．フレームネットは，フレーム意味論に基づく意味解析のための言語資源で，ある語や句の意味を記述するためにその背景場面となる意味フレーム（枠組み）が定義され，その各々に対して，属する語彙やフレーム要素（意味タグ）を定義している．

現在，英語版フレームネットには英語に関して記述された意味フレームが約 800，フレーム要素による注釈付きコーパスが約 15,000 文存在している．これらのコーパスは主に意味解析器等の学習用データとして使われる．日本語版フレームネット [2] も存在するが，まだ開発の歴史が浅く実用に耐える規模を備えていない．また，それらの拡張は基本的に人手で行われるためコストが高く，短期間に実用的な資源を入手するのは難しい．

そこで本研究では，英語版フレームネットのタグ及び注釈付きコーパスを用いて日本語の入力文を意味解析するシステムを提案する．

2 先行研究

英語版フレームネットを多言語に適用する代表的な手法の 1 つとして，対訳コーパスを用いた意味タグの移植に関する研究 [3] がある．具体的には，対訳文に対してそれぞれ構文解析を行い，得られた木構造から様々な素性を取り出して英文と対訳文の対応関係をモデル化する．英文に付けられた意味タグを目的言語の対訳文に写し，そのタグ付き対訳文を目標言語の新たなコーパスとして用いる手法である．(図 1) Tonelli[4] らも同様に，対訳コーパスを用いて意味タグを移植し，イタリア語の新たなコーパスとして用いることを試みているが，意味タグの移植に関しては構文木だけでなく，対応する意味フレームや語彙に関する情報を用いている．以下にその手順の概要を示す．

1. 意味タグを付与された語の中から，意味フレームの語彙などを用いて意味的な主要部分を探し，対応するイタリア語を見つける
2. 対応したイタリア語を含む構文木のノードから遡り，適切らしいノードに意味タグを付与する

Tonelli らの手法の最も着目すべき点は，文の形やパターンといった形式的な情報を用いて意味タグ付与の範囲同定を行うのではなく，意味的に,「何となくそれっぽいところ」を大胆にも予測してしまう点である．また，そのための言語資源として対訳辞書を用いてい

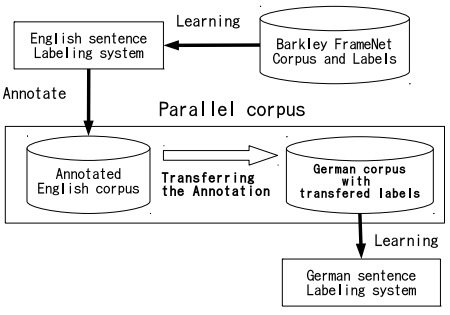

図 1: 対訳コーパスを用いた注釈の移植

る点も重要である．対訳コーパスよりも，単純な対訳辞書の方が網羅性等の点でも信頼性が高い．形式的なレイヤーでは，先に信憑性のある資源で言語間の橋渡しをしてしまい，抽象的なレイヤーでは言語に依存しない意味情報を便りに文を解析する手法である．

しかし，Tonelli らの手法や他の対訳コーパスを用いた手法は，タグを移植した対訳コーパスの文を用いて対象言語の意味解析器を学習させる事を前提としているため，ある程度の規模の対訳コーパスが必要になる。また，対訳コーパスのタグ移植が完全でないため，必然的に学習された意味解析器の性能は低下する．

そこで，本研究では Tonelli らの手法を基にして，対訳コーパスを使わず，対訳辞書と意味フレームの情報を足がかりに英語版フレームネットを直接用いて，日本語の意味解析をする手法を提案する．

3 提案手法

3.1 手法概要

本研究で提案する意味解析法は以下の通りである．先ず入力文を表現するのにふさわしい意味フレームを先に推測する．次に，Tonelli の方法と同様に「主要な情報で，何となく似ている部分」を探す．これは，主要な部分，つまり意味タグが付与された箇所に関して，意味的に類似している語を探すことである．最後にそれを足がかりにして，形式的に妥当性を確かめ，正しい場所にタグを移す．

3.2 単語の対応関係の抽出

まず，前処理として入力文を解析し，各単語に対応する英単語群を抽出する．入力文は構文解析を行って係り受け関係を木構造で抽出する．入力文の形態素は m_{j1} から m_{ja} に分けられる．次に，形態素 m_j の各々対して和英辞書を用い，先ず日本語として使われ得る概念，用法の一覧を列挙する．これらは， c_1 から c_{jb} にまで分けられる．各々の日本語概念 c_j は，和英辞書のレコードから，その日本語と同じ意味を表記できる複数の英語表現 t_{j1} から t_{jc} に対応付けられる．t_j は 1 つの英単語であったり，1 つの日本語概念を表す句であったりするので，t_j はいくつかの英単語 w_{j1} から w_{jd} に分けられる.(図 2)

これを具体的な例で追って見てみると，次の様になる．

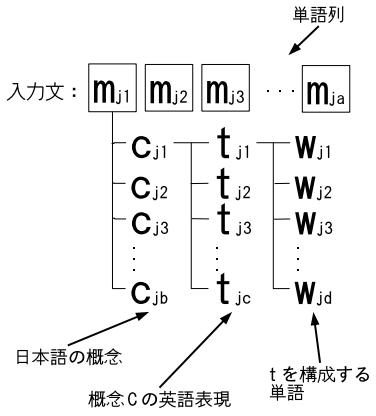

図 2: 入力文と訳語候補のデータ構造

形態素「食べる」には，日本語の用法として「食事を取る」,「生活をしてゆく」等の用法がある．これらが c_j になる．例えば「生活をしてゆく」を英語に訳すとこれにも様々な表現があり，”live”，”be supported（食べさせてもらう）” 等の英訳が存在する．これらが t_j になる.”be supported”の場合には，1 つの英訳として含まれている語が 2 つあるので，$w_{j1} =$ ”*be*” と $w_{j2} =$ ”*support*” というように分解される．

3.3 入力文に該当する意味フレームの同定

ここでは，入力文がおおよそどの話題なのか，どのフレームなのかを定める．

文の意味を定めるのに重要な役割を果たしている部分，動詞に着目する．動詞の形態素 m_{jv} の 訳語 w_{jv} と意味フレームの Lexical units の動詞を総当たりで比較し，最も一致した数が多かった意味フレームを，入力文が該当する場面として定める．この後の処理は，全てこのフレーム内のコーパスに対して操作を行う．

ここで，Lexical units とは，その意味フレームを想起させる語が，英語版フレームネットに予め備えられている．その場面であると強く印象づける言葉が集められている．

3.4 入力文の単語とコーパスの英単語間の尤度

次に，意味タグが付与されたノードに対して，意味的に近い入力文のノードを探す．

コーパス文には英単語 w_{e1} から w_{ef} までが並んでいて，それらは p_{e1} から p_{eg} の句に区切られている．

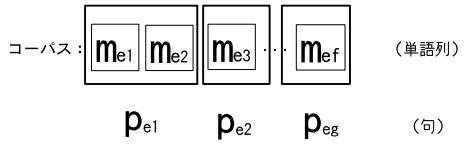

図 3: コーパスのデータ構造

(図 3) それらの句 p_e 中には，意味タグを付与されているものがある．

日本語形態素の m_j と，入力文の内容に合致するとして同定されたコーパスの英単語 w_e の尤度の計算はワードネットを用いた手法 [5] で行う．この手法は，synset と呼ばれる概念の組を与える事で，ワードネットにおける 2 つの概念間の距離を測る．先ず，英訳語中の単語 w_j と コーパスの単語 w_e の尤度を考える．

英訳表現 t_j とコーパスの句 p_e を brill's tagger[6] を用いて pos タグを付与する．与えられた pos タグを用いて"word#pos"の組を作り，ワードネットにおいて"word#pos"に該当する全ての synset を列挙する．t_j と p_e の synset をそれぞれ st_j, sy_e すれば，尤度を総当たりで計算した中での最大値がわかる．

$$Sim_{word}(w_j, w_e) = \max(Sim(st_j, st_e))$$

これを，英訳語中の単語 w_j と コーパスの単語 w_e の尤度とする．

次に，1 つの日本語形態素の概念を表現している英語表現 t_j とコーパス中の単語 w_e の尤度 $Sim_{phrase}(t_j, w_e)$ は以下の様に定める．

$$Sim_{phrase}(t_j, w_e) = \sum_i Sim_{word}(w_{ji}, w_e)(w_{ji} \in t_j)$$

1 つの日本語概念 c_j とコーパス中の単語 w_e の尤度 $Sim_{concept}(c_i, w_e)$ も同様に定める．

$$Sim_{concept}(c_j, w_e) = \sum_i Sim_{phrase}(t_{ji}, w_e)(t_{ji} \in c_j)$$

日本語形態素 m_j とコーパス中の単語 w_e の尤度 $Sim_{je}(m_j, w_e)$ は， m_j の曖昧性を解消する意味から， c_j の中で最大のものを 1 つ選び，これを尤度とする．

$$Sim_{je}(m_j, w_e) = \max Sim_{concept}(c_{ji}, w_e)(c_{ji} \in m_j)$$

コーパス中で意味タグが付与されている句 p_{et} の中にある語 w_e 全てと，入力文の形態素 m_j 全てに関して総当たりで尤度計算を行い，

$$Sim_{p_e}(p_{et}) = \max Sim_{je}(m_j, w_e i)(w_{ei} \in p_{et},)$$

で表される値を，入力文に対する句 p_{et} の尤度とする．

1 つの意味フレームには複数のコーパスが含まれているため，次段で意味タグを付与する範囲を定める前に，どのコーパスを参考にして範囲を定めるか，を決めなければならない．そこで，各コーパスにおいて，先の $Sim_{p_e}(p_{et})$ の総和を入力文とコーパスの尤度とする．

$$\text{入力文とコーパスの尤度} = \sum_t Sim_{p_e}(p_{et})$$

ただし， p_{et} は，意味タグが付与された句である．

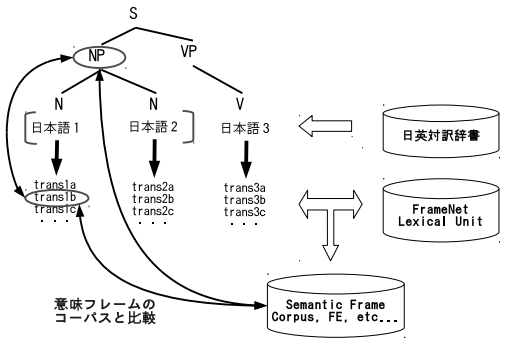

図 4: 提案手法

タグを付与する範囲の決定は，先ず，入力文が該当した意味フレームの中で，最も高い尤度のコーパスと入力文を比較する．Tonelli らの手法に基づき， $w_e \in p_{et}$ で $\max Sim_{je}(m_j, w_e)$ となる語 w_e に対応する日本語の形態素 m_j から構文木のノードを辿る．句 p_{et} には英語版フレームネットに予め構文解析が施されており，Phrase Type タグが付与されている．これを用いて，w_e から辿って p_{et} と同じ性質（品詞）を持つノードに p_{et} の意味タグを付与する．(図 4)

本手法は英語版フレームネットの資源を直接用いるため，対訳コーパスを用いた方法とは異なり，人手により構築された資源で直接意味解析が可能である．そのため，対訳コーパスを用いた間接的な手法よりも，高精度な解析が期待できる．また，本手法は多言語化が可能であると考えられる．本手法の多言語対応は，対象言語の構文解析と訳語候補選定のための辞書が利用可能であれば，どの言語に対しても可能であると考える．

4 実験と結果

日本語フレームネット [2] のコーパスの中から，タグを外した生コーパスを無作為に 143 文を使用し，そ

れらを入力として提案手法によりフレームの推定を行った．日本語フレームネットにおいて，入力された生コーパスに設定されている意味フレームと同様のフレームが本手法によって推定できれば正解とした．結果を [表 1] に示す．

総数	正解	不正解	同定不可
143 文	23 文	120 文	31 文

表 1: 実験結果

適合率は 25.5%，再現率は 16%となった．ここで，同定不可とは合致すると推定されるフレームが定まらなかった事を示す．この同定不可だった 31 文の殆どは，形容詞＋助動詞の形をとるものだった.(図 5) これらは両方合わせて 1 つの動詞に相当する働きを持つものだが，動詞のみをフレーム判定に用いていたため，これらのような述部を持つ入力文に対して，フレームを推定できなかった．このパターンの推定は，形容詞も推定に含めるようにすることで同定不可を回避できると考える．

形容動詞（ be ＋ 形容詞）への未対応

例） 顔が真っ赤だった。
コップはテーブルの上になかった。

図 5: 形容詞＋助動詞の文型

もう一つ，精度を下げた要因として複文や重文など，動詞が複数ある入力への対応である.(図 6) どちらの動詞の文を主体として捉えるべきか，と言う問題が生じた為，今回の実験では最初に見つけた動詞をそのままフレーム推定に用いてしまった．これらの入力の場合，意味フレームに設定されている Lexical units との尤度が最も高かった入力文の語に対して，最も近いノードにある動詞を選ぶことで，この問題を回避できるのではないかと考えている．

副文や重文への対応

例）私は画廊を出て、
ようやくホテルにたどり着いた。

図 6: 重文の例

5 まとめと今後の課題

本稿では，英語版フレームネットを利用した日本語意味解析の手法を提案した．この手法は，現存する英語版フレームネットが拡張されれば，日本語の意味解析にも直接その効果を得られる手法である．また，対訳辞書と構文解析器さえ利用可能であれば，日本語以外の言語にも同様のアプローチが可能であると考える．

現在行ったフレーム推定実験では，改善の余地が多くある事が判明している．今後はそれらを改善し，次のタグ付け及びその範囲同定の実験も行う予定である．その後，日本語フレームネットの生コーパスに本手法でタグ付けを行い，元の注釈と比較した性能評価を行う予定である．

参考文献

[1] Collin F.Baker,Charles J.Fillmore,John B.Lowe(1998). The Berkeley FrameNet Project *COLING '98 Proceedings of the 17th international conference on Computational linguistics - Volume 1 Association for Computational Linguistics Stroudsburg*, PA, USA 1998

[2] 肥塚真輔, 岡本紘幸, 斎藤博昭, 小原京子 (2007). 日本語フレームネットに基づく意味役割推定. 自然言語処理 14.1:43-66

[3] Pado,S.and M.Lapata(2009). Cross-Lingual Annotation Projection of Semantic Roles. *Journal of Artificial Intelligence Research* 36:307-340

[4] Tonelli,S. and E.Pianta(2008). Frame Information Transfer from English to Italian. *In Proceedings of LREC 2008*,Marrakech,Morocco

[5] Lin D.(1998). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*,Madison, WI.

[6] Brill,Eric(1992). A simple rule-based part of speech tagger. *Speech and Natural Language: Proceedings of a WorkshopHeld at Harriman, New York.Morgan Kaufmann Publishers, Inc.*, San Francisco, California. 112116.