5Q-02

日本語文における機能表現意味ラベル付与と事実性解析への効果

上岡裕大† 成田和弥† 菅野美和† 水野淳太‡ 乾健太郎†
† 東北大学 ‡ NICT
{yudai.k, narita, meihe, inui}@ecei.tohoku.ac.jp junta-m@nict.go.jp

1 はじめに

日本語には，語彙的意味をほとんど担わず，統語的関係や話し手の主観的情報を表す表現が存在する．

(1) パソコンが壊れ てしまっ た かもしれない。

例えば，(1) では「てしまっ」「た」「かもしれない」がそれぞれ〈無意志〉〈完了〉〈推量-不確実〉という話者の主観的情報を表している．これらの表現のうち，「てしまっ」や「かもしれない」のように複数の語が組み合わさってはじめて意味をなす表現は複合辞と呼ばれる．本稿では，機能語と複合辞をまとめて機能表現と呼び，これらの意味を認識する処理を機能表現解析と呼ぶ．

機能表現解析は，事実性解析や機械翻訳を始めとする応用分野で必須となる基盤技術である．Narita et al.[1] は，機能表現を元に事実性の解析に取り組み，機能表現が持つ曖昧性に起因する事実性解析誤りが多いことを指摘している．しかしながら，機能表現解析に関する先行研究は意外にも少なく [2][3]，解析器開発の基礎となる大規模コーパスも存在しない．

本研究では，機能表現の意味ラベル体系の設計，コーパス構築，機能表現解析器の構築を行う．また，応用課題の一つである事実性解析に適用し，その効果を検証する．

2 意味ラベルの設計とコーパス構築

今回は，応用課題の一つである事実性解析に影響を与える機能表現を正しく解析することを目的として，述部の機能表現を対象にラベルを設計した．機能表現の意味ラベル体系は，松吉ら [4] が機能表現辞書『つつじ』で意味カテゴリとして整理している．しかし，収録されていない表現があるだけでなく，区別の困難な意味カテゴリも含まれている．そこで，本研究では，『つつじ』の 89 種類のカテゴリを元に，〈無意志〉，〈完了〉など，67 種類の意味ラベルを定義した．ラベルの定義は，事実性解析結果をフィードバックさせながら追加，統合，細分化を繰り返して行った．

(2) 仙台では毎年七夕祭りが開催され ている。

例えば，「ている」という機能表現は，『つつじ』では〈継続〉のカテゴリに分類される．しかし，(2) では，「開催する」は継続的に行われているのではなく，習慣として行われる．このような表現は該当するカテゴリが存在しないため，新たに〈習慣〉ラベルを定義し，これを付与する．

このラベル体系に基づき，現代日本語書き言葉均衡コーパス (BCCWJ)[1] の Yahoo!知恵袋中の述部機能表現に対してラベルを付与した[2]．ラベルは形態素単位で付与し，複合辞は IOB2 形式で付与した．

現在，機能表現を含む文を中心にランダムに抽出した 1,545 文のアノテーションが完了している．コーパス全体に現れる機能表現数は 5,993 個であり，その異なり総数は 584 個であった．

表 1: 範囲同定の性能評価

	精度	再現率	F 値
ベースライン	90.90	74.42	81.84
CRF	95.39	95.93	95.66

表 2: 機能表現解析器の評価

	精度	再現率	F 値
ベースライン	71.73	61.72	66.35
CRF	79.83	81.18	80.50

3 評価実験

機能表現解析の現状を明らかにするため，条件付確率場 (CRF)[3] を利用した系列ラベリング問題として評価実験を行った．2 節で構築したコーパスに対して 10 分割交差検定を行った．学習素性には，形態素素性およびその組み合わせを使用した [5]．CRF の有効性を確認するため，比較手法として機能表現辞書および直前の形態素に関する接続制約に基づく最長一致で解析を行うベースラインを用意した．機能表現に曖昧性がある場合は，候補のうちコーパス中での出現頻度が最も高いラベルを選択する．機能表現辞書および接続制約は，コーパスから得られた情報を『つつじ』に追加したものを使用する．いずれの手法においても，文内の機能表現列の開始，終了位置は正解を与えた．具体的には，述語の位置を与え，その直後から文末までの形態素列が解析対象となる．評価は機能表現を 1 単位として行った．また，機能表現解析の難しさは複合辞の範囲同定と曖昧性解消である．そこで，曖昧性解消までしない範囲同定の性能評価も行った．

機能表現解析の結果を表 1，2 に示す．これらの結果より，CRF を用いることでベースラインよりも高い性能での機能表現解析が可能であることが分かった．CRF の結果から，範囲同定は比較的高い性能で行えるが，曖昧性解消は難しいことが分かった．

(3) a. いつも読ん でいる 雑誌でもかまいません。
(正解:習慣 出力:結果状態)
b. 両親とも働い ている のが条件です。
(正解:継続 出力:結果状態)
c. 感情の高ぶりがよく描かれ ている。
(正解:結果状態 出力:結果状態)

[1] http://www.ninjal.ac.jp/corpus_center/bccwj/

[2] 構築したコーパスは，BCCWJ との差分データとして，アノテーション仕様と合わせて次の URL で公開している．
http://tinyurl.com/ja-fe-corpus

[3] 実装には，CRFSuite (http://www.chokkan.org/software/crfsuite/) を使用した．

例えば，文 (3a)，(3b) 中の「ている」という機能表現は，いずれも〈結果状態〉と解析された．〈結果状態〉は，正しくは (3c) のような場合に付与されるラベルである．誤りの原因は，周辺単語やその形態素素性から区別することが難しいためである．これらを正しく解析するためには，文内の副詞や述語の種類 (動作を表すか状態を表すかなど) を考慮する必要がある．コーパス中の 5,993 機能表現のうち 2,739 表現は少なくとも 2 つ以上の意味を持つ可能性があり，曖昧性がある．

4 事実性解析への適用

事実性解析において，機能表現解析の効果を検証する．本稿では，他の述語の影響を排除するため，機能表現が付随する主事象（主節に含まれる事象）である 1,475 事象のみを解析対象とする．事実性は，Narita et al. [1] と同様に，確信度（CT，PR，U）と肯否極性（+，−）の組によって表す．即ち，CT+，PR+，PR−，CT−，U の 5 種類のラベルのいずれかに，各主事象の事実性を分類する．今回構築したコーパスには拡張モダリティタグ [6] が付与されているため，Narita et al. と同様に，拡張モダリティタグをもとに正解となる事実性ラベルを定めた．1,475 事象中の各ラベルの分布を表 3 に示す．機能表現が付随する事象のみを対象としているため，CT+ではなく，U が最多となっている．

4.1 解析モデル

事実性解析のモデルは，Narita et al. と同様に，主事象に付随する機能表現の意味ラベルを利用することで決定する．例えば，〈否定〉の機能表現が付随している場合には肯否極性を反転する，という事実性更新ルールを適用する．更新ルールは以下の 3 種類を用いる．カッコ内は対応する意味ラベルを表す．

1. 肯否極性：+→ −, − →+（〈否定〉〈不可能〉など）
2. 確信度：CT→PR（〈推量-不確実〉〈意志〉など）
3. 確信度：CT→U, PR→U（〈疑問〉〈依頼〉など）

無標のラベルである CT+から始めて，文末から順に機能表現を参照し，更新ルールの割り当てられた機能表現があれば該当する更新ルールを適用する．すべての機能表現の更新ルールを適用することで，主事象の事実性を決定する．なお，疑問符も事実性に影響を与える要素として考えられるが，機能表現が事実性に与える影響について分析するため，本稿では採用していない．

4.2 評価・考察

表 4 に，機能表現を利用した事実性解析器の評価として，各ラベルごとの精度，再現率，F 値のマクロ平均を示す．機能表現の意味ラベルとしては，3 節で用いたベースラインによる解析結果，CRF による解析結果，および正解ラベルを用いた．

CRF による解析結果を利用した場合，ベースラインによる解析結果を利用した場合と比較して性能が向上した．

(4) 5階くらいから落ちて助かったんでした よね。
（ベースライン: U，CRF: CT+，正解: CT+）

(4) では，主事象「助かる」の事実性は CT+である．ベースラインでは頻度が高い意味ラベルを採用するため，「よね」を〈疑問〉と判断している．そのため，主事象「助かる」の事実性は U と誤解析される．一方 CRF では「でした」や句点といった周辺情報をもとに，「よね」を〈態度〉と正しく判断することができたため，事実性も正しく解析することができた．このように，機能表現解析をより精緻に行うことが，事実性解析に対して有効であることが確認できた．

表 3: 事実性ラベルの分布

事実性ラベル	CT+	PR+	PR−	CT−	U
事例数	476	215	51	107	626

表 4: 機能表現解析結果に基づく事実性解析の評価

機能表現解析手法	精度	再現率	F 値
ベースライン	48.34	40.30	41.55
CRF	55.70	48.38	50.42
正解ラベル	57.36	52.75	54.15

正解ラベルを利用した場合に着目すると，機能表現が正しく与えられているにも関わらず，事実性が正しく解析されない事例が少なくないことがわかる．誤り分析を行ったところ，同じ意味ラベルの機能表現をもっていても，異なる事実性をもつ事象が見られた．

(5) どうやって色を判別してる んでしょうか？
（正解ラベルに基づくシステム: U，正解: CT+）

(5) では，下線部に対して〈疑問〉が付与されているため，主事象「判別する」の事実性は U と解析された．しかしながら，前提として起こった事象である「判別する」の方法を問う文であるため，CT+が正解である．このような事象の事実性を解析するために，〈疑問〉を〈疑問-方法〉のように機能表現レベルで細分化すべきなのか，あるいは事実性解析の段階で文脈を用いて区別すべきなのかは，議論の余地が多分に残されている．

5 まとめ

本稿では，機能表現意味ラベル付与コーパスを構築し，機能表現解析の現状について述べた．機械学習を用いた解析では，F=80.50 で解析することができた．曖昧性のある機能表現の分類が課題であり，今後は，CRF を用いたチャンキングを行った後，曖昧性解消問題として機能表現を分類するなど，曖昧性のある機能表現の分類方法を検討していきたい．また，事実性解析への適用実験から，機能表現解析が事実性解析に有効であることが確認できた．しかし，正解の機能表現ラベルを用いても正しく事実性解析が行えない事例も少なくなかった．今後は，どこまでを機能表現解析で扱うべきかを検討していく．

謝辞

本研究は文部科学省科研費（23240018），および JST 戦略的創造研究推進事業 CREST の一環として行われた．

参考文献

[1] Kazuya Narita, Junta Mizuno, and Kentaro Inui. A lexicon-based investigation of research issues in Japanese factuality analysis. In *In Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pp. 587–595, 2013.

[2] 鈴木敬文, 阿部佑亮, 宇津呂武仁, 松吉俊, 土屋雅稔. 代表・派生関係を利用した日本語機能表現の解析方式の評価. 言語処理学会第 18 回年次大会予稿集, pp. 598–601, 2012.

[3] 今村賢治, 泉朋子, 菊井玄一郎, 佐藤理史. 述部機能表現の意味ラベルタガー. 言語処理学会第 17 回年次大会論文集, pp. 2–5, 2011.

[4] 松吉俊, 佐藤理史, 宇津呂武仁. 日本語機能表現辞書の編纂. 自然言語処理, Vol. 14, No. 5, pp. 123–146, 2007.

[5] 上岡裕大, 成田和弥, 水野淳太, 乾健太郎. 述部機能表現に対する意味ラベル付与. 情報処理学会研究報告 第 216 回自然言語処理研究会, 第 2014-NL-216 巻, pp. 1–9, 2014.

[6] 松吉俊, 江口萌, 佐尾ちとせ, 村上浩司, 乾健太郎, 松本裕治. テキスト情報分析のための判断情報アノテーション. 電子情報通信学会論文誌 D, Vol. J93-D, No. 6, pp. 705–713, 2010.