連想及び語の共起情報に基づく発話内容の自動生成*

秋山 直樹, 唐澤 博†
山梨大学大学院‡
E-mail:{akiyama,karasawa}@jewel.yamanashi.ac.jp

1 はじめに

本研究では従来の質問応答型の対話システムから，創発的対話システムの構築に向けて，文生成の観点からアプローチする．

Allenの提唱した計画立案は，システムの意図を起点として，因果関係などの発話間の論理的な関係を連鎖させることによって対話が行われる．実際に，対話を扱う多くの研究では，この手法に基づいた対話管理がなされ，駅や郵便局の窓口業務のように，目的が明確である場合は有効な手段となり得る．

しかし，日常的な対話ではそのような具体的な目的を持つものだけではなく，情報のやりとり自体が目的となるものも多く見られる．そこで本研究では計画立案のような目的指向型の対話を目的とするのではなく，文脈から連想される長期記憶を素材とし，語の共起情報を用いて対話を途切れさせずに発展させることを目的とする．

2 システム概要

本研究のシステム構成は図1で示される．

- **文解析・文生成サーバ**
  本研究室で開発された既存のシステムを利用する．文解析サーバは発話文から意味ネットワークを，文生成サーバは意味ネットワークから発話文をそれぞれ得る．

- **連想構造変換・意味ネットワーク変換**
  連想構造変換は意味ネットワークを連想構造に変換する．意味ネットワーク変換はその逆変換を行う．

- **連想システム WAVE**
  ユーザの発話や興味知識，文脈をキーとし，連想型知識ベースから長期記憶を想起する．

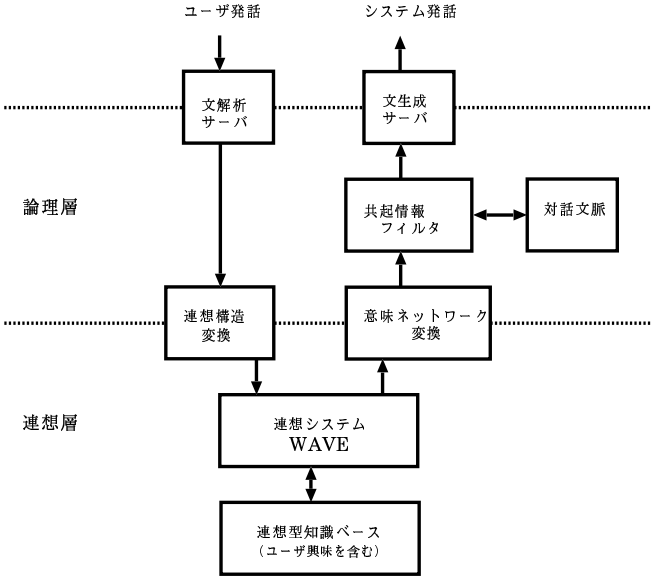

図 1: システム概要

- **共起情報フィルタ**
  連想想起された意味ネットワークと対話文脈間にある共起情報を用いてクラスタリングを行い，システム発話内容を導く．

- **対話文脈**
  1セッション内のユーザとシステムの発話ログが意味ネットワークで表現されている．

- **連想型知識ベース**
  連想型知識ベースはシステムが獲得した知識の概念や事象などを要素とし，それらを相互結合させた連想構造をとる [1].

  意味ネットワークを連想構造化したものを長期記憶として連想型知識ベースに蓄積する．また，対話時にはこの知識ベースのアクセス機構に連想システム WAVE を適用して長期記憶を想起する．

*Automatic utterance generation based on association and word co-occurrence

†Naoki Akiyama,Hiroshi Karasawa
‡University of Yamanashi, 4-3-11 Takeda, Kofu, Yamanashi 400-8511, Japan

3 連想システム WAVE

本研究室で開発された連想システム WAVE は，角田ら [2] によって提案された人間の推論が連想のステップを積み重ねて成り立つという仮説に基づいたアーキテクチャの一部が実装されたものである [3].

3.1 活性値計算

複数のキーから連想型知識ベース内の各要素の活性値計算をする．各要素の入力活性値を I_i とすると，その要素の活性伝搬後の活性値は次式で計算される．

$$O_j = \sum_i W_{ij} I_i \qquad (1)$$

$$(W_{ij} = \frac{1}{(1+n)}, n = \text{要素あたりのリンク数})$$

3.2 コピーバック処理

活性伝搬後，各要素の活性値に 1 未満のコピーバック係数をかけ，活性値を低下させることにより入力の時間差を考慮する．そして，その値を次の入力値に加算することで文脈を生成する．

4 共起情報フィルタ

共起情報フィルタでは，連想想起された意味ネットワークと対話文脈との間にある結束性を考慮したシステム発話内容を生成し，対話の一貫性を保つ．本研究では望月ら [4] の文書中に表れる語間の結束性の強さを共起スコアで表す手法を適用する．

4.1 共起スコア計算

大規模文書コーパス内の語の共起情報から，対象となる意味ネットワーク中に現れる全ての語の共起スコア計算を行う．文書数 N のコーパス内の文書を次元とし，文書ごとの語の出現頻度を要素とする N 次元ベクトルを考える．

任意の語 X と語 Y の共起スコアを，それぞれのベクトル間の類似度によって計算する．類似度の尺度には次式のコサイン距離を用いる．

$$coscr(X,Y) = \frac{\sum_{i=1}^{N} x_i * y_i}{\sqrt{\sum_{i=1}^{N} {x_i}^2} * \sqrt{\sum_{i=1}^{N} {y_i}^2}} \qquad (2)$$

ただし，x_i と y_i は文書 i における語 X と語 Y の出現頻度を表し，N はコーパスの全文書数を表す．

4.2 クラスタリング

共起スコアを基に，対象となる意味ネットワーク中の語のクラスタリングを行う．1 語 1 クラスタから開始し，クラスタ間の類似度が閾値以上である限りクラスタリングを繰り返す．

クラスタ間の類似尺度には次式の最短距離法を用いる．つまり，2 つのクラスタ間の類似度は，各クラスタ内の要素の中で最も共起スコアの高い要素ペアの値とする．

$$sim(C_i, C_j) = max_{X,Y}\ coscr(X \in Ci, Y \in C_j) \quad (3)$$

(X, Y はそれぞれクラスタ C_i 内，C_j 内の語を表す．)

また，連想想起された意味ネットワークと対話文脈との間にクラスタが構成されない場合は，出力が得られるまで閾値を減少させて再構成を繰り返す．

4.3 システム発話内容の決定

クラスタリング結果から，最も領域の大きいクラスタを含む意味ネットワークをシステム発話内容として出力する．これにより対話文脈との結束性が強いシステム発話が可能となり，対話の一貫性が維持できる．また，クラスタリング時の閾値変更により，話題や場面の変化などで対話を途切れさせずに継続できる可能性も高いと考えられる．

5 おわりに

自然言語理解対話システムの文生成に連想処理と語の共起情報処理のハイブリッド機構を導入する手法について述べた．

参考文献

[1] 秋山, 唐澤: 連想に基づく発話内容の自動生成, 情報処理学会第 65 回全国大会講演論文集 (2), pp165-166,2003.

[2] 角田, 田中： PDAI&CD に基づく意味の学習および文脈依存の多義性解消, 電子情報通信学会技術研究報告,Vol.DE93-1,pp1-8,1993.

[3] 藤田, 唐澤: 場面情報に基づく辞書の開発と談話理解への適用, 情報処理学会第 64 回全国大会講演論文集 (2), pp.35-36,2002.

[4] 望月, 奥村, 岩山： Japanese Lexical Chainers < http://www.tufs.ac.jp/ts/personal/motizuki/software/chainers/ >, 参照 2003-11-17.