

6K-01

自動語尾抽出に基づく対話システムにおけるキャラクタ性の構築

長島大和^{†1} 阿倍博信^{†1}東京電機大学^{†1}

1. はじめに

機械に自然言語を入力すると、それに対応した応答文を返す対話システムにおいて、話し方が一定でない場合、情緒の不安定さ、説得力のなさ、人間関係の距離感がうまく取れていないといった問題が発生する場合がある[1]。これらの問題に対して、対話システムに一貫したキャラクタ性を持たせることが有効であると考えられる。キャラクタ性とは、『そのキャラクタを特徴付ける重要な要素』のことであり、例えば、「おれは今腹が減っているのだ。」、「あたしは今お腹が空いてるの。」という2つの文章の場合、一般的に前者は男性、後者は女性の発言であるというように認識できる。このように、発言者に付与される性別などの属性と結びついた要素のことをキャラクタ性と呼ぶ。

こうしたキャラクタ性を対話システムに付与することにより、その機械と対話する際の人間らしさの向上と、対話の満足感の向上をはかることができる。

2. 関連研究

関連研究として、主に対話モデル及び音変換の面からそれぞれキャラクタ性を再現することを目的とした研究が存在する。

小倉らが開発したDeepEVE[2]では、対話モデルの学習を工夫することにより、小説の登場人物のような喋り方をする対話システムの構築を目指している。当該研究では、小説内部より抽出した対話を自動的にキャラクタごとに分類し、それを元に対話モデルを生成することで、キャラクタ性を持った喋り方を実現している。しかしながら、データ量が不足している問題により、純粋な機械学習モデルではキャラクタの再現が難しいことが予想されている。

また、刀山らは、日本語の文章の内部のどこにキャラクタ性を埋め込むことができるのかについて調査している[3]。当該研究では、日本語の対話を分析した結果として、特にキャラクタ性が強く出る場所は終助詞を含む文末表現と、特定の機能表現であると考察されている。

3. 提案方式

3.1 基本方針

本研究では、キャラクタ性の構築を目的として、対話の中に含まれている語尾を自動的に抽出する方式について提案する。

その理由であるが、小倉らの研究[2]より、キャラクタ性の含まれるデータ量は多くの場合において機械学習に必要なデータ量に達せず、別のアプローチが必要であると考えたからである。

また、刀山らの研究[3]より、キャラクタ性の多くは文末表現にあらわれていることから、自動的に語尾を抽出し、発言を個別に検証し分析することが、対話システムにおけるキャラクタ性の実装に必要であると考えたからである。

この方式の利点としては、キャラクタの発言のみを集めることで動作し、対話ペアの作成を省略することが可能である点があげられる。これにより、一人のキャラクタに対しての特徴を明示的に抽出し、それを元にキャラクタの分析に役立てることが想定できる。

3.2 自動語尾抽出方式の概要

図1に本研究で提案する自動語尾抽出方式の概要について示す。

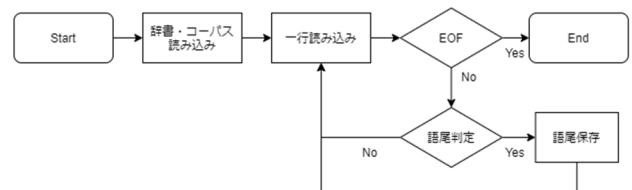


図1 自動語尾抽出の概要

最初に、キャラクタの発言をまとめたテキストファイルを読み込み、文中表現などを一度整形し、また、相槌などに代表されるような短い発言を除外するために一定文字数以下の応答を除外する前処理を実行する。

次に、後述するキャラ語尾の辞書を読み込んだ後、条件付き確率場を用いた形態素解析器を利用し解析を行う。この時点で、多くの場合においてキャラクタ性を持った語尾は「助動詞」「終助詞」などに分類される文末表現が中心となると刀山らの研究[3]より判明しているため、この点を考慮し過検出を防ぐために提案方式ではこれら2属性を持った単語に限って抽出し語尾候補とする。

また、複数の助動詞・終助詞が組み合わさることで語尾として成立する単語がいくらか存在する。例えば「だもん」と言う語尾は、「だ」という一般的な助動詞と、「もん」という終助詞によって構成される単語である。そのため、単純な単語一つを抽出するのみでは、「だもん」が出現した際に、「もん」のみを抽出してしまうことが予想される。これらは後述するキャラ語尾辞書に記載していくことも考えられるが、作業コストとのバランスを考慮し、事前の単語の属性が語

Creating Characteristics in Dialogue Systems based on Automatic Word-End Extraction
^{†1} YAMATO NAGASHIMA, HIRONOBU ABE, Tokyo Denki University

尾になりうる場合にはそれらをまとめて抽出する。

また、時折語尾が存在しない文章(体言止め、長い領きなど)が存在することが分かっており、これらを適切に処理するため上記の語尾になりうる属性を持つ単語が存在しなかった場合には抽出結果を空とすることで、抽出に悪影響が出ないように対応する。

これらの処理をテキストファイルの EOF まで継続して行い、最終的に語尾の抽出結果としてファイル出力する。

3.3 キャラ語尾

語尾抽出の際の問題として、キャラクタを際立たせるために付与されるキャラ語尾が存在する。例えば、犬のキャラクタであれば「～ワン」、猫のキャラクタであれば「～ニャ」などと言った語尾が付与される場合がある。これらの表現幅は非常に幅が広く、創作物などのキャラクタから語尾を抽出する際に課題になることが予想される。そこで、本研究では事前にアニメ・漫画などの作品から 200 単語程度のキャラ語尾を手動で抽出し、事前にユーザ辞書を作成し、語尾抽出の際に読み込むことで対策を行う。

4. 評価

4.1 評価データ

下記の 3 種類のコーパスからそれぞれ 2 種類ずつデータを評価データとして選定した。関西弁コーパス[4]はヘファナン・ケビンが作成したコーパスである。名大会話コーパス[5]は国立国語研究所が作成したコーパスである。Manga109[6][7]は東京大学が作成した 109 の漫画で構成されたデータセットであり、これからキャラクタの台詞を抽出した。これらを、それぞれ関西弁コーパスを KSJ、KYT、名大会話コーパスを NUCC1、NUCC2、Manga109 を M109B、M109M として評価データとした。

4.2 評価結果

前章で提案した自動語尾抽出方式を Python で実装し、評価データを用いて有効性について評価を行った。その結果について表 1 に示す。また、提案方式との比較のため、文末の 1 単語のみを語尾として自動抽出するアルゴリズムを従来方式として実装し、同条件にて評価を行った。その結果を表 2 に示す。評価結果において、それぞれ TP(True Positive)、FP(False Positive)、FN(False Negative)、TN(True Negative)、Recall(再現率)、Precision(精度)、F(F 値)とする。

表 1. 提案方式の評価結果

評価値	KSJ	KYT	NUCC1	NUCC2	M109B	M109M	平均
TP	68	70	85	80	96	73	78.6
FP	62	27	50	40	47	59	48.1
FN	40	68	42	54	55	47	51
TN	31	35	27	34	55	30	35.3
Recall	0.629	0.507	0.669	0.597	0.635	0.608	0.607
Precision	0.523	0.721	0.629	0.667	0.671	0.553	0.628
F	0.571	0.596	0.649	0.63	0.653	0.579	0.613

表 2. 従来方式の評価結果

評価値	KSJ	KYT	NUCC1	NUCC2	M109B	M109M	平均
TP	5	5	5	3	21	8	7.8
FP	29	27	18	37	31	24	27.7
FN	128	134	149	131	140	138	136.7
TN	39	34	32	37	61	39	40.3
Rec	0.038	0.036	0.032	0.022	0.13	0.055	0.052
Prec	0.147	0.156	0.217	0.075	0.404	0.25	0.208
F	0.06	0.058	0.056	0.034	0.197	0.09	0.083

5. 考察

表 1, 表 2 を確認した結果、従来方式では抽出不可能だった語尾が提案方式では抽出率が高く、F 値も向上していることから、提案方式の有効性について確認することができた。

ただし、提案方式の場合でも、例えば「～ちゃうやろうな」という文末の文章の場合、今回採用した助動詞・終助詞に基づく方式では「うな」が語尾候補として抽出されてしまうため、語尾として適切でないものが抽出される場合があった。

特に、M109M は F 値が 0.09 から 0.579 に改善されていることが確認でき、その理由としてはキャラ語尾辞書の存在が考えられる。漫画のように特徴的なキャラ語尾を抽出しておき、事前に辞書として構築しておくことは、語尾抽出方式における抽出精度向上に効果があると考えられる。

6. おわりに

本研究では、対話システムにおいてキャラクタの発言からキャラクタ性を構築する際に重要となる語尾を自動的に抽出する方式について提案した。

提案方式について評価を行った結果、F 値として 0.613 という値を得ることができ、従来方式と比較して有効性について確認することができた。

今後の予定であるが、提案方式によって抽出された語尾を用いたキャラクタごとの語尾のカテゴリ化機能やその結果を用いてキャラクタ性を付与した対話システムの構築及び評価を行っていく予定である。

参考文献

- [1] 東中竜一郎, 稲葉通将, 水上雅博: Python で作る対話システム, オーム社 (2020).
- [2] 小倉拓人, 谷津元樹, 原田実: 人物ベースの Seq2Seq モデルを用いた対話システム DeepEVE, 情報処理学会論文誌 Vol.60, No.3, pp.967-975, (2019).
- [3] 刀山将大他: 文のどこにキャラクタ性を埋め込む自由度があるか, 言語処理学会 第 22 回年次大会 発表論文集, pp.721-725, (2016).
- [4] ヘファナン・ケビン: 関西弁コーパスの紹介, 総合政策研究 41 号, pp.157-164 (2012).
- [5] 藤村逸子, 滝沢直宏: 言語研究の技法:データの収集と分析. pp. 43-72, ひつじ書房 (2011).
- [6] K.Aizawa, A.Fujimoto, A.Otsubo, T.Ogawa, Y.Matsui, K.Tsubota, H.Ikuta: Building a Manga Dataset "Manga109" with Annotations for Multimedia Applications, IEEE MultiMedia (2020).
- [7] Y.Matsui, K.Ito, Y.Aramaki, A.Fujimoto, T.Ogawa, T.Yamasaki, K.Aizawa: Sketch-based Manga Retrieval using Manga109 Dataset, Multimedia Tools and Applications, Springer (2016).