5Q-04

特定の話題に関する対話エージェントの実現に向けた特徴語及び状態抽出法

松林　圭[†]　松原 良和[†]　今野 陽子[‡]　小野 良太[‡]　井上 祐寛[§]　山下 晃弘[†]

[†]東京工業高等専門学校情報工学科　[‡]株式会社調和技研　[§]株式会社クレスコ

1　はじめに

近年，企業内や顧客対応の現場においてチャットサービスが用いられることは少なくない．チャットサービス上の投稿データ分析やコミュニケーション支援の研究も多数取り組まれている．一方でコスト削減や品質向上を目指してチャットにおける対話の自動化を目的とする対話エージェントに関する研究が盛んになってきた．対話エージェントには予約システムやリマインダなど特定のタスクに特化しているタスク指向型エージェント，人との雑談を行いより自然な会話を目的とした非タスク指向型対話エージェントがあり，特に前者においては技術指導やコンサルタントなどの専門性の高い分野での応答の自動化が期待されている．

本研究では，ある限られた特定領域の話題において自動的な会話や応答を実現する対話エージェントの設計・構築方法について検討した．本稿では，検討した手法を用いて開発した食事と健康に関する対話エージェントの開発事例と，実用性の検証結果について報告する．

2　対話エージェントの全体構成

提案手法の全体構成を図1に示す．提案手法では，まず専門的な会話の一つの発話中においてその発話で中心的な意味となる名詞を「発話キー」と定義し，その発話キーに対する状態を示す単語を「状態キー」と定義する．発話から発話キーと状態キーを抽出し，それを用いて会話の分類，状態の判断を行い，複数の応答候補を抽出する．応答候補の中からどの応答を実際に採用するかは評価関数によって決定するが，本稿では，その前段となる発話キーと状態キーの抽出手法に主眼を置く．本研究では，ある特定領域における発話の発話キーは，その領域に特徴的な単語が多く出現すると仮定する．例えば，料理の領域を話題とした会話であれば，料理に関係する単語が発話キーとして多く出現すると仮定する．

Feature Word and State Extraction Method for Dialog Agent on a Specific Topic
† Department of Computer Science, National Institute of Technology, Tokyo College.
‡ CHOWA GIKEN Coporation.
§Cresco ltd.

そこで，インターネット上から一般的な文書とその領域に関係する文書を収集し，その領域の特徴語を抽出することで発話キー候補単語の辞書を予め作成し，その辞書に基づいて発話キーを抽出する方法を提案する．

発話キー候補単語は上記の収集した文書を形態素解析システムの MeCab[1]を用いて名詞を抽出し，特徴量の計算に基づいて決定する．特徴量の計算アルゴリズムは自然言語処理の特徴抽出として一般的に精度が高いと言われている TF-IDF とプロフィール推定において高い精度を持つと言われている[2]赤池情報量基準(AIC)を用いて計算を行い精度の比較をした．また，インターネット上から収集する文書には表記の揺れや新語が含まれる可能性があるため，形態素解析には定期的に新語に対応を行っている辞書 NEologd を用い，名詞は日本語 WordNet を用いて単語をより広い意味の上位語へ変換を行った．状態キー候補単語の抽出には SVM に基づく係り受け解析器 CaboCha[3]を用いてその名詞を修飾する動詞と形容詞を抽出し，結果を状態キーとした．

応答候補を抽出する際には発話キー，状態キーの一致する文書か Word2Vec を用いて発話キー，状態キーをベクトル化したものを Nearest Neighbor 法によって類似度が高い発話キー，状態キーを抽出し，応答候補を求める．

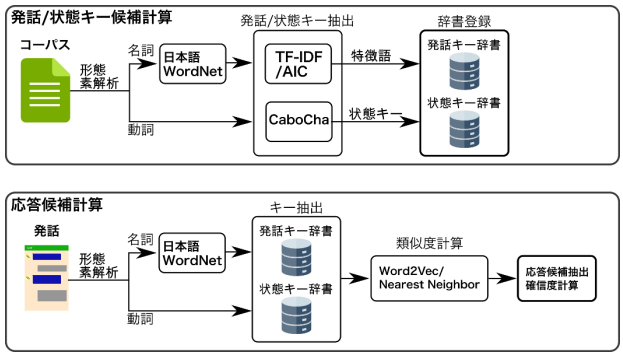

図 1　提案手法の全体図

3　食事健康対話エージェントへの適用

提案手法の有効性評価のため，実際に食事と健康をテーマとする会話データを用いて検証を行った．検証は，会話データから人手で発話キ

ーを抽出した場合と，前述の手法で辞書を生成して抽出した場合で比較検証を行った．

3.1　使用データについて

提案の手法を用いて発話キーを抽出するため，インターネット上からヘルスケアやダイエットなどの食事と健康に関する文書と，一般的な文書の収集を行った．食事と健康に関する文書についてはウェブサイトを目視によって対象領域の文書か判断を行い約 1 万件収集した．また，一般的な文書については①Wikipedia の全記事，②SNS(Twitter)上の記事約 1000 万件，③食事と健康以外のウェブサイトの記事約 3 万件，の 3 つのコーパスを作成して利用した．なお，収集した単語は揺らぎの影響を軽減するために日本語 Wordnet による上位語への変換を行った場合と行わなかった場合の両方で検証を行った．

3.2　計算結果

図 2 に単語をそのまま計算した結果(図上)と Wordnet で上位語に変換した結果(図下)の一部をそれぞれ示す．

Wordnet で上位語に変換するか否かに関わらず，TF-IDF では一般的な文書としていずれのコーパスを用いても計算結果上位の単語はほぼ変化はなかった．一方，AIC ではコーパスによって単語の順位に大きく変化が出た．

TF-IDF		AIC	
Wikipedia	計算値	Wikipedia	計算値
効果	0.4373	効果	97786
ダイエット	0.2819	ダイエット	91380
カロリー	0.2348	ありません	84494
場合	0.1823	おすすめ	81214
摂取	0.1671	気	76725
そう	0.1613	摂取	74560
筋肉	0.1477	健康	71898

TF-IDF(WordNet)		AIC(WordNet)	
Wikipedia	計算値	Wikipedia	計算値
事象	0.3873	絶食	90948
要素	0.2489	ありません	84494
滋養分	0.2307	おすすめ	81214
絶食	0.1851	要素	77929
超分子	0.1692	生体機能	76149
生体機能	0.1682	事象	75843
アクティビティー	0.1658	提案	71870

図 2 AIC/TF-IDF の計算結果の一部

3.3　人手で抽出した発話キーとの比較

精度の比較として，健康と食事に関する会話データから人手で抽出したときの発話キーを正解データとして提案手法で抽出された発話キーがどのくらい正解データを網羅しているかを計算した．その際人手で抽出した文書では複数の文書に同じ発話キーが存在することがあり，単語が重複していた場合と単語の重複を除いた場合をそれぞれ計算し，その結果を図 3 に示す．図 3 は上に WordNet を使用していないもの下に使用したものを示し，そのうち右が重複を除いたもの，左に重複が含まれているものをそれぞれ示している．提案手法で計算した単語の上位 100 件ごとに範囲を増やした結果，使用データ③で AIC を求めたものを除いてそれぞれのグラフ内で同じような傾向が現れた．

WordNet を使用していない方の網羅率はほぼ比例しながら増加し，上位 1000 単語で Wordnet を使用しない方が約 35%，使用したほうが約 40%，一方重複あり(発話キーの出現数からみた網羅率)は，上位 700 単語から Wordnet を使用しない方が約 55%，使用したほうが約 65%となった．

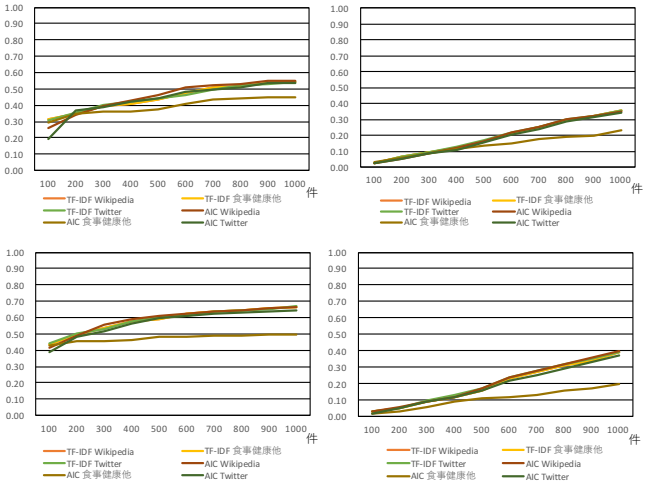

図 3 発話キーの網羅率計算結果

4　まとめ

ある限られた特定領域の話題において自動的な会話や応答を行う対話エージェントについて発話キーと状態キーを抽出することによる手法を提案し，食事と健康分野において実際の会話データから実用性の検証を行い，発話キーを 700〜1000 単語用いることで 60%以上の精度で発話キーを正しく抽出できていることが明らかになった．

参考文献

[1]Mecab, http://taku910.github.io/mecab/
[2]池田 和史,服部 元,他: マーケット分析のための Twitter 投稿者プロフィール推定手法,情報処理学会論文誌,コンシューマ・デバイス & システム,Vol.2,No.1,pp82–93 (Mar. 2012)
[3]CaboCha, https://taku910.github.io/cabocha/