Vol.2010-ICS-158 No.4
2010/3/1

常識的連想による ニュースヘッドラインからの会話文生成

吉岡孝治[†]　吉村枝里子[†]　土屋誠司[†]　渡部広一[†]

人とロボットの会話を用いたコミュニケーションについて研究している．情報量が膨大で，日常生活に必要不可欠なニュースに着目し，ヘッドラインの会話への変換手法を提案する．ヘッドラインは助詞や用言の欠落，新聞独自のルールが存在する．本手法では語の共起情報や意味情報を考慮する連想メカニズムを利用し，ヘッドラインを知識ベースに整理し会話に変換した．また常識を組み込み，人が抱く感覚・知覚を考慮した豊かな会話を実現した．

Generation of Conversation from Headlines Based on Commonsense Association

Koji Yoshioka[†]　Eriko Yoshimura[†]　Seiji Tsuchiya[†]
and Hirokazu Watabe[†]

We study communication with conversation between human and robots. News has a huge amount of information and essential to daily life. So we propose a method to convert headlines of newspapers article into conversations. There are the lacks of partice and verb, unique rules in Newspapers. In this paper, we use association judgment mechanisms considered co-occurrence information and semantic information of words, organize the knowledge base of headlines. And we convert headlines into conversation. We also incorporate commonsense, realize the rich conversation considered sensation and perception.

1. はじめに

情報化社会の進歩により，将来人のパートナーとして自律的に行動を行うロボットの開発が期待される.そこで問題になるのが人とロボットとのインタフェースである．キーボードやリモコンを利用したロボット操作では操作方法習得の訓練が必要になり，人にとって負担になる．またディスプレイや定型文での音声ガイダンスによるロボットからの応答も人がロボットにインタフェースを合わせるため負担が大きくなる．そこで人同士のコミュニケーション手段である会話をロボットとのインタフェースにも適用することにより，人とロボットの円滑なコミュニケーションに繋がると考えられる．その結果，人とロボットとの距離が近づき優しいロボットの実現に繋がると期待できる．

人同士の会話の中で，人はニュースを話題にした会話を行う．その理由はニュースが日常生活に必要不可欠であり，会話の糸口や話し相手の関心事に関連しているためだと考えられる．そこでロボットがニュースを話題にした会話を行うことで，より親しみやすくなると考えられる．しかし，単純にニュースを機械的に読み上げた場合，コミュニケーションツールとしての会話の柔軟性を損ない,親しみやすさは半減する．そこで本稿ではニュースを手がかりにロボットが会話の発展性を高める会話文の自動生成を目的とする．ニュースのソースを新聞記事とすると，記事のタイトルに相当するニュースヘッドラインがニュース記事独特の文法規則によって記述されるため，ロボットにニュースヘッドラインからの情報を理解させることが難しくなる．ニュースヘッドライン独特の文法規則とは例えばニュースヘッドライン「民主党：新人議員の研修会開始　「小沢イズム」徹底教育？」に対し,「研修会」と「開始」の間の『助詞欠如』,「開始」の後の『語尾欠如』等がある．そのためニュースヘッドラインを利用してロボットが発話する場合，まずニュースヘッドライン独特の文法規則に従い，ニュースヘッドラインを理解し，発話の形式に変換する必要がある．また人は「楽しい」や「悲しい」といったニュースから受けるイメージを会話の中で利用する．このニュースのイメージに基づく会話を生成することで会話の発展性を高めることができると考えられる．ここでニュースヘッドラインの意味理解とそこからの会話文生成は，語の表記情報のみを考慮した文法規則のルール化や単純なテンプレートの組み合わせでは，語の意味情報を考慮できない．そこで概念ベースや関連度計算からなる連想メカニズムを利用することで，語の意味情報を考慮した処理を行う．

本稿では，会話中で話題となるニュースを基にした復元会話と，そのニュースヘッドラインのイメージに基づくテーマ拡張会話を生成する．そのため，表記情報を考慮

[†] 同志社大学大学院工学研究科
Department of Knowledge Engineering and Computer Sciences, Graduate School of Engineering, Doshisha University

した文法規則や会話生成ルールに加え，語の意味情報を考慮した連想メカニズムを適用する．

2. 研究概要

2.1 前提条件

本稿では Web 上より自動的に獲得したニュースヘッドラインを利用する．このテキストデータは Web 上のニュースサイトから獲得できたという前提でシステムを構築する．取り扱うテキストデータは毎日新聞の Web サイト「毎日.jp」[1]のニュースヘッドラインである．

2.2 システム概要

本システムは人とロボット会話中，ロボットからのニュースを話題にした復元会話とテーマ拡張会話の生成を目的としている．例えば，ニュースヘッドライン「新人議員の研修会開始　「小沢イズム」徹底教育」に対し，復元会話では「民主党が新人議員の研修会を開始するんですよ」，テーマ拡張会話では「民主党は話題になっていますね」を生成する．本システムでは，会話文を生成するために①ニュースヘッドラインのテキストデータ獲得，②ニュース記事記述文法を考慮したフレーム分解，③会話文生成処理，④出力の 4 つの部分に分かれている．システムの概略を図 1 に示す．

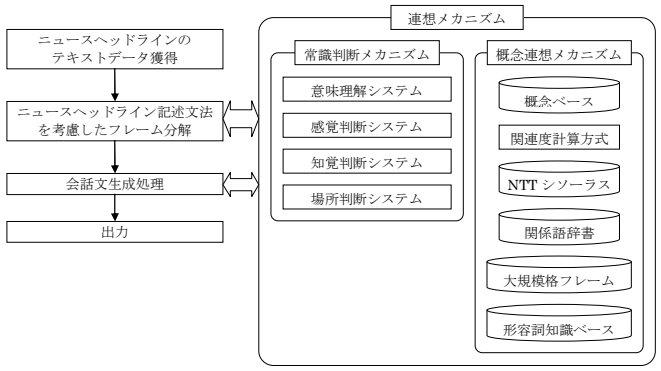

図 1　システムの概略

②ニュース記事記述文法を考慮したフレーム分解では各ニュースヘッドラインを 6W1H（Who，What，When，Where，Whom，Why，How）と用言及びテーマに分解し整理する．語の表記情報であるニュースヘッドライン記述文法の整理に加え，大規模格フレームを利用し語の共起情報，常識判断メカニズム・概念連想メカニズムを利用し語の意味情報である常識を考慮した処理を行う．そして③会話文生成処理ではフレーム分解を利用し，ニュースヘッドラインを会話文形式に復元する復元会話，ニュースヘッドライン中のテーマから人間が想起するイメージを利用した会話を自動生成するテーマ拡張会話を生成する．会話文生成時，テーマと予め用意したテンプレートとの組み合わせを行う．その際，常識判断メカニズム・概念連想メカニズムを利用し，柔軟な組み合わせから会話を生成する．入出力の具体例を図 2 に示す．

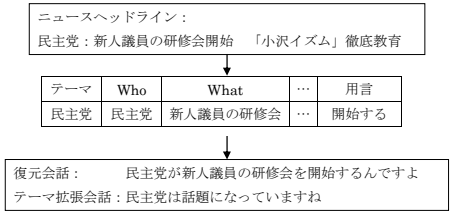

図 2　入出力の具体例

3. 連想メカニズム

3.1 概念ベース

概念ベース[2]とは複数の国語辞書や新聞などから機械的に構築した単語（概念）とその意味特徴を表す単語（属性）の集合からなる知識ベースである．概念にはその重要性を表す重みが付与されている．概念ベースには約 12 万語の概念表記が収録されており，1 つの概念に平均約 30 個の属性が存在する．ある概念 A は属性 a_i とその重み w_i の対の集合として，（式 3.1）で表される．

$$A = \{(a_1, w_1), (a_2, w_2), (a_3, w_3), \cdots, (a_i, w_i), \cdots, (a_n, w_n)\} \quad (3.1)$$

任意の一次属性 a_i は，その概念ベース中の概念表記の集合に含まれている単語で合成されている．したがって，一次属性は必ずある概念表記に一致するため，さらにその一次属性を抽出することができる．これを二次属性と呼ぶ．概念ベースにおいて，「概念」は n 次までの属性の連鎖集合により定義されている．

3.2 関連度計算方式

関連度計算方式[3]とは概念ベースを利用し概念と概念の関連の強さを定量的に評価するものである．関連度の値は 0~1 の連続値をとり，1 に近づくほど関連が強い．

3.3 NTT シソーラス

NTT シソーラス[4]は一般名詞の意味的用法を表す 2710 個のノード上位下位関係，全体部分関係が木構造で示されたものである．ノードに所属する名詞として約 13 万語

のリーフが分類されている．に NTT シソーラスの木構造の一部を図 3 示す．

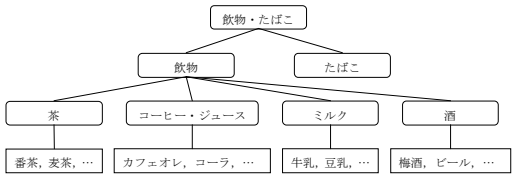

図 3　NTT シソーラスの木構造（一部）

3.4 Web から自動構築した大規模格フレーム

Web から自動構築した大規模格フレーム[5]とは，動詞とその動詞に関係する名詞を用法ごとに整理したものである．この格フレームは，Web 上の約 5 億文の日本語テキストから自動的に構築されている．また，格フレームに含まれる動詞の数は約 5 万語である．この格フレームを用いることで，動詞からその動詞に結びつく名詞，格，頻度を取得できる．また，名詞から動詞を取得することも可能である．格フレームの頻度とは Web 上において，その名詞と動詞が出現した回数を指す．例として，格フレームに名詞「鉛筆」を入力した結果を表 1，動詞「飲む」を入力した結果を表 2 に示す．

表 1　入力「鉛筆」における格フレームの出力

動詞	格	頻度
描く	デ格	371
書く	デ格	327
削る	ヲ格	217

表 2　入力「飲む」における格フレームの出力

名詞	格	頻度
酒	ヲ格	40517
薬	ヲ格	22145
ビール	ヲ格	17105

3.5 形容詞知識ベース

形容詞知識ベースとは現代形容詞用法辞典[6]を参考にし，形容詞の見出し・意味・固有語及びイメージについて整理した知識ベースである．固有語は辞典に掲載されている例文を参考に形容詞と係り受け関係にある名詞を収集した．イメージとは形容詞の持つイメージを 1～7 までの 7 段階に数値化した値で，1 が強いマイナスイメージ，7 が強いプラスイメージを意味する．

3.6 意味理解システム

意味理解システム[7]とは単文入力に対し，6W1H（Who, What, When, Where, Whom, Why, How）と用言に分解し整理するシステムである．意味理解システムの出力例を図 4 に示す．

（Ex）妹が昨日，本屋で母に絵本を買ってもらった
↓

Who	What	When	Where	Whom	Why	How	用言
妹	絵本	昨日	本屋	母			買う

図 4　意味理解システムの出力例

3.7 感情・知覚判断システム

感覚・知覚判断システム[8][9]とは名詞に対し，人間が常識的に連想する感覚や知覚に関する語を取得するシステムである．出力例を表 3 に示す．

表 3　感覚・知覚判断システムの出力例

入力された名詞	感覚から想起された語	知覚から想起された語
林檎	赤い，甘い，丸い	-
怪我	痛い	気の毒な，つらい，痛々しい，憂鬱な

3.8 場所判断システム

場所判断システム[10]とは名詞に対しその語が場所語かどうかを判断し，場所語である場合には，その場所に存在するもの（主体語）とその場所で行う行動（目的語）を連想するシステムである．場所語とは病院・山・オフィス等のように名詞単独で場所を示す語であると連想できる語と定義する．

4. ニュースヘッドライン記述文法

ニュースヘッドラインを会話文に変換する際，ニュースヘッドラインの文を理解することが必要となる．そこで，意味理解システムを利用し 6W1H+用言のフレームに分解する．しかし，ニュースヘッドライン記述には助詞の省略や用言に該当する語尾部分の省略（欠落語）といった独自の文法が採用されているため，文法的に正しい単文の入力を想定している意味理解システムを直接利用することは困難である．そのためまずニュースヘッドラインの記述文法及び毎日新聞特有のルールを抽出し整理する．その整理した文法・ルールをニュースヘッドラインに適用した後，意味理解システムを利用する．

4.1 ニュースヘッドラインの形式

毎日新聞のニュースヘッドラインは特定の記号（「：」,「、」 ,「…」）を利用し記述されている．これらと特定の記号に意味があると考え，特定の記号の使用方法に従い，ニュースヘッドラインを分類する．分類パターン表を表 4 に示す．この分類パターン表には 12 の分類パターンが存在する．また表 4 中の記号構成 A・B・C・D は特定の記号間の文字列を意味しブロックと定義する．

表 4　パターン分類表の一部

パターン No	記号構成	例文
1	A：B　C	民主党：新人議員の研修会開始　「小沢イズム」徹底教育？
2	A：B	大リーグ：フィーリーズがナ・リーグ優勝決定戦へ
3	A：B、C	インフルエンザ：患者数、都市部を中心に倍増
4	A：B…C	中国：ウイグル暴動で 6 人に死刑判決…中級人民法院
5	A：B、C　D	サッカー：日本、トーゴに 5-0 で快勝　キリン C 杯

4.2　形式の分類別出現頻度

表 4 の各パターンは毎日.jp の Web サイトに登場する頻度が異なる．そのため，全体精度を求める際に必要となる各パターンの出現頻度を調査する．対象としたデータは 2009 年 10 月 13 日から 11 月 12 日のニュースヘッドライン 1608 件とした．出現頻度の結果を図 5 に示す．図 5 より出現頻度に偏りがあることがわかる．本稿では，出現頻度の上位 5 つ（全体の 88.1%）であるパターン 1~5 を研究対象のデータとする．

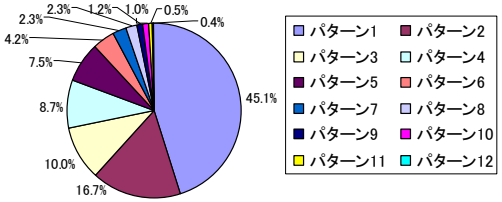

図 5　出現頻度調査結果

5.　ニュースヘッドライン記述文法と形式を考慮したフレーム分解

5.1　ニュースヘッドラインフレーム

ニュースヘッドラインを分解・整理するフレームとして，ニュースヘッドラインフレームを定義する．このフレームは 6W1H+用言と，テーマを加えた計 9 つのフレームが存在する．ニュースヘッドラインフレームへの格納例を図 6 に示す．

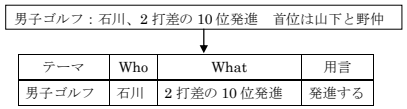

図 6　ニュースヘッドラインフレームへの格納例

5.2　フレーム分解アルゴリズム

主題となる文章取得処理，欠落語の処理後，意味理解システムで処理し，空フレームの補完を行いニュースヘッドラインフレームに格納する．アルゴリズムの全体像と具体例を図 7 に示す．各処理についての詳細は次節より記述する．

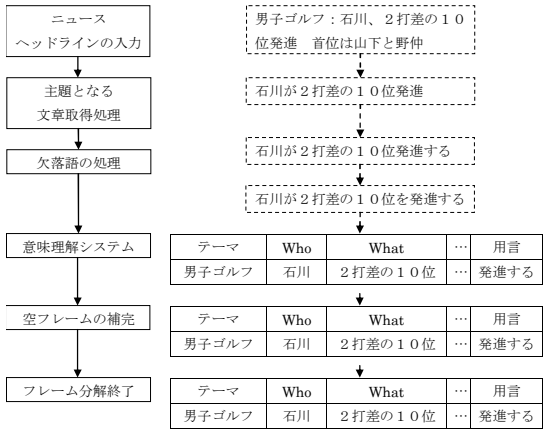

図 7　フレーム分解アルゴリズムと分解例

5.3　主題となる文章取得処理

文は述語が一つだけからなる単文と，複数の述語からなる複文が存在する．ニュースヘッドラインも同様に単文や複文から構成されている．ニュースヘッドラインの特徴として複文の場合，文章の一つが主題を表し，それ以外の文章が補足の文章の役割をしている．パターンの特徴を活かし，ニュースヘッドラインから主題となる文章を切り出し会話生成に利用する．

表 4 を参考に，記号構成の特徴を利用することでニュースヘッドラインを主題となる一文に整理する．ニュースヘッドラインを一文へ整理するための助詞の補完・ブロックの結合のルールを表 5 に作成した．

表 5　パターンの特徴を活かした主題となる一文抽出ルール一覧

条件	ルール
ニュースヘッドラインがパターン 1/2/4 である	ブロック B を主題となる一文と見なす
ニュースヘッドラインがパターン 3/5 である	ブロック B とブロック C の間に助詞「が」を補完し，ブロック B とブロック C を主題となる一文として結合する
全てのパターン	ブロック A をテーマフレームに格納する

5.4　欠落語の処理

5.4.1 助詞の補完

文章は「開始する」や「食べる」のように用言で終わらなければならない．そして

文法的に正しい文章ならば用言の直前に助詞が表れる．しかし，ニュースヘッドラインは助詞を省略するため，用言の直前が目的語（名詞）になる．そのため主題となる一文の末尾が用言であり用言の直前が名詞ならば，助詞の補完が必要になる．

助詞の補完では大規模格フレームを利用し助詞の欠落を補う．例として「研修会開始」に対し助詞を補完する場合，「開始」と直前の「研修会」の間に欠落した助詞を推測する．「開始」から大規模格フレームを利用し，候補となる名詞（候補名詞）と格（候補格）のセットを獲得する（表 6）．その後，頻度順に「研修会」の末尾「会」と候補名詞の関連度を計算する（表 7）．そして閾値以上の候補名詞（運用）がヒットした段階で，候補名詞と対の候補格を助詞の補完に利用する格（ヲ格）として採用する．

表 6　大規模格フレーム「開始」の結果

候補格	候補名詞	頻度
カラ格	時間	26614
ヲ格	サービス	11661
ヲ格	運用	3963

表 7　「会」と候補名詞との関連度

候補名詞	関連度
時間	0.25
サービス	-1
運用	0.33

5.4.2 用言の処理と用言の補完

用言の処理を行うために，主題となる一文の語尾に着目したルールを設定する．そしてそのルールに従い，用言を処理する．主題となる一文の語尾は動詞である場合の他に，語尾が「へ」や「に」や名詞が存在する場合がある．そこで，語尾を動詞化する処理を加え，一文として完結させる．語尾を利用した処理では表 8 に示すルールに従って変換する．

表 8　語尾を利用した処理の一部

語尾情報	語尾の直前	処理	変換例
名詞-サ変接続	-	語尾の後に「する」を追加	大騒ぎ→大騒ぎする
上記以外の名詞	名詞-一般	語尾の後に「です」を追加	難色→難色です
助詞「へ」	名詞-サ変接続	「へ」を「する」に変換	交渉へ→交渉する
助詞「へ」	名詞-一般（動詞化可能な場合）	語尾直前の「イ行」を「ウ行」に変換	見送りへ→見送る
助詞「へ」	名詞-一般（動詞化不可能な場合）	用言の補完	市場へ→市場へ行く

助詞の補完と同様に，主題となる一文に対し大規模格フレームを利用し用言の欠落を補う．用言が欠落している場合，「セリビア決勝トーナメントへ」のように，名詞+助詞の形で語尾が終了している．そこで，大規模格フレームより名詞「トーナメント」，格「へ格」で検索し，用言「進出」を獲得する．

5.5 ニュースヘッドラインフレームへの格納方法

6W1H+用言フレームには意味理解システムを用いて情報を格納し，テーマフレームには分類パターンにおけるブロック A を格納する．また利用するブロック以外にもブロックがニュースヘッドラインの補足情報として存在する．そこで意味理解システムでのフレーム分解後にさらに上記の特徴を活かし補完を行う．具体例を図 8 に示す．

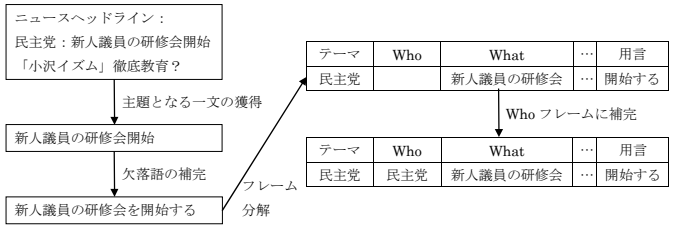

図 8　空フレームの補完例

6. 会話文生成処理

6.1 復元会話

6.1.1 復元会話の前提条件

復元会話はニュースヘッドラインフレームに整理された語句を利用し，各語句を下記のアルゴリズムにより結合し，会話文を生成する．そのため，ニュースヘッドラインを正しくフレーム分解できることが前提条件となる．

6.1.2 復元会話生成の概要

ニュースヘッドラインには語尾を省略するルールがあるため，ニュースヘッドライン「パキスタン：自爆攻撃で 41 人死亡」のように時制が明記されない場合がある．人は「死亡」が過去の出来事と連想し，時制を判断して会話している．そのためニュースヘッドラインの表記に表れない時制に注意する必要がある．また What フレームと用言フレーム間の助詞は一定でないため，助詞の補完が必要である．

6.1.3 時制への対応

ニュースヘッドラインの時制を基本的に現在形とし，現在形では不適切な場合に時制の変更を行う．時制は現在・過去・未来を扱う．過去時制の特徴として，「死亡」や「墜落」のように既に起こった事象特有の用言を過去用言と定義し過去代表用言知識ベース（表 9）を作成した．ニュースヘッドラインフレーム内の用言フレームと，過去代表用言知識ベース内の語の関連性から時制を判断する．

表 9　過去代表用言知識ベース

過去代表用言一覧
死亡，逮捕，起訴，けが，批判，退院，判決，捜査，訪問，求刑，勝訴，控訴，決定，先勝，提訴，表明，申す，保釈，取り消す，敗訴，発足，優勝，完成，連覇，紛失，勝つ，制覇，完成，探知，発見，回復，突破，釈放，和解，盗難，詐取

未来時制の判定はニュースヘッドライン内の「16 日から…」のような日付・日時と現在の日時との照合を行う．日付・日時が明記されない場合は現在として処理する．

6.1.4 復元会話生成時の助詞の補完

元のソースであるニュースヘッドラインの助詞がある場合はその助詞を採用し，ない場合は 5.4.1 節のアルゴリズムを用いて，助詞を補完する．

6.1.5 復元会話生成方法

復元会話を生成するために，ニュースヘッドラインから獲得できた全てのフレームを結合する．表 10 によりフレームを結合する．

表 10　復元会話生成におけるフレームの結合ルール

フレーム	ルール	助詞の補完例
テーマ	テーマ+で	新型インフルで
Who	Who+が	民主党が
When	When+に	23 日に
Why	Why+で	遅刻で
Where	Where+で	大阪地裁で
What	6.1.4 節参照	新人議員の研修会を

フレーム	ルール	助詞の補完例
Whom	Whom+に	三笠フーズ社長に
How	How+で	包丁で
用言（現在）	する→するんですよ	開始するんですよ
用言（過去）	する→したんですよ	逮捕したんですよ
用言（未来）	する→するようです	前倒しするようです

6.2　テーマ拡張会話

テーマ拡張会話を生成するため，イメージ付与，テンプレート組み合わせを行う．

6.2.1 テーマへのイメージ付与

ニュースヘッドラインを表現するテーマフレーム内の語句よりイメージを付与する．「ゴルフ=楽しい」,「時間=悲しい」のようにイメージと人の感覚・知覚には関連がある．そのため感覚・知覚判断システムにより語（想起語）を想起し，想起語に対し形容詞知識ベースを利用し，イメージを数値化する．処理の例を図 9 に示す．ただし感覚・知覚判断システムで出力が得られない場合は，イメージ付与を行わない．

図 9　キーワードに対するイメージ付与の例

6.2.2 イメージ別テンプレート組み合わせ

プラスイメージ時に使用するテンプレートを 5 セット格納したプラステンプレート知識ベースと（表 11），マイナスイメージ時に使用するテンプレートを 20 セット格納したマイナステンプレートを作成した（表 12）．

表 11　プラステンプレート知識ベースの一部

ID	テンプレート	ジャンル名
1	は楽しそうですね	スポーツ
2	は良いですね	スポーツ
3	は期待できますね	スポーツ

表 12　マイナステンプレート知識ベースの一部

ID	テンプレート	ジャンル名
1	は心配ですね	政治，問題，事故，事件，病気，災害
2	は問題ですよね	政治，問題，事件，経済
3	は残念ですよね	事故，事件，病気

テンプレートを選択する際,テーマの意味を拡張するため概念ベースにおける属性,シソーラス直親ノード，同義語をテーマの拡張語とする．そして拡張語と知識ベース内のジャンルとの関連度を評価し，閾値以上のテンプレートを会話生成に採用する．

6.3　会話文生成処理における実験と考察

本節では，復元会話とテーマ拡張会話生成における実験と考察をする．復元会話はニュースヘッドラインを正しくニュースヘッドラインフレームに格納されたときに，生成可能である．そのため，まずニュースヘッドラインフレームへの分解について考察する．

6.3.1 ニュースヘッドラインフレームへのフレーム分解実験と考察

実験

ニュースヘッドラインのフレーム分解について実験を行った．テストデータは　毎日新聞の Web サイト（毎日.jp）より，各パターンを 50 文，合計 250 文用意した．評価方法はフレーム分解を人手で実行した場合とシステムで実行した場合の差異を確認した．完全に一致した場合（正解）を「○」，1 つのフレームに違いがある場合を「-1」とした．実験結果を図 10 に示す．

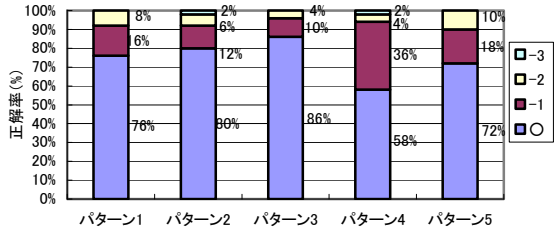

図 10　フレーム分解実験結果

出現頻度を考慮した全体精度は 75.8%となった．

考察

フレーム分解は会話生成処理の基本になるため低い精度のままでは復元会話に直接影響を与える．そのためまずフレーム分解の精度向上がシステム全体の精度向上に繋がると考えられる．

出現頻度の最も大きいパターン 1 での精度向上が，フレーム分解の精度向上に繋がる．パターン 1 では，助詞の補完での誤りが 6 件存在し，全体の誤り件数の 42%を占めていた．ニュースヘッドライン「学力テスト : 40%抽出で文科省調整　費用 30 億円に圧縮」では，主題となる一文の語尾「調節」とその直前の「省」より助詞の補完を行った．この時，調節の格フレームの結果からの「サイズ」と「省」が高関連になり，「文科省を調整」と失敗した．同様にニュースヘッドライン「ルーマニア：不信任決議可決　共産主義体制崩壊後で初」では，主題となる一文の語尾「可決」から得られる格フレームの結果の「会議」と「決議」が高関連になり，「不信任決議で可決」と失敗した．前者の誤りは関連度計算に原因があると考えられる．そして後者の誤りは「不信任決議」が場所を表す語でないという情報が予め必要であった．

6.3.2 復元会話とテーマ拡張会話生成の実験と考察

実験

6.3.1 節でフレーム分解した 250 文のニュースヘッドラインに対し，復元会話及びテーマ拡張会話を生成した．復元会話の評価は助詞・時制を考慮し正しい文章が出力できるかどうかを判定した．正しい出力が得られた場合を「○」とした．復元会話の実験結果を図 11 に示す．

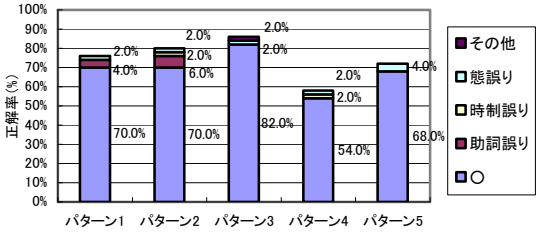

図 11　復元会話の実験結果

6.3.1 節と同様に各パターンの出現頻度を考慮した全体精度は 69.9%となった．この結果より正しいフレームから復元会話を生成する精度は 92.2%となった．復元会話の出力例を表 13 に示す．

表 13　復元会話の出力例

ニュースヘッドライン	出力例	評価
民主党 : 新人議員の研修会開始「小沢イズム」徹底教育	民主党が新人議員の研修会を開始するんですよ	○
新型インフル : 兵庫県西宮市で 8 歳女児死亡　基礎疾患なし	新型インフルで 8 歳女児が兵庫県西宮市で死亡したんですよ	○
ヤクルト : ユウキら 3 選手がＡ型インフル感染	ヤクルトでユウキら 3 選手がＡ型インフルを感染するんですよ	時制誤り
プレスリー : 毛髪が競売に　落札予想額は 70 万円以上	プレスリーで毛髪が競売にかける	態誤り

テーマ拡張会話では 3 人の被験者に生成された文が正しいかどうかの 2 択で解答してもらい，その結果を基に評価した．全員が正しい文章であると判断した場合を正解（○），全員が正しくない文章であると判断した場合を不正解（×）とした．それ以外の場合を（△）とした．実験結果を図 12 に示す．

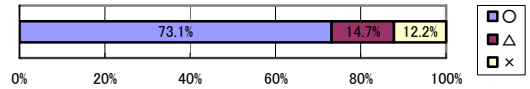

図 12　テーマ拡張会話の実験結果

図 12 より出現頻度を考慮したテーマ拡張会話の精度は 73.1%となった．そしてテストデータ 250 文から 341 文のテーマ拡張会話が獲得できた．成功例と失敗例を表 14 に示す．

表 14　テーマ拡張会話の出力例

ニュースヘッドライン	会話生成時に利用されたジャンル	出力結果	評価
新型インフル：兵庫県西宮市で8歳女児死亡　基礎疾患なし	-	新型インフルが話題になっていますね	○
本棚下敷き：パソコンデータで陳列場所を再現　北海道警	-	本棚下敷きが話題になっていますね	×
男子ゴルフ：石川、2打差の10位発進　首位は山下と野仲	スポーツ	男子ゴルフはすごいですね	○
在日米軍基地：米、環境協定に前向き　知事らが提案	事件	在日米軍基地は残念ですよね	×

復元会話の考察

正しいフレームから復元会話を生成する精度は 92.2%となった．これより正しいフレームに分解することで約 9 割の精度で復元会話が生成できることがわかった．そこで復元会話の精度向上の課題としてフレーム分解の精度向上が考えられる．また他の課題として最も誤り件数の多かった「助詞誤り」がある．

助詞誤りは全誤りの 51.7%を占めた．例えばニュースヘッドラインが「臨時国会：２３日召集に前倒し　与党合意」の場合，「テーマフレーム：臨時国会」，「What フレーム：召集」，「When フレーム：23 日」，「用言フレーム：前倒しする」とフレーム分解される．しかし復元会話生成時，助詞補完を行うことにより「臨時国会で 23 日に召集に前倒しするようですよ」と出力される．ニュースヘッドライン上の助詞「に」をそのまま補完したことによる失敗が見られた．この誤りのように補完する助詞を決定した後，格フレームを利用し What フレームと用言フレームと助詞の関係から，補完する助詞の妥当性を確認する必要があった．

テーマ拡張会話の考察

テーマ拡張会話の全体精度は 73.1%であり，342 文の正解出力が得られた．つまりテストデータ 1 文に対し，平均約 1.4 文の文章が自動生成された．

テーマとテンプレートとの組み合わせ処理での誤りでは，テーマから属性・シソーラス・同義語を利用し拡張語の中に雑音があり，その雑音とテンプレートのジャンルとが高関連度を示したため失敗が起きた．例えばニュースヘッドライン「在日米軍基地：米、環境協定に前向き　知事らが提案」から「在日米軍基地は残念ですよね」が出力された．テーマの語尾「基地」から拡張語「策源地，足場，補給，滑走，空軍」などが獲得されていた．そして拡張語「足場」とジャンル「事件」が高関連度となった．われわれは他のニュースから情報を得ることによって「在日米軍基地」が国際問題や政治問題と連想することができる．しかし本手法では「在日米軍基地」のみから「政治」や「問題」のジャンルを導き出そうとして失敗した．つまりテーマに関する他の情報も考慮し，拡張語を生成する手法が必要だと考えられる．

7. おわりに

本稿では，人とロボットとのニュースを話題にした会話は重要であると考え，ニュースから人が受ける常識的なイメージを利用し会話文を生成する方法を提案した．その際，ニュースヘッドラインの特徴を整理し，Web を利用した語の共起情報，連想メカニズムを利用した意味情報を考慮したフレーム分解を行った．また感覚・知覚判断システムを組み込み，常識的なイメージを考慮した会話を実現した．さらに多様な会話生成のために，本手法での既定のテンプレートに加え，テンプレートの一部を動的に変化させることにより，多様な会話文が生成されることが望まれる．

参考文献

1) 「毎日.jp」，＜http://mainichi.jp/＞
2) 奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14，No.5，pp.41-64，2007.
3) 渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，Vol.13，No.1，pp.53-74，2006.
4) NTT コミュニケーション科学研究所監修，「日本語語彙体系」，岩波書店，1997.
5) 河原大輔，黒橋禎夫，“高機能計算環境を用いた Web からの大規模格フレーム構築”，情報処理学会自然言語処理研究会資料，2006-NL-171-12，pp.67-73，2006.
6) 飛田良文，浅田秀子，「現代形容詞用法辞典」，東京堂出版，2003.
7) 篠原宜道，渡部広一，河岡司，“常識判断に基づく会話意味理解方式”，言語処理学会第 8 回年次大会発表論文集，B6-2，pp.651-654，2002.
8) 渡部広一，堀口敦史，河岡司，“常識的感覚判断システムにおける名詞からの感覚想起手法”，人工知能学会論文誌，Vol.19，No.2，pp.73-82，2004.
9) 米谷彩，渡部広一，河岡司，“常識的知覚判断システムの構築”，第 17 回人工知能学会全国大会論文集，3C1-07，2003.
10) 杉本二郎，渡部広一，河岡司，“概念ベースを用いた常識場所判断システムの構築”，情報処理学会自然言語処理研究会資料，2003-NL-153，pp.81-88，2003.