語彙の階層情報を利用した日本語フレームネットの意味役割推定

曽根孝明，小原京子，斎藤博昭
慶應義塾大学理工学部

1　はじめに

意味役割とは，文の各部分が動詞に対してもつ役割のことである．この意味役割情報は，質問応答，機械翻訳等において有用性が報告されている．意味役割に基づく大規模な言語資源の代表が FrameNet[1]である．その日本語版として，日本語フレームネット[2]が構築されている．本稿では，日本語フレームネットの意味役割推定を目的とする．

1.1　日本語フレームネットについて

日本語フレームネットでは語が使われる典型的な場面であるフレームごとに意味役割が定義され，人手で意味役割がタグ付けされた例文が存在している．定義された意味役割は主要な意味役割である core とそれ以外である non-core に分けられている．フレームに属する動詞は，見出し語とよばれ，到着を表す **Arriving** フレームならば，「着く」,「至る」,「入る」,「たどり着く」などが見出し語として該当する．

また，フレーム間には継承関係がある．例えば，上下の移動を表す **Motion_directional** フレームは動きを表す **Motion** フレームを継承し，**Motion_directional** フレームの各意味役割はそれぞれ **Motion** フレームの意味役割に対応している．

2　関連研究

日本語フレームネットに基づいた意味役割推定としては，肥塚らの研究があげられる [肥塚 2007]．肥塚らは日本語フレームネットの意味タグ付き事例を学習の正解データとし，最大エントロピー法を用いて意味役割を推定した．

肥塚らの手法では，意味役割を付与すべき項[3]が与えられていない場合，適合率 0.63，再現率 0.43 という結果であった．

[1] http://framenet.icsi.berkeley.edu/
[2] http://jfn.st.hc.keio.ac.jp/
[3] 意味役割が付与される単位のこと

3　提案手法

まず日本語フレームネットの意味タグ付き例文から意味役割を日本語語彙大系の意味クラスに対応付ける．この対応付けを用い，日本語語彙大系の意味クラス階層上の距離に基づき類似度を求め，意味役割と格助詞との結びつきやすさである格結合度を事前に決定する．最後に，類似度と格結合度から意味役割を推定する．

以下それぞれについて詳しく説明する．

3.1　意味役割と意味クラスの対応付け

3.1.1　意味タグ付き例文からの意味役割の獲得

意味役割を推定するフレームの意味タグ付き例文を利用し，意味役割と項のペアを得る．例えば図 1 の場合，[*Theme*，取材班は]，[*Goal*，パキスタン側にも] というペアが得られる．

取材班は	パキスタン側にも	入り	ました．
Theme	*Goal*	**見出し語**	

図 1: 意味タグ付き例文

また，対象フレームを継承するフレームが存在する場合には，そのフレームの文からも意味役割と項のペアを獲得する．

3.1.2　日本語語彙大系の意味クラスとの対応付け

3.1.1 節で得られた項と日本語語彙大系の意味クラスを対応付ける．基本的には項の一番最後の名詞に対応する日本語語彙大系の意味クラスを割り当てる．

これにより，日本語フレームネットの意味役割と日本語語彙大系の意味クラスの対応付けが獲得される．しかし，1 つの名詞に対して 1 つも対応する意味クラスが存在しない場合や逆に複数存在してしまう場合がある．それについては固有表現を考慮することと余計な意味クラスの削除を行うことによって対応する．

固有表現について

同じ名詞に対応する意味クラスが複数存在するものや，日本語語彙大系の見出し語として存在せずうまく意味クラスを獲得できないものがあるため，固有表現抽出器を利用する．

固有表現抽出機としては CaboCha を用いる．CaboCha の固有表現抽出は IREX[4]に基づくタグを出力する．これを利用し，LOCATION というタグならば場所に対応する意味クラス，TIME というタグならば時間に対応する意味クラスというようそれぞれのタグに対応する意味クラスを割り当てる．

不要な意味クラスについて

固有表現抽出では対処できないものについては，日本語語彙大系の意味クラス階層の各ノードに重み付けをし，ノード間の距離に基づいて意味クラスを削除することで対応する．

例えば,「本」という語で日本語語彙大系をみると，出版物としての本のほかに，単位としての本も意味クラスとして存在する．そのため，読み物としての「本」という意味で使われている場合に，単純にこれらすべてを対応付けると，誤った意味役割を推定を推定してしまうということが考えられる．

重み付け

以下の重み付けに関する記述は正津らの手法を参考にしたものである [正津 2001].

日本語意味大系は,「名詞」,「固有名詞」,「事象」をルートとする 3 つのツリーからなっている．しかし，深さが一段階変化することによるノード間の意味の変化が一定ではない．例えば,「悪人」と「魔物・化け物」の日本語語彙大系上のノードから，ルートノードである「名詞」までは以下のようになっている．

- 悪人 → 悪人等 → 善人・悪人等 → 人間＜性向＞ → 人間＜能力＞ → 人間 → 人 →主体 → 具体 → 名詞
- 魔物・化け物 → 準人間 → 人 → 主体 → 具体 → 名詞

これを見ると同じ一段階の変化でも「悪人」 → 「悪人等」の間の方が「魔物・化け物」 → 「準人間」の間に比べて意味の変化が少ないことが分かる．同じ一段階で意味の変化に差がある場合，単純にノードの深さだけで類似性を判断しようとすると，間違いが生じることが予想される．

この問題に対して提案手法では，日本語語彙大系の各ノードに重み付けを行うことで対応する．各ノードに対する重みづけは以下のように行う．

$$W(c) = \begin{cases} 50 & (d < 3) \\ 1/d & (d \geq 4, C = \phi) \\ \sum_{c_i \in C} W(c_i) & (d \geq 4, C \neq \phi) \end{cases}$$

ただし，c は重み付けされるノード，d は c の深さ，C は c の子クラスの集合である．意味クラスの深さはルートノードを 0 として計算する．

この重み付けにより，ルートノードから遠ざかるほど，ノードの重みが軽くなることになる．

不要な意味クラスの除去

以下の手順で，不要な意味クラスの除去を行う．

1. 各ノードごとに，他のノードへの距離の合計を計算する
2. 距離の合計の平均を求め，他のノードへの距離の合計が平均以上のノードは削除する
3. 距離が 100 以上の経路を削除する
4. 閉じた経路の中で，各ノードごとに他のノードへの距離の合計を計算し，それが最小のノード以外を削除する

1，2 によって集団から外れているノードを削除し，3，4 でノード群の代表を抽出することを目的としている．

3.2 格結合度

本稿では意味役割と格助詞との結びつきやすさを格結合度と呼ぶことにする．意味役割推定に用いるため格結合度を事前に求めておく．

3.2.1 格助詞と意味クラスの組の獲得

日本語コーパス[5]の小説データのうち，意味役割を推定しようとするフレームの見出し語を含む文から格助詞と意味クラスの組を抽出する．例えば,「姉が病院へ着いた。」という文があった場合，[が，(姉，女)]，[へ，(病院，公共機関)] の組が得られる[6].

[4] http://nlp.cs.nyu.edu/irex/index-j.html

[5] http://www.tokuteicorpus.jp/

[6]「姉」,「女」,「病院」,「公共機関」は意味クラスである

3.2.2 格結合度の計算

3.2.1 節で得られた，格助詞ごとに収集された意味クラスの集合をもとに，格結合度を以下の式で求める．

$$Sim_{case}(fe, case) = \frac{\sum_{c \in C} Sim_{arg}(fe, c)}{|KEY_C|} \quad (1)$$

$$Sim_{arg}(fe, c) = max_{x \in fe} sim(x, c) \quad (2)$$

$$sim(x, y) = \frac{2L}{l_x + l_y} \quad (3)$$

$Sim_{case}(fe, case)$ は格結合度である．fe は意味役割と対応づけされた意味クラスの集合であり，3.1 節の方法で得られる．$case$ は格を表す．C は 3.2.1 節で $case$ ごとに収集された意味クラスの集合であり，KEY_C はその数を表す．

$Sim_{arg}(fe, c)$ は意味クラス群と意味クラスの類似度を表す．fe は意味クラス群であり，c は意味クラスである．

$sim(x, y)$ は 2 つの意味クラスの類似度である．x，y はともに意味クラスを表し，L はルートノードから x と y の共通ノードまでの距離，l_x と l_y はそれぞれの意味クラスまでのルートノードからの距離を表す．

3.3 意味役割の推定

見出し語を含む文に対して，意味役割を推定する手順について述べる．

項候補の決定

意味役割を推定するためには，文を項に分ける必要がある．

項候補を得るには以下の手順をとる．

1. CaboCha を使い係り受け関係を求める
2. 見出し語に係っている文節をその文節にかかっている文節も含め項候補とする
3. 見出し語が係っている文節も項候補とする

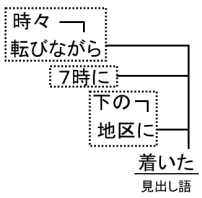

図 2: 対象項候補獲得の例

例えば「時々転びながら 7 時に下の地区に着いた」という文を見出し語が「着く」であるとして考えた場合，図 2 のようになる．実線は係り受けを表す．点線で囲まれた部分，つまり「時々転びながら」,「7 時に」,「下の地区に」が項候補となる．

スコアの計算

項候補に対してすべての意味役割との間で，式 (4) で定義される $score$ を計算する．

$$score = \frac{Sim_{case}(fe, case) + Sim_{arg}(fe, c)}{2} \quad (4)$$

ただし，fe は 3.1 節で得られた日本語語彙大系の意味クラスに対応づけされた意味役割の集合，$case$ は項候補のもつ格，Sim_{case} は式 (1) の格結合度，c は項候補の意味クラス，Sim_{arg} は式 (2) の類似度を表す．

意味役割の重複

基本的には項候補に対し $score$ がもっとも高い意味役割を付与する．しかし，この方法ではある意味役割の項が複数存在する可能性が残る．意味役割には，複数の項が同じ意味役割をもつことはないという特性が存在するため，ある意味役割が複数の項で最も高い $score$ であった場合，最も $score$ が高い項にその意味役割を付与する．

4 実験

実験対象フレームは **Arriving** フレーム，**Motion** フレームとする．意味役割推定をする文としては，2002 年度の毎日新聞の記事から各フレームの見出し語を含む文をランダムに 100 文抽出したものを利用する．また，役割推定の対象とする意味役割として，フレームにおける主要な意味役割である core のみを対象にした場合と core 以外も含めた場合の両方について実験を行う．

4.1 実験対象フレーム

実験対象フレームについての情報を表 1 に示す．また，**Motion** フレームを継承するフレームのうち意味タグ付き例文が存在するフレームについての情報を表 2 に示す[7]．

評価は適合率，再現率，F 値によって行う．

4.2 実験結果

実験結果を表 3 に示す．また，項候補同定誤りを除外した結果を表 4 に示す．

[7]フレーム間の継承関係については，FrameNet version 1.3 に従う．

表 1: 実験対象フレーム

意味フレーム	**Arriving**
意味役割	*Goal, Theme, Cotheme, Depictive, Goal_conditions,Manner, Means, Mode_of_transportation, Path, Source, Time*
見出し語	至る，入る，たどり着く，着く
意味タグ付き例文	154 文
意味フレーム	**Motion**
意味役割	*Area, Direction, Goal, Path, Source, Theme, Carrier, Degree, Distance, Duration, Manner, Place, Purpose, Result, Speed, Time*
見出し語	動く，止まる，停まる，行く，飛ぶ，跳ぶ
意味タグ付き例文	0 文

表 2: Motion フレームを継承するフレーム

意味フレーム	**Motion_directional**
見出し語	おりる，上がる，上る，下がる，下る，沈む，登る，落ちる
意味タグ付き例文	171 文
意味フレーム	**Traversing**
見出し語	たどる，横切る，渡る，通る
意味タグ付き例文	129 文
意味フレーム	**Fluidic_motion**
見出し語	流れる
意味タグ付き例文	11 文

表 3: 実験結果 (項候補同定誤りを含む)

	適合率	再現率	F 値
Arriving(core のみ)	0.72	0.78	0.75
Arriving(core 以外も含む)	0.66	0.61	0.64
Motion(core のみ)	0.64	0.65	0.64
Motion(core 以外も含む)	0.55	0.44	0.49

表 4: 項候補同定誤りを除外した結果

	適合率
Arriving(core のみ)	0.89
Arriving(core 以外も含む)	0.87
Motion(core のみ)	0.78
Motion(core 以外も含む)	0.71

5 考察

表 3 と表 4 の比較から分かるように，項候補の同定誤りを取り除くことで適合率，再現率ともに大幅な向上がみられた．項候補の同定誤りを取り除くことが，意味役割推定の精度を向上させることにつながると言える．

また，表 3 で示されているように，**Arriving** フレームと **Motion** フレームを比較すると，適合率，再現率ともに **Arriving** フレームの方がおよそ 0.1 から 0.2 程良い値となっている．原因としては，以下の 2 つが考えられる．

1 つ目は意味タグ付き例文の有無である．**Arriving** フレームには，意味タグ付き例文が 154 文存在する．一方，**Motion** フレーム自体には意味タグ付き例文が存在せず，継承関係のみを利用して意味役割の推定がなされた．**Motion** フレームでは対象フレーム自体の意味タグ付き例文を参照できなかったことが結果の悪さにつながっていると考えることができる．

2 つ目は似た意味クラスに対応付けられると考えられる意味役割の存在である．**Arriving** フレームには，場所に類する意味クラスと対応付けされる意味役割が，*Goal*, *Source* の 2 種類存在する．一方 **Motion** フレームには，場所に類する意味クラスと対応付けされる意味役割が，*Area*, *Path*, *Goal*, *Source*, *Place* の 5 つ存在する．本手法では，類似度と格結合度は意味クラスに対応付けられた意味役割をもとにして計算されるため，似た意味クラスに対応付けられる意味役割が多いフレームに関しては，意味役割推定の難易度が高くなると考えられる．

6 おわりに

意味タグ付き例文が比較的豊富に存在する **Arriving** フレームにおいて，主な意味役割に関して適合率 0.72，再現率 0.78 という結果を残したことで，日本語フレームネットに基づく意味役割の推定において語彙の階層情報が有用であることが分かった．

また，意味タグ付き例文が存在せず継承関係のみを用いて意味役割を推定した **Motion** フレームにおいても主な意味役割に関して適合率 0.64，再現率 0.65 という結果を残したことで，フレーム間の継承関係が，意味役割の推定に有用であることを示した．

参考文献

- 肥塚 真輔，岡本 紘幸，斎藤 博昭，小原 京子.“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，Vol.14，No. 1，pp.43-66，2007.
- 正津 康弘，白井 清昭，徳永 健伸，田中 穂積.“国語辞典の語釈文の解析と語義のシソーラスへのマッピング”，第 *15* 回人工知能学会全国大会論文誌，pp.115-120，2001.