日常会話に暗示される主題の語間関係を利用した明示化

片桐 望　荒木 修
東京理科大学理学部応用物理学科

1 はじめに

自然言語による対話を行う人工知能は、接客や娯楽など様々な需要がある。しかし、頭脳の構築を基盤とした言語思考の実現には可能性爆発[1]やルール学習問題[2][3]など数多くの難題が存在する。

現在インターネット上などで実用化されているチャット型対話システム[4][5]は、この問題の解決を回避し自然言語による対話の実現を最優先として開発されている。これらの対話システムはキーワード検索によって辞書から選んだ返答文を、乱数で決定し対話を形成している。そのため人間性・偶然性の高い対話を擬似的に表現できる反面、乱数の影響が強すぎると文脈を無視し、返答内容が不適切となる傾向がある。

文脈判断には形態素解析や意味解析を用いる方法などがあるが、これらは複雑で多くの計算処理を伴う。返答に高い確実性が要求される接客対話などには有効であるが、日常会話は情報伝達よりも会話行為そのものが目的であり、相手がどのような話題を持ち掛けているかさえ把握できれば、ある程度的確な返答が可能と考えられる。複雑な文法解析を行わないことで、思考時間を短縮した迅速な応対も期待できる。

本論文では汎用的な雑談対話システムを想定し、構文解析などの高度な文脈処理を行わない簡素な主題検索アルゴリズムを提案する。ここで言う「主題」とは「文章の内容を簡潔に表現できる単語」とする。この場合、基本的に会話中に登場した単語のいずれかが主題となる場合が多いが、会話中に存在しない単語が主題となる場合も考えられる。これを我々は「日常会話に暗示される主題」と呼ぶ。これを導出する方法として、我々は語間関係網[6]の概念を用いる主題検索アルゴリズムを考案した。

2 主題検索アルゴリズム

主題を指す単語は、文との意味関係が強いと考えられる。文は単語の集合体であるので、主題を指す単語は文を構成する各単語との強い関係があると仮定した。従って、単語間の関係をどのように定義するかが重要な鍵となる。単語間の関係を表現する方法としてKarov & Edelmanの単語と文の間の相互依存性を用いるアルゴリズムがある[6]。これは、類似した単語は類似した文に現れ、類似した文は類似した単語で構成される、というものである。橋本はこれを基本とした動的な語間関係の構造、語間関係網を定義している[7]。

本論文ではこれらの概念を元に、語間関係の強弱を①文中で同時に使用された二単語は関係を強める（形容・主述の関係） ②一方の単語が第三の単語と多用されている場合、二単語は関係を弱める（語彙の多様性） ③文中の出現パターンが似た二単語は関係を強める（類義語の関係）――の三点から判定する。以下具体的に説明する。

単語w_iとw_jの対語親密度$I(w_i,w_j)$を式(1)のように定義する。

$$I(w_i,w_j)=\frac{P(w_i \text{ and } w_j)}{P(w_i \text{ or } w_j)} \quad (1)$$

単語wを含む文の出現確率を$P(w)$とする。分子は単語w_i およびw_jを共に含む文の出現確率であり、条件①を簡単に数式化したことに相当する。分母は単語w_iまたはw_jを含む文の出現確率であり、条件②に相当する。単語自身との対語親密度は最大値1をとり、逆に一度も同一文で使用されていない二単語の対語親密度は最小値0をとる。

このアルゴリズムでは、単語wを語間関係網に新規登録する際に単語のおおまかな分類c_wを指定し、記録する方法を用いる[5]。この登録情報を用いて、条件③に相当する単語w_iとw_jの対語類義度$S(w_i,w_j)$を式(2)のように定義する。

$$S(w_i,w_j)=\frac{P(c_{wi} \text{ and } c_{wj})}{P(c_{wi} \text{ or } c_{wj})} \quad (2)$$

分類 c_w の単語 w を含む文の出現確率を $P(c_w)$とする。同じ分類の二単語（類義語）の対語類義度は最大値 1 をとり、一度も同一文で使用されていない異分類に属する二単語の対語類義度は最小値 0 をとる。

以上の関係式を用いて、単語w_iとw_jの語間関係値$R(w_i,w_j)$を式(3)のように定義する。

$$R(w_i,w_j)=\alpha\cdot I(w_i,w_j)+(1-\alpha)\cdot S(w_i,w_j) \quad (3)$$

対語親密度と対語類義度との係数 α による線形結合で表される。単語w_iの語彙は、登録された全ての単語との語間関係値によるベクトルで表現できる。

単語wにおける文中の全ての単語（N_{wj}語）との語間関係値の平均値を対文親密度$A(w)$と定義する。

"A Method for Clarifying the Theme suggested in an Everyday Conversation using the Relation among Words"
Nozomu Katagiri, Osamu Araki: Department of Applied Physics, Faculty of Science, Tokyo University of Science

$$A(w) = \frac{\sum_j R(w, w_j)}{N_{wj}} \qquad (4)$$

より高い対文親密度を持つ単語が文の内容により近い意味を持つ言葉、つまり文の主題候補である。文中に含まれない単語の対文親密度も検索することにより、文中に明示されない主題であっても発見が可能となる。

3 主題検索実験

主題検索アルゴリズムを実装したチャット型対話システムを構築し、新聞記事 33 件 271 文[8]から 1044 語の語間関係を学習させた(図 1)。このシステムに別の記事による 12 文を入力し、各文での主題検索能力を検証した(図 2)。出力された主題候補のうち明示的候補は、学習済みの単語が文中から選出される。これに対し暗示的候補には文中の単語の略称や関連語などが多く選出されているが、対

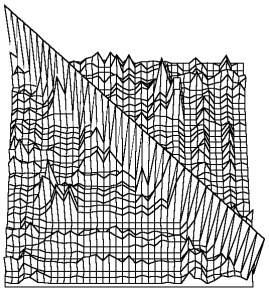

図 1　語間関係値 $R(w_i,w_j)$（一部・α=1.00 の状態のもの）
右および手前方向の軸は単語の index(1-40)

イラン南東部の大地震の救援活動の一環として、
主題候補>>　南東部(468) 救援活動(463) 一環(454) 大地震(424) イラン(377)
南東(386) ケルマン州(321) 南(298) 地震(295) 救援(286)
援助物資を積んだ航空自衛隊のＣ１３０輸送機２機が１日と２日にそれぞれ、
主題候補>>　援助物資(351) 航空自衛隊(329) １日(324) Ｃ１３０輸送機(322)
Ｃ１３０(295) 輸送機(251) 航空(204) 空自(200) 物資(194)
ケルマン空港に到着した。
主題候補>>　ケルマン空港(633) 到着(633)
ケルマン(317) シンガポール(286) 北西(276) 空港(274) 被災地(249)
物資は毛布、テント、ビニールシートなど約１０トン、
主題候補>>　毛布(563) テント(540) ビニールシート(501) 物資(436)
医薬品(246) 援助物資(234) 救援物資(225) 燃料(208) エンジン(208)
１０００万円相当で、
主題候補>>　相当(1000)
主要(271) 緊急(199) 強制的(199) 完全(199) 平和(199)
イラン赤新月社に渡された。
主題候補>>　イラン(1000)
地震(355) 南東部(338) 南東(332) ケルマン(327) 南(313)
赤新月社のケルマン空港責任者メフディ・ムラビ氏は
主題候補>>　ケルマン空港(564) 責任(564)
ケルマン(268) 空港(268) 重さ(233) 北西(199) 被災地(186)
「大変感謝している。出来る限り素早くバムの被災者に届けたい」と話した。
主題候補>>　被災者(460) 感謝(460) バム(448)
被災(319) 前進(162) 被災民(149) 人道援助(149) チャーター(149)
自衛隊機による物資輸送は、
主題候補>>　物資(470) 輸送(469) 自衛隊機(370)
輸送機(305) 航空(235) Ｃ１３０(225) Ｃ１３０輸送機(210) 救援(182)
国際緊急援助隊派遣法に基づき、
主題候補>>　国際緊急援助隊派遣法(1000)
援助隊(447) 国際緊急援助隊(447) 国際(359) 緊急(321) 一環(299)
外務省の要請を受けて防衛庁が派遣を決めた。
主題候補>>　防衛庁(418) 要請(417) 外務省(384) 派遣(368)
外務(305) 援助隊(238) 国際緊急援助隊(238) 協力(219) 自衛隊(210)
同法に基づき自衛隊が海外派遣を行うのは４回目。
主題候補>>　法(416) 自衛隊(411) 海外派遣(387)
派遣(229) 憲法(195) 海外(186) 自衛隊派遣(165) 航空自衛隊(160)

図 2　新聞記事による主題候補検索実験結果（α=0.80）
出力される単語のうち 1 行目は明示的主題、2 行目は暗示的主題候補。カッコ内の数値は対文親密度 $A(w)$×1000

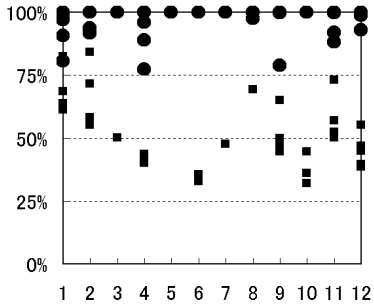

図 3　文の内容と関連の強い候補の対文親密度 $A(w)$最大値比
■が文中には含まれない暗示的な主題候補

文親密度 $A(w)$の値のより低い候補の中には、主観的に文の内容と関連が弱いと思われる単語も見られる。そこで各文で値を比較したところ、主観的に文の内容と関連が強いと思われる主題候補の値は各文の最大値の約 50%以上の範囲に分布していた(図 3)。この分布幅は話題分岐の自由度であると考えられる。閾値の設定により主題候補を絞り込むことは可能だが、明示的主題より比較的低い対文親密度の暗示的主題をどの程度考慮して設定するかが重要である。

4 まとめ

実験結果は日常会話ではなく新聞記事によるものであるが、本論文のアルゴリズムによる主題候補の検索能力は確認できた。最終的にこのアルゴリズムによる主題候補をその対文親密度で加重選択することで、乱数を利用した疑似対話システムであっても、より的確な主題を決定し返答を行うことが可能になる。ただし、語間関係を学習させるために単語の分類を逐次ユーザ入力しなければならない点、単語検索に文字列照合を用いるために語尾変化を伴う動詞・形容詞などが扱えない点など、語間関係網に単語を登録する方法の更なる改良が今後の課題である。

参考文献

[1] M.R.Geneseleth & N.J.Nicolson（古川監訳）(1983) “工知能基礎論” オーム社.
[2] ヴィゴツキー（柴田吉松訳）(1962) “思考と言語” 明治図書.
[3] 市川真一 (1996) “認知心理学４ 思考”, 東京大学出版.
[4] @niftyチャッとも http://chat.nifty.com/chatmo/
[5] Sony Computer Entertainment Inc(1999) “どこでもいっしょ”
[6] Y. Karov & S. Edelman (1998) “Similarity-based word sense disambiguation”, Computational Linguistics, 24, 41-59.
[7] 橋本敬 (1999) “動的言語観に基づいた単語間関係のダイナミクス”, 認知科学, 6[1], pp0-10.
[8] 朝日新聞社アサヒ・コム http://www.asahi.com/