自動抽出した換喩表現を用いた係り受け関係のずれの解消

清田 陽司
科学技術振興機構 さきがけ
(京都大学 学術情報メディアセンター)
kiyota@ar.media.kyoto-u.ac.jp

黒橋 禎夫
東京大学 大学院情報理工学系研究科
kuro@kc.t.u-tokyo.ac.jp

木戸 冬子
マイクロソフト株式会社
fkido@microsoft.com

1 はじめに

テキストを知識源とする質問応答システムは，ユーザの質問に対して単に適合テキストを検索するだけでは十分ではなく，質問とテキストのマッチングを正確に行って，テキスト中から答えそのものを見つけてユーザに返さなくてはならない．これを実現するために，TREC QA track や NTCIR QAC の質問応答タスクの参加システムの大半は構文解析結果にもとづくマッチングを行っている．ダイアログナビ (後述) においても，日本語の係り受け関係を利用している．

しかし実際には，ユーザが質問文に換喩を用いた場合に，構文解析結果にもとづくマッチングが失敗することが多い．換喩とは比喩の一種であり，あるものをそれと関連する別のものに置き換えて表現する現象である [1]．例えば,「漱石を読む」においては「漱石」は「漱石の小説」を指していると考えられる．

換喩の存在によるマッチングの失敗例を図 1 に示す．ここでは，ユーザは「GIF」を「GIF の画像」の換喩として用いていると考えることができる．このとき，ユーザ質問文には「GIF→ 表示」という係り受け関係が存在するが，対応するテキストにはそれは存在しない．そのため，係り受け関係のマッチングを考慮すると，このテキストは低いスコアでしかマッチしない．この問題は，初心者をターゲットとした質問応答システムにおいては特に大きな問題である．なぜならば，初心者はエキスパートと比較して質問文をできるだけ短くしたいという欲求が強く，換喩を頻繁に用いると考えられるからである．

本論文では，コーパスから換喩を自動的に抽出し，係り受け関係のずれの問題を解決する手法を提案する．

2 ダイアログナビ

この節では，ダイアログナビの概要と，本システムにおいて用いているユーザ質問文とテキストのマッチング手法について簡単に述べる．

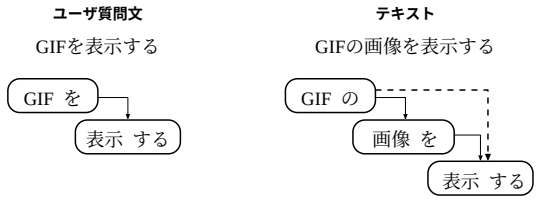


図 1: 換喩の存在によるマッチングの失敗

表 1: ダイアログナビで用いるテキスト知識ベース

知識ベース	件数	文字数	マッチング対象
用語集	4,707	700,000	見出し (1 文)
ヘルプ集	11,306	6,000,000	タイトル (1 文)
サポート技術情報	23,323	22,000,000	文書全体 (複数文)

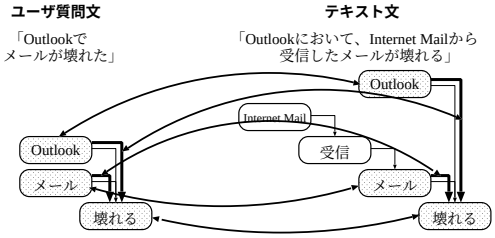
ダイアログナビは，Windows 環境利用者を対象とした対話型の自動質問応答システムである．2002 年 4 月から，http://www.microsoft.com/japan/navigator/ において公開サービスを行っている．本システムは，マイクロソフトがすでに保有しているテキスト知識ベース (表 1) をそのままの形で利用しており，**同義表現辞書**による表現のずれの吸収や**係り受け関係への重みづけ**によるテキストの正確なマッチングを行う．詳細については [2] を参照されたい．

同義表現辞書の一部を図 2 に示す．ユーザ質問文とテキストの間の表現のずれは語のレベルだけでなく，「パソコンを起動する」「Windows を起動する」「電源を入れる」のように 2 文節以上のフレーズレベルにおいても多数存在するので，それらの同義表現を適切にマッチさせるためにグループ化している．

また，係り受け関係を考慮したユーザ質問文とテキストのマッチングの例を図 3 に示す．システムは，ユーザ質問文と各テキスト中のすべての文との間で**類似度**を計算するが，この際に文節間の係り受け関係の一致に重みを与えている．類似度の計算は文節を単位として行っており，キーワード・同義表現の一致にもとづいて両者の文節と係り受け関係を対応づけたのち，互いに対応する文節と係り受け関係の割合 (**被覆率**) を

[読む]
　読む，よむ，読める，よめる，読み込む，よみこむ，読み込める，よみこめる
[メール]
　メール，メイル，電子メール，電子メイル，Mail，E-Mail
[メールを読む]
　メールを読む，メールを受信する，メールを見る，メールを受ける，メッセージを受信する，メッセージを受ける
[パソコンを起動する]
　パソコンを起動する，Windows を起動する，電源を入れる，ブートする，パソコンを立ち上げる，スイッチを入れる

図 2: 同義表現辞書



3	—	**対応をもつ文節の数**	—	3
2	—	**対応をもつ係り受け関係の数**	—	2
3	—	**文節の総数**	—	5
2	—	**係り受け関係の総数**	—	4
1.0	—	**被覆率** ($m = 1.0$)	—	0.56

類似度 $= 1.0 \times 0.56 = 0.56$

図 3: ユーザ質問文とテキスト文の類似度の計算

それぞれ計算し，両者の被覆率の積を 2 文の類似度としている．

被覆率は以下の式によって計算する．ここで，m は係り受けへの重み付けを決める定数 ($m \geq 0$) である．

$$\frac{(\text{対応をもつ文節の数}) + m \times (\text{対応をもつ係り受け関係の数})}{(\text{文節の総数}) + m \times (\text{係り受け関係の総数})}$$

図 3 においては，ユーザ質問文，テキスト文ともに 3 つの文節と 2 つの係り受け関係が対応をもっている．$m = 1.0$ としたとき，両者の類似度は 0.56 となる．

各テキスト中でもっとも類似度の大きな文をテキストの代表文とし，その類似度をテキストのスコアとする．最後に，システムはテキストのリストをスコアの大きな順に出力する．

3　換喩表現の扱い

本節では，図 1 に示した係り受け関係のずれの問題を解決するため，コーパスから換喩とその解釈を自動的に抽出し，マッチングに応用する方法を提案する．

3.1　換喩表現・換喩解釈表現

以下の 2 種類の表現の組み合わせを扱うことで，図 1 に示した係り受け関係のずれに対処する．

(α) $\boldsymbol{A\ P \rightarrow V}$
(β) $\boldsymbol{A}$ (の) $\rightarrow \boldsymbol{B\ P \rightarrow V}$

ここで，$\boldsymbol{A}$ と $\boldsymbol{B}$ は名詞，$\boldsymbol{V}$ は用言，$\boldsymbol{P}$ は格助詞を表す.「の」は接続助詞であり，その有無は問わない．また $\rightarrow$ は係り受け関係を表す．図 1 の例では，$\boldsymbol{A}$ は「GIF」，$\boldsymbol{B}$ は「画像」，$\boldsymbol{P}$ は格助詞「を」，$\boldsymbol{V}$ は「表示する」に相当する．このとき，(α)「GIF を表示する」は換喩であり，(β)「GIF (の) 画像を表示する」はその解釈になっていると考えることができる．

予備実験としてコーパスから (α) と (β) のペアの抽出を行ったところ，得られたペアの大部分は換喩とその解釈として妥当なものであった．よって，このことを用いて換喩とその解釈の自動的な抽出を試みる．

コーパスとしては，ダイアログナビなどの質問応答システムによって収集されたユーザ質問文と，表 1 のテキスト知識ベースを利用する．特にユーザ質問文は，初心者が入力した質問文が大半であるので，大量の換喩的な表現を含んでいると考えられる．

以下，(α) を**換喩表現**，(β) を**換喩解釈表現**とよぶ．

3.2　換喩表現・換喩解釈表現ペアの抽出

KNP によって構文解析済みのコーパス (ユーザ質問文とテキスト知識ベース) から，以下の方法によって換喩表現と換喩解釈表現のペアを自動的に抽出する．

1. 換喩表現の候補 (C_α) の収集: パターン「$\boldsymbol{A_\alpha\ P_\alpha \rightarrow V_\alpha}$」にマッチする表現をすべて集める．ただしコーパス中の出現頻度 (f_α 回) が閾値 (t_α 回) を下回る表現は除外する ($f_\alpha \geq t_\alpha$)．
2. 換喩解釈表現の候補 (C_β) の収集: パターン「$\boldsymbol{A_\beta}$ (の) $\rightarrow \boldsymbol{B_\beta\ P_\beta \rightarrow V_\beta}$」にマッチする表現をすべて集める．ただし，コーパス中の出現頻度 (f_β 回) が閾値 (t_β 回) を下回る表現は除外する ($f_\beta \geq t_\beta$)．
3. C_α に含まれる各々の表現について，C_β 中に対応する表現，すなわち $\boldsymbol{A_\beta = A_\alpha}$，$\boldsymbol{P_\beta = P_\alpha}$，$\boldsymbol{V_\beta = V_\alpha}$ を満たす表現が存在するとき，それらを換喩表現・換喩解釈表現のペアとして抽出する．

ここで，$\boldsymbol{A_\alpha}$・$\boldsymbol{A_\beta}$・$\boldsymbol{B_\beta}$ は任意の名詞，$\boldsymbol{V_\alpha}$・$\boldsymbol{V_\beta}$ は任意の用言 (サ変名詞のうち,「する」が付属語としてつくものを含む)，$\boldsymbol{P_\alpha}$・$\boldsymbol{P_\beta}$ は任意の格助詞，$\rightarrow$ は係り受け関係を表している．以下では簡単のため，係り受け関係「$\rightarrow$」の記述を省略する．また可読性を高めるため，接続助詞「の」を場合に応じて挿入する．

出現頻度の閾値を設けたのは，構文解析の誤りやおかしなユーザ質問文の悪影響を抑えるためである．以下の実験においては，$t_\alpha = t_\beta = 3$ と定めた．

換喩表現					換喩解釈表現					評
A_α	P_α	V_α	f_α		A_β	B_β	P_β	V_β	f_β	価
エラー	が	出る	1681		エラー	表示	が	出る	68	○
					エラー	報告	が	出る	9	○
					エラー	画面	が	出る	6	○
					エラー	情報	が	出る	4	○
					エラー	メッセージ	が	出る	3	○
					エラー	署名	が	出る	3	○
ファイルが		開く	374		ファイル	検索	が	開く	4	×
電源	を	入れる	290		電源	スイッチ	を	入れる	5	○
印刷	を	実行	141		印刷	プレビュー	を	実行	12	×
					印刷	ジョブ	を	実行	4	○
					印刷	処理	を	実行	4	○
					印刷	コマンド	を	実行	3	○
動作	が	遅い	123		動作	速度	が	遅い	8	○
文字	が	ずれる	97		文字	の 位置	が	ずれる	19	○
					文字	の 間隔	が	ずれる	4	○
					文字	列	が	ずれる	3	○
ファイルが		破損	56		ファイル	の 一部	が	破損	3	○
改行	が	変わる	34		改行	の 幅	が	変わる	3	○
ドメインを		追加	7		ドメイン	ユーザ	を	追加	3	×
アドレスを		開く	4		アドレス	帳	を	開く	43	×
					アドレス	帖	を	開く	3	×
ワード	が	消える	4		ワード	の メニュー	が	消える	4	○
					ワード	の フォント	が	消える	4	×
画面	に	従う	3		画面	の 指示	に	従う	96	○
					画面	の メッセージ	に	従う	3	○

図 4: 抽出された換喩表現・換喩解釈表現ペア

ただし，不適切な換喩表現・換喩解釈表現が得られることを防止するため，C_α と C_β の収集においては，名詞句の一部になっていたり，括弧や遠い係り受け関係を含む表現を除外する．

以上の方法をコーパス中の 1,351,981 文 (ユーザ質問文 762,353 文，テキスト知識ベース 589,628 文) に適用した結果，847 個の換喩表現に対して 1,126 個の換喩解釈表現が得られた．図 4 にその例を示す (「評価」欄については次節で説明する).「電源を入れる」「電源スイッチを入れる」,「改行が変わる」「改行の幅が変わる」のように，興味深い例が多く得られていた．

3.3 マッチングへの応用

以上の方法によって得られた換喩表現・換喩解釈表現ペアを同義表現辞書に登録することで，図 1 に示した係り受け関係のずれを吸収することができる．例えば,「GIF を表示する」と「GIF の画像を表示する」を同義表現辞書に登録することで，図 1 における両者の文の類似度 (係り受け関係への重みづけ $m = 1.0$ とした場合) は，0.27 から 1.0 に増加する．

4 評価と考察

提案手法の評価として，抽出された換喩表現・換喩解釈表現ペアの妥当性評価と，ダイアログナビのマッチングに応用した際の有効性の評価の 2 種類を行った．

表 2: 換喩表現・換喩解釈表現ペアの評価結果

(換喩解釈表現の評価)

評価	換喩解釈表現数
○	222 (84%)
×	42 (16%)
合計	264 (100%)

(換喩表現ごとの評価)

評価	換喩表現数
すべて○	160 (80%)
○×混在	4 (2%)
すべて×	36 (18%)
合計	200 (100%)

4.1 換喩表現・換喩解釈表現ペアの評価

提案手法によって抽出された換喩表現・換喩解釈表現ペアをランダムに選び，妥当な解釈がなされているかどうかという観点で評価を行った．

まず，得られた 847 個の換喩表現から 200 個をランダムに選択し，それらに対応する 264 個の換喩解釈表現が換喩の解釈として妥当かどうかを，○ (妥当)，× (誤り) のいずれかで判断した (図 4 の「評価」欄).

この評価の結果を表 2 左に示す．また，上述の 200 個の換喩表現について，対応する換喩解釈表現が「すべて○」「○×が混在」「すべて×」のいずれかを調べた結果を表 2 右に示す．

これらの結果からわかるように，換喩表現・換喩解釈表現ペア単位でみた場合には 8 割以上のペアは妥当なものであった．また，換喩表現のうち 8 割は対応する換喩解釈表現がすべて妥当なものであった．

4.2 マッチングにおける有効性の評価

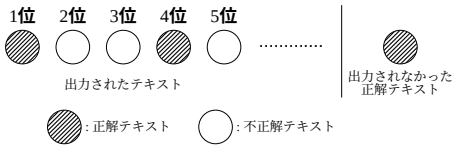
提案手法をダイアログナビに応用した際の有効性を調べるため，テストセットを用いた評価を行った．

テストセットとしては，ダイアログナビの質問文ログから無作為抽出された各々のユーザ質問文に評価者が正解テキストを付与したもののうち，提案手法によって得られた換喩表現・換喩解釈表現のいずれかを含むものを用いた．内訳は，ヘルプ集のテキストを正解とする 31 ユーザ質問文と，サポート技術情報のテキストを正解とする 140 ユーザ質問文である．

各々のユーザ質問文に対するシステムの出力 (スコアによって順序づけされたテキストのリスト) の評価尺度としては，以下の式で定義される ϵ を用いる．

$$\epsilon = \frac{\sum_{i \in \mathcal{R}} \frac{1}{i}}{\sum_{j \in \{1,\cdots,n\}} \frac{1}{j}}$$

ここで，n は入力されたユーザ質問文に対する正解テキスト数，$\mathcal{R}$ は出力されたテキストのリストのうちの正解テキストの順位番号の集合である．ϵ の計算の



$$\epsilon = \frac{1/1 + 1/4}{1/1 + 1/2 + 1/3} = 0.68$$

図 5: ϵ の計算

一例を図 5 に示す．ϵ は，正解テキストがすべて最上位に順位付けされたときに，最大値 1 をとる．

以下の 2 種類の条件で実験を行った．

ベースライン 2 節で述べた方法でユーザ質問文とテキストのマッチングを行う．

提案手法 2 節で述べた方法でユーザ質問文とテキストのマッチングを行う．この際，提案手法によって抽出された換喩表現・換喩解釈表現ペアを同義表現辞書に登録しておく．

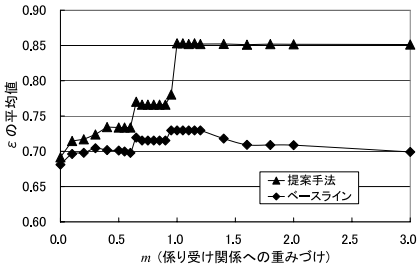
実験の際は，係り受けへの重み付け定数 m を 0 から 3.0 までの範囲で変化させ，それぞれの場合において ϵ の全ユーザ質問文での平均値を計算した．

実験結果を図 6 に示す．この結果から，提案手法を導入することによってシステムの出力が有意に改善されることがわかる．また，図 6 は係り受け関係への重みづけが有用であることを示している．

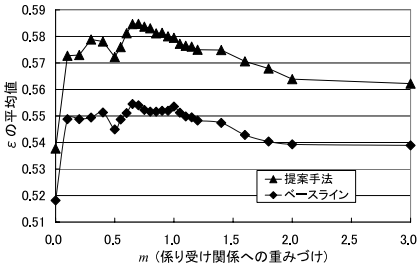
ϵ の改善に有効であった換喩表現・換喩解釈表現ペアの例を図 7 の I1〜I4 に示す．I1 の例では，換喩表現「LAN で接続」がユーザ質問文に含まれ，対応する換喩解釈表現「LAN 経由で接続」が正解テキストに含まれていたため，正解テキストのスコアがその他の不正解テキストのスコアを上回って結果が改善した．逆に ϵ を悪化させた換喩表現・換喩解釈表現ペア (図 7 の W1，W2) は，いずれも換喩の解釈として不適切なものであった．

5 まとめ

本論文では，大量のコーパスから換喩表現・換喩解釈表現ペアを自動的に抽出する方法と，それを質問応答システム「ダイアログナビ」におけるユーザ質問文とテキストのマッチングに応用する方法を提案し，それらの有用性を示した．コーパスとしては，質問応答システムの公開運用によって得られた大量のユーザ質問文と，マイクロソフトが保有するテキスト知識ベースを利用した．質問応答システムをひきつづき運用してさらに大量のユーザ質問文を蓄積することによって，



(ヘルプ集: 31 ユーザ質問文)



(サポート技術情報: 140 ユーザ質問文)

図 6: ダイアログナビのテストセットによる評価結果

	換喩表現	換喩解釈表現
	A_α P_α V_α	A_β B_β P_β V_β
I1	[**ユーザ質問文**]	[**ヘルプ集**]
	LAN で 接続	LAN 経由で 接続
I2	[**ユーザ質問文**]	[**ヘルプ集**]
	ファイル に 関連づける	ファイル の 種類に 関連づける
I3	[**ユーザ質問文**]	[**サポート 技術情報**]
	HTML で 保存	HTML 形式で 保存
I4	[**サポート 技術情報**]	[**ユーザ質問文**]
	アプリケーションが 遅い	アプリケーションの 起動が 遅い
W1	[**ユーザ質問文**]	[**サポート 技術情報**]
	ページ を 表示	ページ の 番号を 表示
W2	[**サポート 技術情報**]	[**ユーザ質問文**]
	ファイル を 印刷	ファイル 一覧を 印刷

図 7: ϵ が変化した換喩表現・換喩解釈表現ペア

さらに多くの換喩表現・換喩解釈表現ペアが得られることが期待できる．

ただし，不適切な換喩の解釈を抽出してしまい，システムの出力を悪化させてしまう例もあった．また，提案手法は多様な換喩現象のごく一部しか扱っていない．これらの問題に対処するためには，換喩を扱うためのより一般的なモデルを考える必要がある．

参考文献

[1] George Lakoff and Mark Johnson. *Metaphors we live by.* University of Chicago Press, 1980.

[2] 清田陽司, 黒橋禎夫, 木戸冬子．大規模テキスト知識ベースに基づく自動質問応答 —ダイアログナビ—．自然言語処理, Vol. 10, No. 4, pp. 145–175, 2003.