述部機能表現の意味ラベルタガー

今村 賢治†, 泉 朋子†, 菊井 玄一郎†, 佐藤理史‡
†日本電信電話株式会社 NTT サイバースペース研究所
{imamura.kenji,izumi.tomoko,kikui.genichiro}@lab.ntt.co.jp
‡名古屋大学大学院 工学研究科
ssato@nuee.nagoya-u.ac.jp

1　はじめに

日本語では，1つ以上の形態素が組み合わさって，助詞や助動詞などの機能語とほぼ同等の構成要素として働く表現が多数存在する．これを機能表現と呼ぶ．

たとえば,「パソコンが壊れてしまったかも知れない。」という文を形態素解析すると,「パソコン/が/壊れ/て/しまっ/た/かも/知れ/ない/。」という形態素列が得られる．このうち,「が」「て」「しまっ」「た」「かも」「ない」が機能語であるが,「て/しまっ」という形態素列が「完了」という意味を,「かも/知れ/ない」という形態素列が「推量」という意味を表現している．また,「た」は1形態素で「完了」を表す．この文における主たる述語は，最後に現れる内容語形態素とすると「知れる」であるが，上記機能表現を1つの単位としてまとめると「壊れる」となる．

ブログ，twitterなど，ユーザが直接発信するメディアの拡大に伴い，そこから意見情報などを抽出する試みが多く提案されている．ユーザの主観的情報は，日本語においては述部に現れることが多く，述部機能表現を正しく特定することは，後段の言語処理にとって，重要な要素となる．

このような機能表現を網羅的に集めた辞書として，日本語機能表現辞書「つつじ」(松吉他, 2007)がある．「つつじ」は，日本語の機能表現の表層形約17,000種に対して，そのID，意味，文法的機能，音韻的変化などが網羅的に収録された辞書である．つつじでは，機能表現の意味として，89種類のラベルを定義している．本稿では，形態素解析結果に対して，述部を同定し，つつじの意味ラベルを機能表現に付与するタガーについて述べる．

2　出力する意味ラベル

本稿の意味ラベルタガーの目的は2点である．

述部を同定し，内容部と機能部を区別する
述部の機能表現に，その意味を表すラベルを付与する

表 1: 出力する意味ラベル

種別	詳細	出力意味ラベル
内容部 (C)	述語	C,PRED
	機能動詞構造	C,VN C,P C,LV
機能部 (F)	つつじ意味ラベル	F, 推量など
	意味ラベル不明	F,NULL
接続部 (J)	つつじ意味ラベル	J, 順接確定など
その他 (*)	その他	*,*

以上の目的を達成するため，出力するラベルを2つのレベルに分けた（表1）．一つは，述部のうち，その表現が内容部に属するのか，機能部に属するのか，接続表現（接続詞など）なのかというレベルである．内容部は，主として述語を同定するために用い，機能部は機能表現列の同定に，接続部は等位接続文を判断するために用いる．

内容部はさらに，通常の述語と機能動詞構造に分けている．機能動詞構造は，動作性名詞と機能動詞が組み合わさっているもので，動作性名詞を述語として扱う方が都合がよいことが多いため，両者を区別する．

機能部および接続部の意味ラベルは，主としてつつじに収録された意味ラベルを付与する．しかし，つつじに存在しない機能語があった場合，形態素単位でNULLを付与し，意味ラベル不明を表している．また，述部ではない形態素には，‘*’ を付与する．

結果，出力する意味ラベルは，つつじの意味ラベル数*2+6種類となっている．

3　意味ラベルタガー

前述のとおり，機能表現は複数の形態素から成り立っている．また，表層的には同じでも，文脈により意味が異なる場合も多い．そのため，つつじを利用して，意味ラベルを付与しようとすると，機能表現の範囲の同定と，曖昧性解消を同時に行わなければならない．

範囲同定と曖昧性解消を同時に行うには，大きく2

つの方法がある．一つは，日本語形態素解析で採用されている方法である．形態素解析では，入力文字列で辞書引きを行い，形態素候補を取得，その中から，統計モデルや接続表を用いてもっともらしい形態素列を出力する方法を取っている（たとえば (工藤他, 2004)）．もう一つの方法は，IOB2 タグ形式 (Sang and Veenstra, 1999) を用いた系列ラベリングのように，範囲とその内容を示すラベルを，個々の入力単位（文字や形態素など）に付与する方法である．

本稿で述べる意味ラベルタガーは，形態素解析と同様な方式を取る．言い換えると，形態素解析結果を入力とする形態素再解析器として構成する．具体的には，辞書とルールを用いて “フレーズラティス” を構築し，統計モデルに基づいて最尤フレーズ列を探索する．ここでいうフレーズとは，複数形態素に対して，1 つの意味ラベルが付与されたもので，表現の単位となる．本稿では出力情報として意味ラベルのみを扱うが，つつじを候補獲得のための辞書として用いるため，同時に ID など，つつじに収録された他の情報も出力することができる．以下，ラティス構築，最尤パス探索，統計モデルの学習について述べる．

3.1 フレーズラティスの構築

形態素列が意味ラベルタガーに入力されると，まず，機能表現辞書つつじと形態素の品詞やパターンマッチを用い，表層形と意味ラベルをセットにしてフレーズ候補を生成し，ラティス構造にする．各フレーズ候補は，ほぼ 2 節で述べた出力意味ラベルに対応する．

内容部の候補

述語
形態素の品詞を参照し，それが述語になりうる品詞（動詞，形容詞，名詞+だ）である場合，PRED を意味ラベルとしてフレーズを作成する．

機能動詞構造
動作性名詞 (VN)-格助詞 (P)-機能動詞 (LV) または，動作性名詞 (VN)-「する (LV)」のパターンを持つ形態素列をパターンマッチで抽出し (Izumi et al., 2011)，フレーズを作成する．

機能部・接続部の候補

つつじに登録された機能表現
複数の形態素の表記を結合した部分形態素列表記でつつじを検索する．マッチするものすべてについて，機能部 (F) のフレーズを生成する．また，マッチしたものが 1 形態素で，かつ接続助詞に属する品詞を持つ場合は接続部 (J) のフレーズも生成する．

機能語形態素
形態素の品詞を参照し，それが機能語である場合，「機能表現であるが，意味は不明」を表す NULL を意味ラベルとして機能部フレーズを作成する．

その他

入力の全形態素について,「述部ではない」ことを表す意味ラベル ‘*’ を使用してフレーズを作成する．

なお現在は，以下のフレーズも補助的に追加している．

訓練コーパスから正解フレーズを抽出・集約し，一致するものがあればラティスに追加する．これは，訓練時に正解パスがラティスに含まれるようにするためである．

このように作成されたラティス構造は，必ず文頭から文末に至るパスが存在するため，あとはどのパスが尤もらしいのか，選択することになる．

つつじを辞書として用い，候補を選択する利点としては，つつじに収録された意味ラベル以外の情報も付与できる他に，考慮すべきタグ連鎖が劇的に少なくなるという利点もある．系列ラベリングは，全ラベル連鎖を考慮するため，IOB2 タグ形式を使用すると，最低でも出力意味ラベルの 2 倍（89 種類なら，最低限 178 種類）のタグ連鎖を考慮しなければならない．つつじを辞書として用いることにより，最尤パス探索および学習時の処理時間が高速になる．

3.2 最尤パス探索

最尤パス探索では，3.1 節で作成したラティス構造から，最尤のフレーズ列を探索する．本稿では，統計モデルとして，3.3 節で述べる識別モデルを使用する．最尤パスは，以下の式を満たすフレーズ列となる．

$$\hat{P} = \underset{P \in P_L}{\operatorname{argmax}} \sum_k w_k f_k(P) \qquad (1)$$

なお，P はフレーズ列，P_L はラティス内の全フレーズ列，$f_k(P)$ は，フレーズ列 P が与えられたときの k 番目の素性，w_k は，素性 $f_k(P)$ に対応する重みである．(1) 式を満たすフレーズ列は，動的計画法を用いて探索する．

最尤パス探索およびモデルの学習に用いる素性は，系列ラベリングの考え方を利用する．すなわち，入力（形態素列）を出力（意味ラベル）に対応づけるため

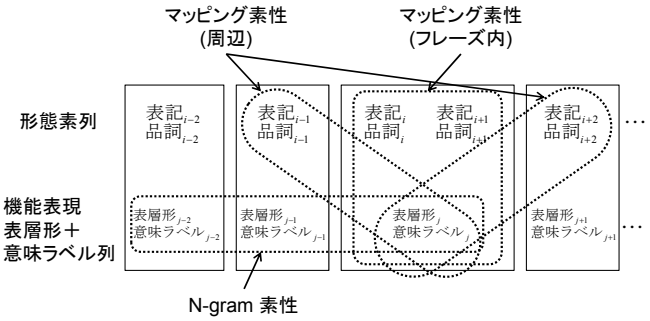

図 1: 素性テンプレートのイメージ図

の素性（本稿ではマッピング素性と呼ぶ）と，出力系列（意味ラベル列）の素性（同 N-gram 素性）である．マッピング素性，Ngram 素性ともに，入力形態素の全情報（表記，品詞，活用型など）と出力の意味情報（意味ラベル，つつじ ID）を使うものと，入力形態素の品詞情報と出力の意味ラベルのみを使うものを両方使用している．

マッピング素性は，基本的には 1 フレーズについて，入力の部分形態素列と，出力の意味ラベルを対応づけているが，系列ラベリングでよく利用されるように，入力の形態素列についてはフレーズ内に閉じる必要がない．本稿でも同様に，フレーズの直前・直後 1 形態素を周辺形態素として，マッピング素性として用いる（図 1）[1]．また，N-gram 素性には，Averaged Perceptron の場合 1〜3-grams，CRF の場合 1〜2-grams を用いる．

3.3 学習/パラメータ推定

パラメータ推定では，3.2 節で述べた素性の重みを学習する．本稿では，学習方式として，Averaged Perceptron (AVRP; (Collins, 2002)) または条件付き確率場 (CRF; (La erty et al., 2001)) を用いる．Averaged Perceptron はオンライン学習の一種で，現在のモデルでの最尤パスが正解になるように学習する．一方 CRF はバッチ学習で，ラティスに含まれる全パスに対して正解が最大確率になるように学習する．どちらも識別学習であるので，正解となる教師データが必要である．

4 実験

4.1 実験条件

コーパス　新聞記事 (News)，ブログ記事 (Blog) について，部分形態素列とその意味ラベルを人手で付与したものを正解コーパスとして学習およびテストを行った．コーパスサイズは，News が 2,115 文，Blog が 2,113 文と，比較的小さい．今回は述部の機能表現を正しく抽出・同定するのが目的であるため，正解コーパスも述部のみにラベル付を行った．したがって，名詞句に後続する格助詞など，述部ではない場所に現れる機能表現にはラベルを付与していない．

機能表現辞書　実験では，ブログなどの表現に対応するため，機能表現辞書つつじを拡張したものを用いる．具体的には，意味ラベルとして，受身，使役，丁寧，尊敬，体験，困難，容易の 7 ラベル，のべ 666 エントリを追加したものを使用した[2]．

実験方法　実験は，News ドメインを教師データとして Blog ドメインをテスト（またはその逆）する，オープンドメインテストと，News, Blog データを混合し，10 分割交差検定（クローズドドメインテスト）の 2 種類を行った．

評価は，まず正解のラベル列とシステム出力のラベル列（部分形態素列を含む）について，編集距離が最小となるようにアライメントを行い，一致したラベルを正答として F 値を算出した．

ベースライン　本稿では，ベースライン手法として，CRF による系列ラベリングを用いる．系列ラベリングツールとして，CRF++ [3]を用いた．素性は，3.2 節で説明したものとほぼ同様なものを用いたが，ツールの制限により，N-gram 素性に関しては，意味ラベルのみを用い，Unigram と Bigram だけを用いた．また，ラベルは，IOB2 タグ形式に展開して実施した．なお，評価の際は，システム出力の BI タグ連続をまとめて，フレーズを再現したのちに F 値を算出している．

4.2 実験結果

表 2 に，実験結果を示す．なお，表中の「全ラベル」とは，すべてのラベルを考慮した精度である．「有効ラベル」とは，機能表現でも述語でもない（その他）ラベルを取り除いて評価した精度であり，実用的な精度を表す．

まず，交差検定（クローズドドメイン）の精度について，提案方式と系列ラベリングを比較すると，全ラベル，有効ラベルともに，精度の差は少ない．提案方式 (AVRP) と提案方式 (CRF) の差は，学習方式の違いに起因すると考えられるが，いずれにしても，学習

[1] MeCab では，周辺素性を用いていない．

[2] この他に引用，前後関係-前の 2 ラベルをコーパスに追加した．

[3] http://crfpp.sourceforge.net/

表 2: 提案方式/系列ラベリングによる意味ラベル付与の精度

訓練	テスト	方式	全ラベル F 値	有効ラベル 適合率	有効ラベル 再現率	有効ラベル F 値
News	Blog	提案方式 (AVRP)	**0.957**	0.867 (10,955/12,631)	**0.824** (10,955/13,298)	0.845
		提案方式 (CRF)	0.956	**0.890** (10,737/12,059)	0.807 (10,737/13,298)	**0.847**
		系列ラベリング (CRF)	0.955	0.887 (10,576/11,921)	0.795 (10,576/13,298)	0.839
Blog	News	提案方式 (AVRP)	**0.949**	0.868 (11,778/13,571)	**0.818** (11,778/14,404)	0.842
		提案方式 (CRF)	0.948	**0.886** (11,653/13,145)	0.809 (11,653/14,404)	**0.846**
		系列ラベリング (CRF)	0.942	0.878 (11,134/12,682)	0.773 (11,134/14,404)	0.822
交差検定		提案方式 (AVRP)	0.958	0.895 (23,087/25,788)	0.833 (23,087/27,702)	0.863
		提案方式 (CRF)	0.959	**0.900** (23,342/25,941)	**0.843** (23,342/27,702)	**0.870**
		系列ラベリング (CRF)	**0.960**	**0.900** (23,274/25,860)	0.840 (23,274/27,702)	0.869

表 3: 部分別同定精度 (提案方式 AVRP)

訓練	テスト	精度 (F 値) 述部	機能表現	内容語	述語
News	Blog	0.830	0.874	0.823	0.859
Blog	News	0.834	0.860	0.827	0.860
交差検定		0.849	0.886	0.846	0.873

とテストが同一ドメインであれば，従来の系列ラベリングを使用すれば，機能表現の意味ラベルを精度よく付与可能である．

一方，オープンドメイン（News 訓練 Blog テスト，またはその逆）では，全ラベルの F 値は，方式間の差異は少ないが，有効ラベルでは提案方式の方が精度が高い結果となった．とくに，再現率が大幅に向上している．つつじは機能表現を網羅的に収録しているため，訓練コーパスに現れなかった機能表現も同定することができた．このように，辞書と統計モデルを併用するタガーは，さまざまなドメインを取り扱わなければならない場合に有効である．

次に，意味ラベルタガーの出力から，内容語列＋機能表現列を 1 つの述部とみなし，述部抽出精度を測定した結果を，表 3 に示す．約 85%について，述部を完全に抽出した．内容部と機能部・接続部を分けた場合，若干機能部の表現列の同定精度がよい．エラーを分析すると，内容部の場合，機能動詞構造が本動詞と認定されてしまう誤りが多く，機能表現では，長い機能表現が複数の短い機能表現に分割される場合が多かった．長い機能表現は，正解コーパス上でも複数に分割されている場合があり，これが影響したものと考えられる．機能表現の単位を再考する必要がある．

5 まとめ

本稿では，形態素列から述部を同定し，機能表現の意味をラベルとして付与する意味ラベルタガーについて述べた．本タガーは，述部の同定と機能表現の意味が同定できるため，文の意味理解，機能表現の正規化 (Izumi et al., 2010) などに利用できる．

本稿の方式は，機能表現辞書つつじと，識別モデルに基づく最尤選択を組み合わせて機能表現を同定する．辞書を用いてラティス作成を行うことにより，以下の利点がある．

- 教師データとは異なるドメインにおいて，系列ラベリングより精度がよい．
- 辞書につつじ ID など，様々な情報が格納できるため，出力結果を他の処理に利用しやすい．
- ラティスサイズが小さくなるため，学習が速い．

参考文献

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP-2002*, pages 1–8.

Tomoko Izumi, Kenji Imamura, Genichiro Kikui, and Satoshi Sato. 2010. Standardizing complex functional expressions in japanese predicates: Applying theoretically-based paraphrasing rules. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 64–72.

Tomoko Izumi, Kenji Imamura, Genichiro Kikui, Atsushi Fujita, and Satoshi Sato. 2011. Paraphrasing japanese light verb constructions: Towards the normalization of complex predicates. *International Journal of Computer Processing Of Languages (IJCPOL)*. to appear.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-2001*, pages 282–289.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of EACL-1999*, pages 173–179.

工藤 拓, 山本 薫, 松本 裕治. 2004. Conditional random fields を用いた日本語形態素解析. 情報処理学会 自然言語処理研究会 NL-161-13, pages 89–96.

松吉 俊, 佐藤 理史, 宇津呂 武仁. 2007. 日本語機能表現辞書の編纂. 自然言語処理, 14(5):123–146, 10 月.