ユーザ嗜好情報を用いたWikipediaからの対話に有効な文の抽出

森下崇弘　米田崇明　篠崎隆宏　堀内靖雄　黒岩眞吾
千葉大学大学院　融合科学研究科

1. はじめに

近年，対話システムに関する研究はその発展に伴い，予約受付や道案内など，特定のタスク達成を目的としたタスク指向型対話システムから，達成すべきタスクを持たず，雑談などを扱う非タスク指向型対話システムへと広がりつつある[1].

非タスク指向型対話システムは，雑談などの対話をすること自体が目的であるシステムである．その為，システムはユーザを飽きさせないように興味を惹く話題を提供し，対話を発展させていくことが求められる．非タスク指向型対話システムとしては，これまでにWeb情報の中から意外性のある内容の文を抽出しそれを応答文生成に利用したものが試みられている[2][3]．しかし，ユーザがどのような情報に興味を惹かれるかは，その内容の意外性のみではなく，ユーザそれぞれの嗜好の違いにより異なることが予想できる．その為，対話システムがユーザの興味を喚起する応答を実現する為には，Webからユーザの嗜好に合った意外性のある情報を提供できることが望ましい．

そこで本稿では，ユーザの嗜好情報を利用した意外性のある文抽出手法を提案する．Web上の情報源としてはWikipedia [4]を利用する．Wikipediaを用いるのは，記事の種類が幅広く多くの固有名詞を取り上げられていることや，日々発生する新しい単語についても随時追加されている利点がある為である．

2. ユーザ嗜好情報の利用

2.1 ユーザ嗜好情報の獲得

本稿では，mixi[5]やfacebook[6]等のSNSのプロフィール登録と同等の方法でユーザの情報を収集した．収集したユーザの情報は，ユーザの基本情報（誕生日，出身地，現住地，職業，所属），ユーザの趣味情報（18個の選択肢からの選択式），ユーザの興味・関心情報（10個の単語，自由入力）である．ユーザの情報をこのような形式で収集した理由は，1)多くの人が慣れ親しんだプロフィール登録形式に似せることで，ユーザ情報登録におけるユーザの負担を軽減できる，2)SNSのプロフィールから得られる情報だけを利用することで，将来的にSNSのプロフィール情報をそのまま利用することが可能となる，の2点である．

表1：趣味の選択肢の用語集合の例

選択肢	用語数	用語集合内の用語例
映画鑑賞	1423	ラブコメディ，時代劇
スポーツ	1428	サッカー日本代表，体操
ゲーム	1350	Wii，ゲームデザイナー
インターネット	1166	ブログ，ネット配信

2.2 ユーザ嗜好情報の用語拡張

ユーザから収集した嗜好に関する情報（ユーザの趣味情報と興味・関心情報）は，Alagine[7]により提供されている文脈類似語データベースとWikipediaのカテゴリ構造をシソーラスとして用いた拡張を行った．

ユーザの趣味情報は，選択肢と同意のWikipediaカテゴリ名内の用語と，下位カテゴリ内の用語を収集し，趣味$\hbar$ごとに用語集合$W_{\hbar}$を作成した．趣味の選択肢の用語集合の例を表1に示す．文抽出時には，ユーザu_iが選択した趣味$\hbar$の用語集合$W_{\hbar}$の集合を，ユーザの趣味情報の用語集合$W_H(u_i)$として利用した．

また，ユーザu_iの興味・関心情報は，ユーザから獲得した10個の興味・関心のある用語$W_{I1}(u_i)$の他に，その用語を元に文脈類似語データベースを用いて，興味・関心用語の類似語集合を収集し，それらの用語集合$W_{I2}(u_i)$を，スコア付けに利用した．

3. Wikipedia からの文の抽出手法

3.1 文の抽出処理の流れ

対話中に出現した単語から，対話システムがユーザの興味を惹く応答を返し，対話を発展させることを想定してWikipediaからの文抽出手システムを開発した．図1に文抽出の概要を示す．システムによる文抽出ではまず，入力された単語（以下，見出し語）から，その単語に該当するWikipediaの記事文書を取り出す．このWikipedia記事は前処理が行われており，タグは削除されテキストのみになっている．次に，そのWikipedia記事の各文にスコア付け

を行う．各文には 4 つの指標によるスコアを組み合わせて総合的なスコア付けを行う．その 4 つの指標とは①ユーザ嗜好情報，②TF-IDF，③語の共起頻度，④文長である．スコアが高い文ほど，ユーザの興味を惹く応答文として利用できると考え，システムはスコアの最も高い 1 文を出力する．

3.2 前処理

本稿ではシステムで使用する Wikipedia 記事を，Wikipedia のダウンロードページ[8]から XML 形式で取得し利用した．本稿で利用したデータは，日本語版 2011 年 9 月 17 日時点のものである．記事には，全て XML のタグを取り除きテキストだけを抽出する前処理を行った．

3.3 文のスコア付け

4 つの指標を用いたスコアを組み合わせて，見出し語の Wikipedia 記事中の各文にスコア付けを行う．4 つの指標によるスコアを標準化した後線型結合したものを，ユーザ u_i の文 S_j に対するスコア $SS(u_i,S_j)$ とし，式(1)のように定義した．

$$SS(u_i,S_j)=R(u,S_j)+TI(S_j)+C(S_j)+L(S_j) \quad (1)$$

ここで $R(u_i,S_j)$ はユーザ嗜好情報によるスコア，$TI(S_j)$ は TF-IDF によるスコア，$C(S_j)$ は語の共起頻度によるスコア，$L(S_j)$ は文長によるスコアである．提案手法の文抽出システムは，見出し語の Wikipedia の記事内の全ての文に対し $SS(u_i,S_j)$ を計算し，スコアの最も高い 1 文を出力する．

3.4 ユーザ嗜好情報によるスコア付け

ユーザ嗜好情報によるスコア $R(u_i,S_j)$ を式(2)に示す．

$$R(u_i,S_j)=U_F(u_i,S_j)+U_H(u_i,S_j)+U_I(u_i,S_j) \quad (2)$$

ここで $U_F(u_i,S_j)$，$U_H(u_i,S_j)$，$U_I(u_i,S_j)$ は，ユーザ u_i の文 S_j に対するユーザ基本情報によるスコア，ユーザ趣味情報によるスコア，ユーザ興味・関心情報によるスコアで，以下の節で説明する．

3.4.1 ユーザの基本情報によるスコア

ユーザ u_i の基本情報による文 S_j に対するスコアを $U_F(u_i,S_j)$ とし，式(3)のように定義した．

$$U_F(u_i,S_j)=\sum_{w_k\in S_j\cap W_F(u_i)} V_F(u_i,w_k) \quad (3)$$

ここで $W_F(u_i)$ はユーザ u_i の基本情報の用語集合である．また $V_F(u_i,w_k)$ は，ユーザ u_i の基本情報の用語集合内の用語 w_k に対するスコアで，予備実験により式(4)のように定義した．

$$V_F(u_i,w_k)=1.0 \quad (4)$$

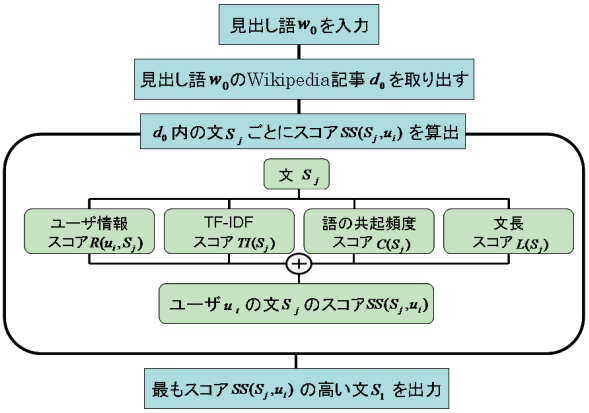

図 1：文抽出の概要

3.4.2 ユーザの趣味情報によるスコア

ユーザ u_i の趣味情報による文 S_j に対するスコアを $U_H(u_i,S_j)$ とし，式(5)のように定義した．

$$U_H(u_i,S_j)=\sum_{w_k\in S_j\cap W_H(u_i)} V_H(u_i,w_k) \quad (5)$$

ここで $W_H(u_i)$ はユーザ u_i の趣味用語の集合である．また $V_H(u_i,w_k)$ は，ユーザ u_i の趣味用語の集合内の用語 w_k に対するスコアで，予備実験により式(6)のように定義した．

$$V_H(u_i,w_k)=1000/N_i \quad (6)$$

式(6)における N_i は，ユーザ u_i の趣味用語の集合内の用語 w_k の総数である．スコアに N_i を用いた理由は，趣味の選択が多いユーザと，少ないユーザでは趣味用語 1 つに対する重要度が異なると考えたからである．後述する 4. の実験において N_i は $2477\le N_i\le 9828$，平均 5566 の値となった．

3.4.3 ユーザの興味・関心情報によるスコア

ユーザ u_i の興味・関心情報による文 S_j に対するスコアを $U_I(u_i,S_j)$ とし，式(7)のように定義した．

$$U_I(u_i,S_j)=\sum_{w_k\in S_j\cap W_{I1}(u_i)} V_{I1}(u_i,w_k)+\sum_{w_k\in S_j\cap W_{I2}(u_i)} V_{I2}(u_i,w_k) \quad (7)$$

ここで $W_{I1}(u_i)$ はユーザ u_i の興味・関心用語の集合である．また $W_{I2}(u_i)$ はユーザ u_i の興味・関心用語の類似用語の集合である．また $V_{I1}(u_i,w_k)$ は，ユーザ u_i の興味・関心用語 w_k に対するスコアで，予備実験により式(8)のように定義した．

$$V_{I1}(u_i,w_k)=1.0 \quad (8)$$

また興味・関心用語の類似用語によるスコア $V_{I2}(u_i,w_k)$ には，Alagine[6]により提供されている類似語データベースに付随する類似度を標準化して利用した．類似度に関する詳細は論文[9]を参照されたい．

3.5 TF-IDF を用いた文のスコア付け

TF-IDF は，重要語に注目した重要文抽出などに用いられている[2]．本研究においても，Wikipedia の記事中の文から重要な文の評価を高くするために，TF-IDF を用いる．本稿では文書 d_t における単語 w_k の TF-IDF 値を $tfidf(w_k, d_t)$ とした．

本研究では，Wikipedia の記事中の文に形態素解析を行い，出現する「名詞」を対象に $tfidf(w_k, d_t)$ を求めた．また，Wikipedia 全記事（726558 記事）のうち，出現する名詞全ての $idf(w_k)$ を算出するのは，膨大な時間がかかってしまう．そのため，まず各テキストに出現する名詞に対して $tf(w_k, d_t)$ のみを算出した．そして $tf(w_k, d_t)$ が設けた閾値 (0.001) より大きかった名詞に対してのみ $idf(w_k)$ を算出し，$tfidf(w_k, d_t)$ を求めた．文 S_j における TF-IDF によるスコアを式(8)に示す．

$$TI(S_j) = \sum_{w_k \in S_j} tfidf(w_k, d_0) \tag{8}$$

このときの文書 d_0 は，文抽出の際の見出し語の Wikipedia 記事を指す．

3.6 語の共起頻度を用いた文のスコア付け

共起頻度によるスコアとして Dice 係数[10]を用いた．Dice 係数は，2 つの単語 w_1，w_2 に対し，$C(w_1, w_2)$ を w_1，w_2 の共起頻度，$C(w_1)$，$C(w_2)$ を w_1，w_2 の出現頻度とした場合，式(9)のように表せる．

$$Dice(w_1, w_2) = \frac{C(w_1, w_2)}{C(w_1) + C(w_2)} \tag{9}$$

文抽出の際の見出し語と，Dice 係数による共起スコアが高い単語によって構成される文は，人間にとって陳腐で平凡な内容であると考えられる．そこで本手法は，意外性という観点から，見出し語との Dice 係数が高い単語を含む文のスコアが低くなるようにスコアの式を定義した．文 S_j における Dice 係数によるスコアを式(10)に示す．

$$C(S_j) = -\sum_{w_k \in S_j} Dice(w_0, w_k) \tag{10}$$

ここでの w_0 は，文抽出の際の見出し語である．

3.7 文長を用いた文のスコア付け

Wikipedia には説明的な文が多く，中には非常に長い文が含まれている．対話において長すぎる文は理解に時間がかかり対話の流れを悪くすると考えられる為，応答文は短い方が望ましい．そこで本研究では短い文に対して高いスコアを付与した．文 S_j における文長によるスコアを式(11)に示す．

$$L(S_j) = -\frac{Len(S_j) - Len\_Mean_0}{Len\_Mean_0} \tag{11}$$

表 2：実験 II における手法と使用したスコア

手法名	使用したスコア
従来手法 1	TF-IDF，文長
従来手法 2	TF-IDF，文長，語の共起頻度
提案手法	TF-IDF，文長，語の共起頻度，ユーザ嗜好情報

ここでの $Len\ (S_j)$ は文 S_j の文長であり，Len_Mean_0 は，文抽出の際の見出し語の w_0 の Wikipedia 記事 d_0 中の文の平均文長である．

4. 評価実験

4.1 比較手法

本稿で提案したユーザ嗜好情報を用いた手法により抽出した文が，対話に有効であることを確認するため，文書情報のみを用いた従来手法との比較実験を行った．比較した 3 つの手法と用いたスコアを表 2 に示す．従来手法 1，従来手法 2 は先行研究[2][3]など，意外性のある情報の抽出に関する研究で用いられている文書情報によるスコアを用いた手法である．提案手法は，それらのスコアにユーザ嗜好情報によるスコアを加えたものである．

4.2 実験方法

図 2 に示す実験 I，実験 II を行った．図で網かけした部分が被験者の作業である．被験者は大学生，大学院生の 20 名（男性 10 名，女性 10 名）である．実験に用いた評価用の見出し語は，Wikipedia の月間アクセス数上位 1000 語から選出した 50 語を使用した．

実験 I では，被験者 1 人あたり 5 語の見出し語に対して，抽出元である Wikipedia 記事全文から，「興味を惹かれる文」を 1 文以上選択してもらった．そして各手法で抽出された文が，その選択されたいずれかの文である率を計算した．Wikipedia 記事の全文は見出し語ごとに紙に印刷し被験者に提示した．その際，1 文単位で評価してもらうため，文の順番はランダムに提示した．評価で用いた見出し語の Wikipedia 記事の平均文数は 148 文だった．

実験 II では，被験者 1 人あたり 25 語の見出し語に対して，主観評価，比較評価を行なってもらった．主観評価では，各手法による抽出文を読み，6 段階（5 : とても興味を惹かれる，4 : 興味を惹かれる，3 : どちらともいえない，2 : 興味を惹かれない，1 : 全く興味を惹かれない，0 : 文の内容がわからない）でそれぞれの文を評価してもらった．そして比較評価では，それら 3 つの抽出文の中から最も興味を惹かれた文を 1 文だけ選択してもらった．なお，比較評価において，2 つの手法の抽出文として同じ文が抽出された場合，両手法とも選択されたとして数を数え

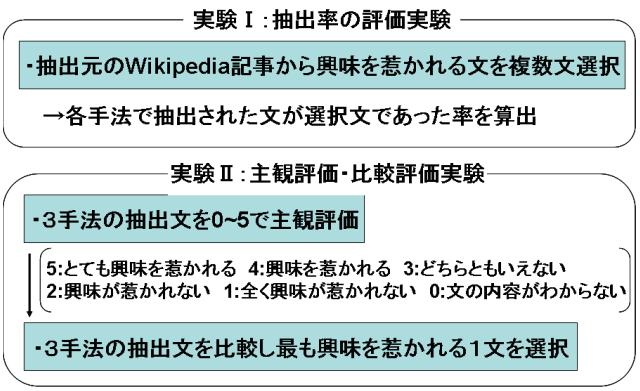

図 2：実験Ⅰ・Ⅱの概要

た．なお，3 つの手法ですべて同じ文を抽出することはなかった．

4.3 実験結果

表 3 に，実験Ⅰにおいて，各手法で抽出された文が，被験者が選択したいずれかの文であった率を示す．表 3 より，提案手法は，選択文を 48%の率で抽出できており，2 つの従来手法と比べ高い率で抽出できたことがわかる．また表 4 に，実験Ⅱにおける，主観評価の全被験者の平均評価値と，主観評価の 6 段階評価において最も高いスコアである「5：とても興味を惹かれる文」と評価された文の数，そして比較評価で 3 つの抽出文のうち「最も興味を惹かれる文」に選ばれた数を示す．表 4 より，提案手法の平均評価値が最も高い 3.36 となっていることがわかる．さらに，最高評価値を獲得した文の数も最も多くなった．また，比較評価においても，3 手法の中で最も多く「最も興味を惹かれる文」として選択されていることがわかる．

4.4 考察

全ての実験で，ユーザ嗜好情報を用いた提案手法は，文書情報のみを用いた 2 つの従来手法より高い精度と評価を得た．このことにより，ユーザ嗜好情報を用いた本手法は，ユーザの興味を惹く文の抽出に有効であることがわかった．

一方で，実験Ⅰにおいて提案手法は最も高い抽出率を得られたものの，その値は 48%に留まった．実験Ⅱの主観評価における提案手法のスコア値の分布を調査したところ，評価 5（とても興味を惹かれる）が 19.8%，評価 4（興味を惹かれる）が 34.8%，評価 3（どちらともいえない）が 21.2%，評価 2（興味を惹かれない）が 12.0%，評価 1（全く興味を惹かれない）が 7.4%，評価 0（文の内容がわからない）が 4.8%となった．このことから興味を惹かれる（評価 5，4）と評価された率は全体の 54.6%で，実際に抽出された文に対する評価も，50%程度であることがわかった．

また，実験Ⅰにおいて，被験者が興味を惹かれる文として選択したものの中には，「血液型」や「学歴」，「性別」に関する内容など，本手法で獲得した

表 3：実験Ⅰの結果

	提案手法	従来手法 2	従来手法 1
抽出率	48%	26%	30%

表 4：実験Ⅱの結果

	提案手法	従来手法 2	従来手法 1
主観評価の平均値	3.33	2.99	2.89
評価 5 の獲得数	99	48	28
比較評価の選択数	254	141	130

ユーザの情報を用いただけでは抽出でき得ない文があることがわかった．このことから，本手法では利用しなかったユーザの情報の利用を検討することが今後の課題であると考える．

5. まとめと今後の課題

本稿では，ユーザ嗜好情報を用いて，非タスク指向型対話システムの対話に有効な文を，Wikipedia から抽出する手法を提案した．

本稿では，ユーザの興味を惹く文が，対話の発展において有効な文であるという仮定の元，ユーザ嗜好情報を用いた文抽出手法を提案し，実験の結果，全ての実験で従来手法と比べ提案手法が高い評価と精度を得た．これらの結果から，ユーザ嗜好情報の利用は，ユーザの興味を惹く文の抽出をする上で有効であることがわかった．

参考文献

[1] 稲葉，平井，鳥海，石井，“非タスク指向型対話エージェントのためのランキング学習を用いた統計的応答手法”，人工知能学会第 63 回言語・音声理解と対話処理研究会，SIG-SLUD-B102-05，2011.
[2] 太田，鳥海，石井，“発話生成を目的とした Wikipedia からの文抽出”，人工知能学会第 23 回全国大会，2G1-NFC5-11，2009.
[3] 村岡，楠村，水口，久寿居，“情報推薦のための意外性判定方式の提案と評価”，情報処理学会第 73 回全国大会，5B-6-1-527，2011
[4] Wikipedia，http://ja.wikipedia.org/
[5] mixi，http://mixi.jp/
[6] facebook，http://www.facebook.com/
[7] ALAGIN Forum，http://alaginrc.nict.go.jp/
[8] Wikipedia ダウンロードページ，http://download.wikipedia.org/jawiki/
[9] 風間，Saeger，鳥澤，村田，“係り受けの確率的クラスタリングを用いた大規模類似語リストの作成”，言語処理学会第 15 回年次大会 2009.3.
[10] 松本，白井，“質問の曖昧性を検出し複数の解答を提示する質問応答システム”，言語処理学会第 12 回年次大会，pp. 935–938, 2006.