1Q-2

文末機能表現シソーラスの網羅性の検証

松木 久幸 † 佐藤 理史 † 駒谷 和範 †

† 名古屋大学大学院 工学研究科 電子情報システム専攻

1. はじめに

日本語には，文末の用言に接続する表現が多数存在する．本研究では，これらの表現を**文末機能表現**と呼ぶ．文末機能表現には，同じ意味を表す表現が複数存在するため，データマイニングなどの応用タスクにおいては，これらの表現を一つの代表表現へと言い換えること（文末機能表現の正規化）が求められている [1].

本稿では，文末機能表現の正規化を実現するために我々が作成した文末機能表現シソーラスとそれに基づく述部正規化システム [2] を評価する．具体的には，述部正規化システムに含まれる文末機能表現の検出システム（以降，**検出システム**と呼ぶ）を用いて，与えられた文に含まれる文末機能表現を検出できるかどうかを調べ，それを通してシソーラスの網羅性を検証する．以下では，このために行った２つの実験とその結果について述べる．

2. 実験１：検出システムの評価

実験 1 では，正解を付与したテストデータを用いて，検出システムとベースラインシステムを比較する．

2.1 正解データの作成

Yahoo!知恵袋の質問データ（以降，知恵袋と呼ぶ）と毎日新聞 2000 年（以降，毎日新聞と呼ぶ）のそれぞれから，ランダムに 1000 文 (句点以外の記号で終わる文は除外) を選択し，それぞれの文に，人手で正解を付与した．具体的には，文末機能表現が含まれている場合は，その部分にマークを付与し，含まれていない場合は，含まれてないということを表す記号を付与した．

2.2 検出システムとベースラインシステム

我々が作成した文末機能表現検出システムは，文末機能表現シソーラスに基づいて，文末機能表現を検出し，その表現が，シソーラスのどのエントリに対応するかを出力する．そのため，シソーラスに未収録の文末機能表現が存在すると，そのような表現を検出することができない．一方，形態素情報のみを用いて文末機能表現を検出する方法を採用すれば，このような問題は発生しないが，誤検出の問題が発生することが予想される．

本実験のために，形態素情報のみから文末機能表現を検出する**ベースラインシステム**を作成した．このシステムは，以下の方法で，文末機能表現を検出する．

1. 入力文を形態素解析する (Mecab+JumanDic).
2. 各形態素に対して，以下のルールに従い **P**・**F**・**O** のいずれかのラベルを付与する．
   - 品詞が形容詞・動詞・判定詞のいずれかの場合，ラベル **P** を付与する．
   - 品詞が助動詞・助詞・接尾辞・形式名詞のいずれかの場合，ラベル **F** を付与する．
   - 品詞が上記以外の場合，ラベル **O** を付与する．
3. ラベルが付与された形態素列（$m_n...m_0$）に対して，以下を満たす形態素 m_i が存在するかチェックする．
   条件 1 m_i のラベルが **P**
   条件 2 $m_{i-1}...m_0$ のラベルがすべて **F**
4. 存在した場合は，m_i を文末用言，$m_{i-1}...m_0$ を文末機能表現と判定し，文末機能表現と文末用言の表層形・活用形の３つ組を出力する．存在しなかった場合，その文には文末機能表現が存在しないと判定し，記号 ‘–’ を出力する．

2.3 結果

人手で付与した正解と，検出システムおよびベースラインシステムの出力を比較した結果を表 1 に示す．この表において，‘Correct’ の列は，システムの出力が正解と一致した数を表し，‘Incorrect’ の列は，正解と一致しなかった文の数を表す．それらの下の記号 ‘E:D’，‘N:N’，‘E:D’，‘E:N’，‘N:D’ は，正解とシステムの出力の詳細を表すラベルで，‘:’ の左側の記号が正解の詳細（文末機能表現が存在する場合は ‘E’，しない場合は ‘N’），右側の記号がシステム出力の詳細（なんらかの表現を検出した場合は ‘D’，文末機能表現が存在しないと判定した場合は ‘N’）を表す．ここで，‘Incorrect’ の ‘E:D’ は，たとえば，正解の文末機能表現が「〜はじめてもらいたい」である文において，システムが「〜もらいたい」のみを検出した（検出部分が短い）場合が相当する．

我々の検出システムは，知恵袋で 95.9%（959/1,000），毎日新聞で 98.3%（983/1,000）の文に対して正しい出力が得られた．文末機能表現が含まれる文に限定した場合は，知恵袋で 95.0%（747/786），毎日新聞で 94.4%（252/267）となる．いずれの場合も，我々の検出システムは，ベースラインシステムよりよい性能を示している．

文末機能表現が含まれる文に対して，検出システムが正しく検出できなかった要因を調べた（表 2）．正しく検出できなかった文（知恵袋で 39 文，毎日新聞で 15 文）の要因のうち，半分（知恵袋で 16 文，毎日新聞で 7 文）が，当該文末機能表現がシソーラスに収録されていないことであった．

3. 実験２：大規模テキストを用いた検証

実験２では，大規模テキストを用いて，シソーラスに未収録の文末機能表現（出現頻度が高いものに限定）がどれくらい存在するかを見積もる．

表 1: 正解との比較

		Correct		Incorrect		
		E:D	N:N	E:D	E:N	N:D
知恵袋	正解	786	214	–	–	–
	検出システム	747	212	27	12	2
	ベースライン	640	202	95	51	12
毎日新聞	正解	267	736	–	–	–
	検出システム	252	731	10	5	2
	ベースライン	207	726	19	41	7

表 2: 正解が出力されなかった要因

一致しなかった要因	知恵袋	毎日新聞
形態素解析誤り	18	8
シソーラスに未収録	16	7
同定システムのバグ	3	0
入力文が正しくない	2	0

3.1 方法

大規模テキストに対して，人手で正解を付与することは，コストが大きい．次善の策として，検出システムとベースラインシステムの出力結果を比較し，シソーラスに未収録の文末機能表現の数を推定する．具体的には，以下の手続きで推定する．

1. 大規模テキストの各文を検出システムとベースラインシステムに与え，２つのシステムの出力の組を作成する．
2. これらの出力の組のうち，検出システムの出力が，ベースラインシステムの出力より短かいものを選択し，異なり（活用形＋文末機能表現）でまとめる．
3. 出現頻度が 10 以上の組を人手で調べる．

大規模テキストとして，Yahoo!知恵袋の質問データの一部（1,500,000 文），毎日新聞 1 年分（1,330,802 文）を使用した．いずれも，句点以外の記号で終わる文は除外した．

3.2 結果

検出システムとベースラインシステムの出力組の数を表 3 に示す．この表において，‘B’ はベースラインシステムが表現を検出した場合，’S’ は検出システムが表現を検出した場合，’N’ はシステムが表現を検出しなかった場合を表す．これらの組のうち，上記の手続きにおいて，人手での調査対象となったものは，知恵袋で 536 組，毎日新聞で 191 組あった．そのうち，文末機能表現と判定した表現は，知恵袋で 72 個，毎日新聞で 29 個であった．

実験 1 の条件下で，ベースラインシステムと検出システムのオーバーラップ（異なり数）を数えた（表 4）．実験 1 のテキストにおいて，検出システムで検出できない文末機能表現のうち，ベースラインシステムのみが検出できる文末機能表現の占める割合は，知恵袋で 6/28，毎日新聞で 1/11 であった．実験２で用いたテキストにおいても，この割合が保存されていると仮定すると，頻度 10 回以上の文末機能表現のうち，検出システムでは検出できない表現は，知恵袋で 336 ($= 72 \times 28/6$) 個，毎日新聞で 319 ($= 29 \times 11/1$) 個であると推定できる．

これらの推定値を用いて，出現頻度が 10 回以上の文末機能表現のうち，シソーラスに未収録の表現がどれくらい存在するかの割合を求めた（表 5 の E）．検出システムが検出した表現のうち，頻度が 10 回以上であったのは，知恵袋で 2,689 個，毎日新聞で 1,674 個である．出現頻度が 10 回以上の文末機能表現のうち，11.1%（336/3,025），19.6%（319/1,993），がシソーラスに収録されていなかった．今後は，これらの表現を収録するように，シソーラスを改良する必要がある．

表 3: ベースラインシステムと検出システムの出力組

		出力が一致		出力が不一致		
		B:S	N:N	B:S	B:N	N:S
知恵袋	総数	900,596	349,449	138,884	25,882	85,189
	異なり数	6,147	1	8,910	4,874	2,022
毎日新聞	総数	290,886	960,942	20,091	10,647	48,236
	異なり数	4,529	1	4,003	2,411	2,305

表 4: ベースラインシステムと検出システムの比較

知恵袋：E (786 文)

	T_{thes}	F_{thes}
T_{base}	113	6
F_{base}	47	22

毎日新聞：E (267 文)

	T_{thes}	F_{thes}
T_{base}	79	1
F_{base}	38	10

表中の数字は，異なり数を表す．

表 5: 未収録語数の推定 (頻度 10 回以上)

	知恵袋	毎日新聞
A:手順２のリストに含まれる文末機能表現の異なり数	72	29
B:実験 1 における，ベースラインのみが検出できる割合	6/28	1/11
C:シソーラスに未収録の表現の異なり数の予測値 (A/B)	336	319
D:検出システムが検出した表現の異なり数	2,689	1,674
E:シソーラスに未収録の表現の割合 (C/(C+D))	11.1%	19.6%

謝辞 本研究は，JSPS 科学研究費基盤研究 (B)「平易な日本語表現への工学的アプローチ」課題番号 24300052 の助成を受けている．本研究では，現代日本語書き言葉均衡コーパス DVD 版・毎日新聞 2000 年・Yahoo!知恵袋データを利用した．

参考文献

[1] T. Izumi, K. Imamura, G. Kikui, and S. Sato. Standardizing Complex Functional Expressions in Japanese Predicates: Applying Theoretically-Based Paraphrasing Rules. In *23rd International Conference on Computational Linguistics*, p. 64, 2010.

[2] 松木久幸, 佐藤理史, 駒谷和範. 文末機能表現シソーラスと述部正規化システム．第２回コーパス日本語学ワークショップ, pp. 185–194, 2012.