会話文脈に応じた関連情報提示タスクのための文脈類似度計算手法の開発

Developing calculation method of contextual similarity for providing information related to conversational context

白松 俊[†*]　駒谷 和範[†]　尾形 哲也[†]　奥乃 博[†]
Shun SHIRAMATSU, Kazunori KOMATANI, Tetsuya OGATA and Hiroshi G. OKUNO

[†] 京都大学大学院 情報学研究科
Graduate School of Informatics, Kyoto University

Abstract: We aim to develop a system that provides text information related to conversational context. Topics of conversation, i.e., targets of joint attention among conversational participants, depend on conversational context and dynamically change according to conversation progress. The conventional scales of importance of words (e.g., TF·IDF) , however, cannot estimate dynamic transition of the topics in conversational context. Our system requires the calculation method of contextual similarity, which can handle dynamic transition of the topics in conversational context. We developed such a calculation method on the basis of reference probability, i.e., a scale of salience of words. To investigate the effictiveness of our method, we performed an experiment to search information relevant to transcription of spoken utterances in CSJ (Corpus of Spontaneous Japanese). The experimental result showed some qualitative validity and many future issues on our system.

1　はじめに

談話参与者 (発話者, 受話者; 筆者, 読者) の共同注意が向かう対象 (トピック集合) は，談話文脈の流れに従い，発話単位毎に遷移する．本研究では，(1) 複数ユーザ間の会話文脈の流れに従って遷移するトピック集合を自動推定し，(2) その瞬間のユーザ間会話文脈に類似する文脈に位置するテキストを検索・提示するシステムの開発を目的とする．つまり，図 1 に示すように，会話文脈から自動推定したトピック集合をクエリとして検索することで，ユーザが明示的にクエリーを与える必要のない情報提示，すなわちクエリ・フリーな情報提示サービスを提供するシステムの実現を目指す．これは，会話の流れに連動して関連情報や広告をシステムが勝手に提示・更新することにより，ユーザ間の話が途切れたときに，新たな話題としての「ネタ振り」効果を期待したサービスである．

そのために，会話に現れる各単語に対するユーザ達の「注目度」の遷移を定量的に推定し，その瞬間の注目度が大きい単語群をクエリーとして用いる．われわれは，各単語に対して発話単位ごとに変遷するユーザ達の「注目度」を，単語の**顕現性 (salience)** と呼ぶ．

*〒 606-8501 京都市左京区吉田本町 京都大学
情報学研究科 奥乃研究室
E-mail: siramatu@kuis.kyoto-u.ac.jp

一般的な検索においては，Bag of Words と呼ばれるベクトル，すなわち，単語の TF·IDF によって重み付けされたベクトルが，文書の表現として用いられる．しかし，TF·IDF のような単語の頻度に基づく重みは，文脈の流れに従って遷移する顕現性を推定する用途には不向きである．なぜなら，顕現性は発話単位毎に遷移するので，頻度の計測範囲を細粒度にする必要があるが，細粒度ではどの単語の頻度も 0〜1 回になり，単語間の顕現性の差が検出できなくなるからである．さらに，顕現性は，文法役割などの出現頻度以外の要因 (たとえば日本語の場合，係助詞の「ハ」に係る名詞句は主題として働き，後続文脈でも注目されやすい) にも依存するので，それら他の要因も統合した顕現性の推定手法が必要である．

この問題を解決するため，本研究では**参照確率 (reference probability)** を顕現性の尺度として用いる．参照確率とは，「注目されている単語ほど，継続して参照されやすい」という仮定に基づいて設計された尺度である．具体的には，与えられた先行文脈に依存して定まる，ある単語が後続発話で参照される条件付き確率である．この確率は発話単位ごとに更新されるので，原理的にトピックの細かい推移を追うのに適している．さらに，頻度以外の要因 (文法役割など) も統合可能な統計的計測方法があるので，様々な素性を追加

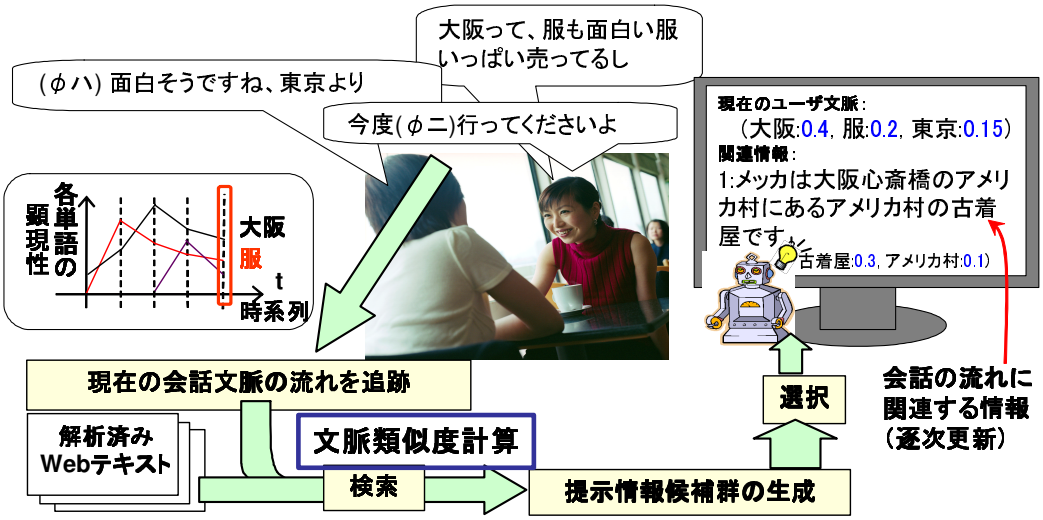

図 1: 会話文脈に応じた関連情報提示システム

することで，より精緻な顕現性の推定が可能である．

本稿では，参照確率を用いて各発話における単語の重要度を推定することにより，トピックの細かい推移を反映したクエリを自動生成し，文脈間の類似度の計算手法を開発する．さらに，会話文脈との類似度に基づくクエリ・フリーな検索の実験を行い，その有効性を検討する．

2 関連研究

以下の研究領域は，本研究で扱う「ユーザ間の会話文脈に応じた関連情報の提示」というタスクと関連が深い．

- TDT (Topic Detection and Tracking)
- クエリ･フリー情報検索 (Query-Free Information Retrieval)
- LSA (Latent Semantic Analysis)
- 意味ゲームに基づくセンタリングモデル (Meaning-Game-based Centering Model)

以下では，これら関連する研究領域について述べる．

2.1 Topic Detection and Tracking

TDT (Topic Detection and Tracking) [1] とは，Web 上のニュース記事群からトピックを抽出し，ニュースの時系列順に追跡するタスクである．TDT は，ニュースのテキストを個々のニュースに分割する Segmentation，分割されたニュースをトピック毎に分類する Topic Detection，新たに配信されたニュースを分類する Topic Tracking という 3 つのタスクから構成される．

TDT は，本研究の目的 (ユーザ間の会話文脈に応じた関連情報の提示) と関連の深い研究領域である．しかし，ユーザ間の会話から自動生成したクエリを用いて関連情報を提示するという問題は扱っていない．

2.2 クエリ・フリー情報検索

クエリ・フリー情報検索 (Query-Free Information Retrieval) [2, 3] とは，ユーザがクエリを明示的に与えなくとも，システムが自動的にクエリを推定して情報を検索する技術である．標準的なアプローチとしては，ユーザが選択した文書全体を現す Bag of Words ベクトル (TF·IDF に基づく) を生成し，これをクエリとして用いて類似文書を検索するという手法がある．しかし，本研究で扱うタスク (ユーザ間の会話文脈に応じた関連情報の提示) においては，文書全体や会話全体でなく，文脈の流れに従って発話単位ごとに変遷するトピック集合を推定できて初めて，会話文脈の流れに応じた情報検索が可能となる．よって，本研究の目的のためには，文脈の流れに従って変遷する単語の顕現性を表せる尺度が必要となる．

2.3 Latent Semantic Analysis

LSA (Latent Semantic Analysis; 潜在的意味解析) [4] とは，単語と文書の共起行列に特異値分解を施し，その結果を用いて関連の強い単語群を同一次元に縮退させることで，関連語間の関係を考慮した類似度計算

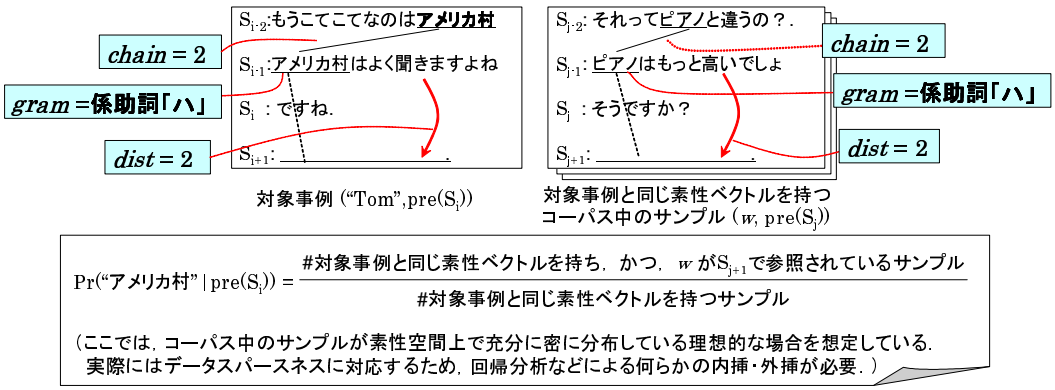

図 2: 参照確率計測の基本的アイディア

を可能にする手法である．本研究で扱うタスク (ユーザ間の会話文脈に応じた関連情報の提示) においても，関連語間の関係を考慮した類似度は必要となると予想される．

しかし，文脈に従って変遷する顕現性の推定という本研究での課題に対しては，LSA は直接の解決を与えるものではない．

2.4 意味ゲームに基づくセンタリングモデル

意味ゲームに基づくセンタリングモデル (Meaning-Game-based Centering Model) [5] とは，センタリング理論 (Centering Theory) [6] を統計的・定量的に一般化したモデルである．センタリング理論とは，「顕現性が高い単語は継続して参照されやすく，代名詞化されやすい」という，参照結束性 (referential coherence) に関する理論である．同理論は，発話単位毎のトピック遷移をモデル化しているが，顕現性を先験的なルールによって定義していたので，本研究で必要となる文脈類似度計算には適していなかった．この課題を解決すべく，われわれは過去の研究 [5] において，センタリング理論をゲーム理論に基づいて統計的・定量的に一般化した．これが，意味ゲームに基づくセンタリングモデルである．

同モデルでは，単語の顕現性を**参照確率 (reference probability)** という尺度で定量化した．参照確率とは，与えられた先行文脈の下で，当該単語が後続の文で参照される条件付き確率である．これは，「注目されている単語ほど，次の文でも継続的に参照されやすい」という仮定に基づいて設計された尺度である．日本語と英語のコーパスを用いた実験 [5] の結果，以下の性質を持つことが確かめられた．

- 先行文脈中で最近出現した単語の方が，もっと前に出現した単語よりも参照確率が高い．
- 先行文脈中で多く出現した単語の方が，もっと少なく出現した単語よりも参照確率が高い．
- 係助詞「ハ」に係る単語は，格助詞「ガ」に係る単語よりも参照確率が高く，「ガ」に係る単語は「ヲ」に係る単語よりも参照確率が高い．

これらの性質は，顕現性に関する言語学的な知見と整合している．これらの性質から参照確率は，注目対象の動的な遷移を考慮した文脈類似度計算に適していると考えられる．

3 参照確率ベクトルに基づく文脈類似度

本節では参照確率の定義および計測手法を説明し，それに基づく文脈類似度の計算方法を述べる．ただし，以下では会話やテキストの談話構造を文単位の列 $[S_1, S_2, \cdots, S_n]$ によって表し，S_i までの先行文脈 $[S_1, \cdots, S_i]$ を $\mathrm{pre}(S_i)$ と表記する．

参照確率の定義

参照確率とは，文 S_i までの先行文脈 $\mathrm{pre}(S_i)$ が与えられた下で，w が後続文 S_{i+1} で参照される条件付き確率 $\Pr(w|\mathrm{pre}(S_i))$ である．この確率が，文 S_i 時点における単語 w の顕現性を表す．

参照確率の計算

図 2 は，コーパスを学習データとして用いて参照確率を測定するための基本的アイディアを表している．文 S_i における “アメリカ村” の参照確率，つまり，後続文 S_{i+1} で “アメリカ村” が参照される確率を求めるために，図 2 では以下の 3 つの素性を用いている．

- *dist*=2 (“アメリカ村” が最後に出現した S_{i-2} から S_{i+1} までの間の発話数)

表 1: 参照確率 $\Pr(w|\mathrm{pre}(S_i))$ の計算に用いる素性

dist	$\log((\#\ \mathrm{pre}(S_i)$ 中で最近 w が現れた位置と S_{i+1} の間の文単位の数)+1)
gram	$\mathrm{pre}(S_i)$ 中で最近 w が現れた箇所における文法役割
chain	$\log((\#\ \mathrm{pre}(S_i)$ 中で w が現れた回数)+1)
last_topic	w が $\mathrm{pre}(S_i)$ 中で最近の主題 (係助詞「ハ」) か (yes/no)
last_sbj	w が $\mathrm{pre}(S_i)$ 中で最近の主語 (格助詞「ガ」) か (yes/no)
p1	w が一人称を表すか (yes/no)
pos	w の品詞
in_header	w がタイトルや見出し中で参照されているか (yes/no)

- *gram*=係助詞「ハ」 (“アメリカ村” が最後に出現した S_{i-2} での “アメリカ村” の文法役割)
- *chain*=2 (“アメリカ村” の S_i までの出現回数)

この素性ベクトルと同じ素性ベクトルを持つサンプル $(w,\ S_j)$ をコーパスから抽出し，その中で w が S_{j+1} で参照されているサンプルの割合を，S_{i+1} で “アメリカ村” が参照される確率の近似値と見なすことができる．

しかし，実際には対象事例と全く同じ素性ベクトルを持つサンプルがコーパス中に充分にあるとは限らないため，何らかの内挿・外挿法が必要となる．本稿では，素性ベクトル上でコーパス中のサンプル集合に $k = 50$ の k-NN 平滑化を施し，それを学習データとしたサポートベクター回帰 (SVR; Support Vecter Regression) を内挿・外挿法として用いることで，参照確率を計測する．ただし，SVR には Tiny SVM [7] を用いる．これにより，素性空間上でコーパスのサンプルが疎な分布を示していても参照確率が計算できるようになったので，本稿では表 1 の素性を用いる．学習データとして用いるコーパスとしては，照応・共参照の正解タグが付与されたコーパスが必要となる．それは，ユーザの会話と，検索対象のテキスト文書のそれぞれに似たコーパスであることが望ましい．しかし，まだそのような照応タグ付きコーパスは無い．そこで本稿では，毎日新聞記事に GDA (Global Document Annotation) [8] の照応タグが付与されたコーパスを，参照確率の計算のための学習データとして用いる．

文脈類似度の計算

まず，コーパスに含まれる N 個の単語に対して N 次元空間を割り当てる．このとき，ある時点の当該文 S_i における単語 w の参照確率 $\Pr(w|\mathrm{pre}(S_i))$ から成る N 次元ベクトルにより，文 S_i 時点における文脈を表現する．このベクトルを，本稿では**参照確率ベクトル**と呼ぶ．

ユーザ間の会話に含まれる発話単位 U が有する文脈と，検索対象のテキスト文書に含まれる文単位 S が有する文脈とを，それぞれ参照確率ベクトル $\boldsymbol{v}(U)$, $\boldsymbol{v}(S)$ で表現する．この参照確率ベクトル間のコサイン類似度 $\frac{\boldsymbol{v}(U)\cdot\boldsymbol{v}(S)}{|\boldsymbol{v}(U)||\boldsymbol{v}(S)|}$ を，動的な遷移を考慮した文脈の類似度として用いる．

4 書き起こし会話データの会話文脈に応じた関連情報の検索実験

参照確率ベクトル間のコサイン類似度を文脈類似度と見なす手法を用い，会話の書き起こしに対する関連情報提示実験を行った．会話の書き起こしとしては，日本語話し言葉コーパス (CSJ; Corpus of Spontaneous Japanese) [9] vol. 17 のディレクトリ D03F0040 に含まれる自由会話を用いた．提示する関連情報の候補として，会話に含まれていた名詞句 330 単語 (複合語も含む) の各単語を含む Web 上の HTML 文書 18,429 文書を収集した．そこに含まれていた 397,089 文を，提示する関連情報の候補として用いた．

4.1 実験手順

以下に，本稿で行った実験の手順を示す．

1. D03F0040 の会話に含まれていた 330 単語を 1 単語ずつクエリとして検索エンジン Google で検索して得た上位 100 件の URL のうち，HTML 文書のみを収集した．
2. 収集した HTML のうち，XHTML に変換できた文書集合に対し，係り受け解析器 CaboCha [10] で自動解析した統語構造を GDA タグとして付与した．
3. 統語構造を自動付与した 18,429 文書に含まれていた名詞句 17,934 単語 (複合語も含む) に，17,934 次元空間[1]を対応づけた．
4. 18,429 文書に含まれていた 397,089 文単位の各々に対し，その文が有する文脈を表す参照確率ベクトルを付与した．
5. D03F0040 の会話に含まれていた 495 発話単位の各々に対し，その発話単位が有する文脈を表す参照確率ベクトルを付与した．
6. 495 発話単位の各々に対し，提示テキストの候補 397,089 文から，参照確率ベクトルのコサイン類似度が高い候補 k 件を，提示する関連情報として選択した．

[1]本稿で用いた手法で文脈類似度計算に使われるのは，会話に含まれる 330 単語に対応する次元のみである．しかし，今後，関連語を考慮した文脈類似度計算手法の実験をするためには，会話に含まれない単語についても次元を割り当てておく必要がある．

4.2 実験結果

以下では，実験によって得られた検索結果の例を示す．以下の書き起こしデータの，発話 U_{116} 時点での会話文脈に対する関連情報を検索した結果を示す．

D03F0040 に含まれる会話例

⋮
(U_{112}) B: 何かね東京の人ってやっぱり
(U_{113}) B: 有名になったものは認めるけれども
(U_{114}) A: そうそうそうそう言ってた
(U_{115}) B: 新しいものに対しては最初はね
(U_{116}) A: **ねそうみたいですよね**
(U_{117}) B: 厳しいですよね
(U_{118}) B: でも関西人は素直に
(U_{119}) B: すっと入りますよね
(U_{120}) A: いいものはいい
⋮

ただし，U_{111} までの先行文脈においては，主に関西人，関西弁，東京との対比などが顕現性が高いトピックとなっている．

以下の○印で示されたゴシック体の文は，発話 U_{116}「**ねそうみたいですよね**」の時点での会話文脈に近い文脈を有すると判定された候補である．参考のために，その前後の文も示してある．

発話 U_{116} 時点の文脈に近いと判定された文

1 位 (コサイン類似度: 0.418)
まず、ゆうておくけど、関西弁ゆうても大阪弁、京都弁、そして、ワイの住んどる滋賀にも独特の関西弁がある
○ **まあ、細かいことは気にしたらあかんけど、実際にはちゃうことを覚えといてや**
たこ焼きが好きやぁ？

2 位 (コサイン類似度: 0.416)
まあ、細かいことは気にしたらあかんけど、実際にはちゃうことを覚えといてや
○ **たこ焼きが好きやぁ？**
そんなもん、あたりまえや

3 位 (コサイン類似度: 0.415)
関西人が認める東京の店
○ **関西人が認める東京の店**
東京の店はハコだけや！

1 位，2 位の文単位は，関西や関西弁についての文脈に位置し，3 位の文単位は，関西人が東京の店に関して語っている文脈に位置するので，会話文脈と類似した文脈を有する妥当な検索結果であると言える．

「**ねそうみたいですよね**」という内容語を全く含まない発話 U_{116} に対し，文脈を反映した検索結果が得られた理由は，U_{116} 時点における文脈を表す参照確率ベクトル $\boldsymbol{v}(U_{116})$ が示唆している．

$\boldsymbol{v}(U_{116})$ **の要素** (値が大きい順): 関西人:0.3246 東京:0.2934 大阪:0.2928 関西:0.2150 久しぶり:0.1810 町:0.1810 人種:0.1807 友達:0.1807 火:0.1807 精神:0.1807 最初:0.1701 人:0.1671 おばちゃん:0.1589 日本語:0.1149 関西弁:0.1149 ⋯

すなわち，U_{116} 時点の文脈を表す参照確率ベクトルには，先行文脈に含まれる単語群 (関西人，東京，大阪，⋯) が，高い参照確率 (顕現性) を保ったまま含まれていた．このことが，文脈を反映した検索結果をもたらしたと考えられる．

ただし，1 位と 2 位は，同じ HTML 文書の中で隣接して位置していたので，本来ならば別の候補としてではなく，まとめて見せるべきである．この不備は，前述した TDT タスクにおける Segmentation のような処理をしていないことが原因である．

5 残された課題

書き起こし会話データの会話文脈に応じた関連情報提示実験の結果，示唆された課題を以下に列挙する．

1. **提示手法の妥当性の定量的な評価**．検索結果を無作為抽出し，人手で再現率と適合率を数えることにより，関連情報提示の性能を評価する必要がある．同時に，ユーザ間の会話に対する「ネタ振り」効果の定量的評価手法を設計する必要がある．
2. **話題の切り替わり点を考慮していない**．話題の切り替わり点においては，単語群が持つ参照確率 (顕現性) の分布が大きく変化すると予想される．よって，話題の切り替わり点を考慮した参照確率の推定手法の設計が重要な課題である．
3. **テキストセグメンテーション**．同じ文書の隣接する文が別々に提示される不備を解消するためには，意味段落のセグメンテーションが不可欠である．この処理により，前述した話題の切り替わり点の検出が可能になるという利点も考えられる．
4. **関連語による連想を考慮していない**．たとえば，「アメリカ村」や「古着屋」の顕現性が高い文脈においては，各々の関連語である「大阪」や「服」の顕現性も同時に上がる傾向があると考えられる．しかし，参照確率は当該単語単独の参照パターンだけから計算されるので，関連語の影響を反映していない．そのため，文脈と深く関連するが背景知識として省略されている語などの顕現性

は不当に低く推定されてしまい，検索結果の再現率向上の障害になると考えられる．

この問題に対しては，LSA が一つの解決策を与える．しかし，数万次元から成る共起行列に対して特異値分解を施すには数 GB のメモリ空間が必要となるので，そのような場合に LSA を適用するのは簡単ではない．そこで，今後は [11] でわれわれが提案した連想を加味した顕現性推定手法に基づき，たとえば「大阪」や「服」がトピックになっている会話文脈に対し,「アメリカ村」や「古着屋」がトピックになっているテキストを提示できるかを検証する必要がある．

5. **音声認識誤りを考慮していない**．本稿では書き起こしデータを用いて実験を行ったが，実際には音声認識誤りが重大な障害となると予想される．自由会話を認識するためには大規模語彙を持つ統計的言語モデルが必要になるが，語彙サイズが大きくなると認識精度が下がるので，解決が非常に困難な課題である．

6. **話し言葉に特有の素性を用いていない**．話し言葉においては，文法役割などよりも，イントネーションなどの音韻的な素性が参照確率の推定に有効であると予想される．

6 おわりに

本稿では，会話文脈の流れを考慮した文脈類似度計算手法を，参照確率に基づいて設計・開発した．また，本手法を用いて日本語話し言葉コーパスの書き起こしデータに対する関連情報提示実験を行い，ある程度妥当な関連テキストを検索できることを確認するとともに，多くの課題を示唆する結果を得た．

今後は，前節で述べた多くの課題を解決し，会話文脈と連動して関連情報を提示するシステムを完成させる．まずはチャットなどの文字メディアに対してのサービスを運用すると同時に，音声の自由会話に対して適用する実験も試みる予定である．

謝辞

参照確率計測用の学習データとして用いた新聞記事 GDA コーパスについて，研究利用を許諾して下さった三菱電機株式会社と，調整の労をお取り頂いた GSK に感謝する．本研究は，科研費 (特別研究員奨励費) の支援を受けた．

参考文献

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic Detection and Tracking Pilot Study Final Report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, 1998.

[2] P. Hart and J.Graham. Query-free information retrieval. *IEEE Expert*, Vol. 12, No. 5, pp. 32–37, 1997.

[3] M. Henzinger, B-W. Chang, B. Milch, and S. Brin. Query-free news search. In *Proceedings of the 12th International World Wide Web Conference*, pp. 1–10, 2003.

[4] T.K. Landauer and S.T. Dumais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, Vol. 104, No. 2, pp. 211–240, 1997.

[5] S. Shiramatsu, K. Komatani, K. Hasida, T. Ogata, and H.G. Okuno. Meaning-Game-based Centering Model with Statistical Definition of Utility of Referential Expression and Its Verification Using Japanese and English Corpora. In *Proceedings of the 6th DAARC*, pp. 121–126, March 2007.

[6] B. Grosz, A. Joshi, and S. Weinstein. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, Vol. 21, No. 2, pp. 203–225, June 1995.

[7] T. Kudoh. TinySVM: Support Vector Machines. http://chasen.org/˜taku/software/TinySVM/, 2002.

[8] 橋田浩一. GDA: 意味的修飾に基づく多用途の知的コンテンツ．人工知能学会誌, Vol. 13, No. 4, pp. 528–535, 1998.

[9] K. Maekawa. Corpus of Spontaneous Japanese: Its Design and Evaluation. In *Proceedings of the ISCA & the SSPR2003*, pp. 7–12, 2003.

[10] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析．情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[11] 白松俊, 駒谷和範, 尾形哲也, 奥乃博. コーパスからの関連語獲得に基づく連想を加味した顕現性の推定．言語処理学会 第 13 回年次大会 発表論文集, pp. 522–525, 2007.