SDRT による談話の意味記述

— e-Learningにおける回答中の類似度推定に向けて —

吉 武 春 光

1．まえがき

e-Learningが盛んである．本学でもe-Learningシステムを導入し，2005年 4 月からe-Learning授業を開始している．筆者は2000年より，授業内容を出来る限り電子化し，それをホームページ上で公開する作業を続けてきた．更に，2003年より，簡易的な掲示板システム（Nucleus Blog）を使って，Web上で授業の一部を行ってきた．その後，2005年 4 月からは，商学部経営学科の科目「ビジネス情報技術入門」「情報ネットワーク論」などで，複数のe-Learningシステムを使用した本格的な授業を行っている．これらの経験を経て，筆者は，e-Learning を使った授業において幾つかの問題を感じた．その中でも，特に，教員の負担に関する問題が一番大きかった．そこで筆者は，長年取り組んできたコンピュータによる自然言語処理の技術を使って，教員の負担を減らす目的で，e-Learning授業を支援するプログラムを開発することを目指すことにした．本稿は，学生が書いた回答文章の予備分析，並びに，支援プログラム開発に向けての理論的なモデルの検討を行うものである．

2．大学におけるe-Learning使用における類似度の問題

一般のe-Learning未経験者がe-Learningを捉えているのは次のような形態であろう．「第 3 者が作成したコンテンツや問題を使用し，受講者が勝手に自学自習を行う．」その延長線上の捉え方として，教員の負担が減るので教員一人

で数多くの受講生を受持つことが可能になると．しかし，これは企業における e-Learningをそのまま大学に当てはめているだけであり，現実とは大きくかけ離れている．大学におけるe-Learningの場合，企業とは異なる問題点が生じている．

既存のコンテンツや問題は，教員の教育目的に完全に一致はしていないので，一部分しか使えない．どうしても，教員独自のコンテンツや問題を作成する必要がある．e-Learningのコンテンツや問題を作成する為には多大な時間とスキルが必要になる．

一般にe-Learningではきめ細やかな学生対応が可能であるが，逆に言えば，教員がコンピュータに向って学生を相手に採点，指導する時間が増えることになる．学生の理解度向上という観点では好ましいが，教員の負担増になる．文系大学の場合，１科目当りの受講学生数が多いので，教員がe-Learningを採用する為には教員の負担を減らす為に補助教員としてのTAを使用することが必須である．

企業と違い，学生はe-Learningを使う必然性がない．例えば，自分のためになるので，e-Learningを使って勉強しておくようにと指示しても，学生はe-Learningを使おうとはしない．e-Learningの進行状況やオンラインテストの結果を成績に反映させるようにすると，学生は俄然e-Learningを使うようになる．つまり，学生のモチベーションを向上させる工夫をすることがe-Learning活用の大事なポイントとなる．

e-Learningシステムでは，学生に対して幾つかのタイプの問題を出すことが可能である．１つは，予め作成しておいた回答に合致しているかどうかをシステムが自動的に調べて採点を行い，即座に結果を表示するタイプのものである．このタイプでは，予め用意した回答中から選ばせるか，短い語を入力させるようになっている．もう１つは，ある課題のテーマについて学生に自由記述回答をさせて，後で教員が，それを読んで採点を行うものである．前者においては，出題に際しての教員の負担が大きく，努力に見合う成果を得ることが出来るのは受講生が多い講義であろう．一方，後者は，出題に際しての教員の負担は少ないが，採点時の負担が多い．そこで，自動的に内容を評価採点してくれるシ

ステムや，その内容が，どの程度，模範解答に似ているのかを提示してくれる支援システムの構築が望まれる．なお，この傾向はe-Learningに限らず，従来の紙媒体を使用した試験においても同様であろう．

実際に後者のタイプの採点をしていて気になるのは，次の２つの点である．

1）検索エンジンで見つけたWeb上のページの内容を，そのまま，或いは，僅か変更しただけで，回答文として提出してあるもの．

2）友人から教えてもらった内容を，そのまま丸写しで，或いは，僅か変更しただけで，回答文として提出してあるもの．

これらは本質的には情報倫理の著作権の問題に帰着するのであるが，現実的に採点に携わっていると，正当に学生を評価できないことに問題を痛感する．そこで，教員の負担低減と学生の努力を正確に評価する為には，検索エンジンで調べたWebページの内容と回答文との間の類似度，更には，回答文同士の類似度を計る仕組を考えるべきである．この問題は，従来からの紙記述の回答においてもあったが，電子化されていない為に教員が類似度を推測するのに止まっていた．しかしe-Learningにおいて提出された，電子化された回答であれば，コンピュータ処理により容易に，かつ，自動的に類似度を測定することが可能になるはずである．

本研究が目指す方向性は，上記の通りであるが，まず手始めとして，e-Learningシステムにおいて，学生が回答した文同士の類似度，教員が作成した模範回答と学生が回答した文との間の類似度を，自動的にコンピュータ処理により算出し，教員に提示することを試みる．検索エンジンで調べたWebページの内容との類似度計算は，その延長線上に考えられる．

3．学生から提出された回答文の分析

筆者は 2005年度の情報ネットワーク論という科目において，e-Learning課題として「日本の文字コードの種類を述べた上で，半角カタカナと呼ばれるコードの問題点を説明せよ.」という問題を出題した．この課題に対して受講生から全部で78回答を得た．まずは，回答文を，どう計算機処理すべきかを見極めるために，この回答を手で分析した．

3.1 学生が回答した文章の例

この課題に対する学生からの回答文の特徴的なものを次に幾つか示す．

箇条書きタイプ（回答文番号183）

◆日本の文字コードの種類◆
シフトJIS；日本のパソコン用の文字コードとして作られた．
EUC-JP；UNIX用
ISO-2022-JP；電子メール用（最上位ビットは0，つまり7bitでおさまる）
◆半角カタカナと呼ばれるコードの問題点◆
電子メールに，JISX0201の8単位符号表の右半分が入っていると，表示の際に文字化けを起こすものが多い．
※インターネットの世界では使用禁止．

文章タイプ（回答文番号128）

文字コードとは文字ごとに数字を割り当て，表現したものであり，その種類としては，ASCII，ISO８８５９－１，JIS X ２００１がある．
それぞれ，ASCIIは米国の標準コード，ISO8859-1はASCIIをヨーロッパ各国用に拡張したコード，また，JIS X 0201はASCIIにカタカナ用の拡張を施したものである．ただし，JIS x 0201はISO8859-1の拡張部分と重複したコードがあるため，あるコードをISO 8859-1で表現されたときと，JIS x 0201で表現されたときではまったく違う文字として現れる．
よって，電子メール等のインターネット世界においては，半角カタカナを使用してはならない．もし使用してしまったら受信者側で文字化けしてしまうことがしばしば起こる．
また，日本語には漢字が存在するために，ASCIIのコードでは表現するのは不可能であり，漢字を表現するために漢字のためのコードを作成し，さらにASCIIと漢字コードとを切り替えるコードが作成され，それはISO-2022-jpというコードである．
一方で，世界中の文字をひとつにまとめた体系としてユニコードが米国主導で開発されたが，実態にそぐわない部分がある．

要約タイプ（回答文番号139）

ASCIIコードにカタカナをつけたものをJISX0201と言う，JISによる日本の漢字コードをJIS X 1208と言い，実用的なコードは三つシフトJIS，EUC-JP，ISO-2022-JPが存在する．JISX0201は半角カタカナを格納するASCIIコード部分にヨーロッパでは他の固有の文字が入っており，これが文字化けを起こす．

3.2 学生が回答した文章の分析と結果の検討

78回答を手で分析した結果を次の表１に示す．特徴の延べ数は116個であった．

特徴としては，入力の文字種などの多義によるもの（60%）が多く，次いで，箇条書きやWeb書式を表意文字として使っているなどの，脱文法的な表現に依存する問題（38%）が多かった．

計算機処理の立場で判断すれば，入力文字種の問題（60%）は，計算機処理により容易に解決できる問題であり，具体的には，計算機処理の前処理段階に対処方法を組み込むことが可能である．一方，箇条書きなどの脱文法的な表現に依存する問題（38%）は，単文の意味処理や複文から成る談話の意味処理を行う必要がある．更には，２% ほど存在している，記号による代理表現を適切に処理するためには，これらの意味処理の延長線上に可能となる推論処理が必要である．逆に言えば，意味処理を考える場合，推論処理までを視野に置いたモデルが必要となる．

出題の際に，回答の記述形式や回答の長さなどを，きめ細かく指定し，不適切な回答を提出した学生に，再提出を求めたりして，回答を一定レベルに揃えることは可能であるが，現実的ではない．そこで，本研究では，学生が回答した文章を，そのまま解析するという方針で臨むことにした．

表 1　学生の回答文の分析結果

	現　　象	延べ出現数	小計
入力文字種の問題（入力段階で対応可能）	外字	24	70 (60%)
	半角と全角の数字	16	
	大文字と小文字のアルファベット	9	
	似た記号の混用（半角と全角，-と_）	8	
	半角と全角のアルファベット	6	
	漢数字	2	
	中丸を読点の代りに	2	
	カンマと読点の混用	1	
	シフトの代りにshift	1	
	EUC－JPの代りにEUC	1	
回答全体の意味をどう表現するか？談話処理が必要	箇条書き（全部，または，部分的に）	13	35 (30%)
	記号が意味を持って使われている．	10	
	全角のスペースが述語の代り．	1	
	名詞止め	6	
	Web入力画面の書式を意味として使っている．	5	
推論処理が必要．	（）による修飾	1	2 (2%)
	強調として『と』で囲んである．	1	
その他	意味が通じない．	5	9 (8%)
	文末が読点になっている．	1	
	回答の先頭に問題文を書いている．	1	
	不要な全角のスペースが入っている．	1	
	「つかう」という平仮名	1	
合計			116

4．本研究が採用する自然言語処理の方式の検討

自然言語をコンピュータ処理する場合，形態素処理，構文処理を経た後，意味処理を行うやり方と，意味処理を行わずに構文処理の結果を統計的に処理するやり方とがある．本研究では，推論処理までを目指したいので，意味処理を行うやり方を採用することにする．しかしながら，意味処理として確立した手法が存在している訳ではない．

4.1 コンピュータにより自然言語処理の処理方法

コンピュータによる自然言語処理は，一般的には次の処理段階を経る．

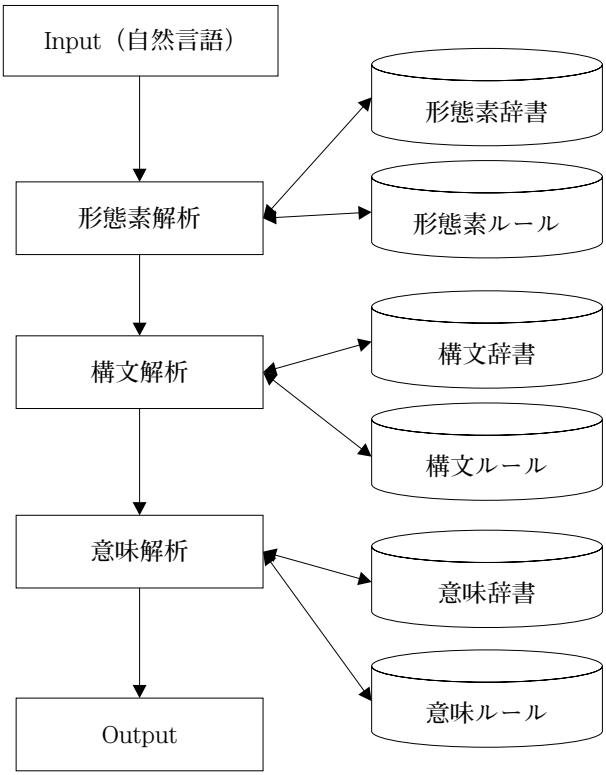


図 1　自然言語処理の処理段階

ここで，意味解析の結果として出力される意味表現は，使用する意味のモデルに基づくものとなる．そこで，まずは，使用する意味表現モデルの検討を行う．

4.2 類似度計算に必要な意味処理

類似度を自動測定する研究は，色々と行われている．しかしながら，それらは構文的な特徴に基づくものである．

文献（荒牧，黒橋，柏岡，加藤，2005）においては，単語ごとの一致を重視するのではなく 2 つの文の間で大きく一致しているかどうかを類似度が高いかどうかの判断基準とすべきだとしている．しかしながら，e-Learningにおいて学生が回答した文章を見る限りにおいては，文法に適合していない文が多く，文章同士の比較による類似度計算には無理があると思われる．文献（山下，富士，大倉，潮田，2003）では，翻訳の訳例検索における「一致率」と「スコア」という 2 つの観点の類似度計算について述べてある．「一致率」は，文において単語が一致している率が高いものの方を類似度が高いと判断するというものである．「スコア」は，部分一致箇所が多いものの方を類似度が高いと判断するというものである．

しかし，文献（高橋，乾，松本，2002）の課題においても指摘されていように，人間が文章を比較したときに類似しているかどうかを判断するのは，構文的に類似しているかどうかという観点と意味的に類似しているかどうかという観点の両面が必要である．そこで，本研究では，構文処理の後で，意味処理を行うことで，より人間の判断に近い類似度の計算を試みることにした．

なお，3.1 節で述べたように，学生の回答文は，箇条書きであったり，文章であったりするが，結局は，複数の文によって成り立っている．そこで，類似度を計算する際には，複数の文によって成り立っている談話としての類似度の計算が必要となる．

4.3 意味記述モデルの検討

一言で「意味処理」というが，根底には，人間が文章を読んだときに，その文章を理解するプロセスを解き明かすことが必要になる．その為の研究は，1990

年代の自然言語処理の聡明期に，工学，文学，医学，心理学などの幅広い分野で多角的に行われた．しかしながら，決定的な解法は見いだされていない．現状では，目的とする自然言語の計算機処理を行うのに必要十分な意味記述と意味処理を行うのが最善である．但し，domain specificな意味記述＆処理ではなく，出来る限り他のdomainへも適用可能なように汎用性に注意を支払う必要がある．

簡単な意味処理としては，次がある．

1）意味素性や意味分類

2）格フレームや意味ネットワーク

1）は，予め，語の意味を，意味的にそれ以上は分解できない意味素性(primitives又はfeatures）に分類しておき（これを意味分類と呼ぶ)，語を辞書に登録する際に，意味として意味素性を入れておき，実際の解析の際には，その意味素性を使って語の選択を行うという手法がある．しかしながら，語の意味を他の語を使って規定する必要がある為，トートロジーに陥る可能性がある．更には，どの程度の種類の素性を用意すれば十分なのか，語に，その素性を当てはめる際に，ユニークに当てはめることは可能かどうか？などの問題が生じる．2）は，用言を予め分類しておき，その関係を係り受け関係やネットワークを使って表すというものである．

更には，意味そのものを何か独立した体系にて記述するという研究も存在している．例えば，筆者も関わった文献（横田，吉武，田町，1986）においては，意味を心像意味論の観点で取り扱っている．しかしながら，心像意味論に従う意味記述も，どの範疇まで記述するかという自由度が高く，コンピュータ処理が複雑になる．心像意味論がベースとしているKatz＆Foderの意味論は，生成意味論と呼ばれる方向性である．つまり，個々の語の意味を文法に従って予め記述しておけば，文法に従って，その語の意味を組み合わすことによって文全体の意味を生成することが可能になるという主張である．しかし，実際に，心像意味論を用いて語の意味を記述すると，何をどこまで記述しておく必要があるのか，という点が不明確である．認知心理学の分野では色々と論争が繰り返されたが，Katz亡き後は，その方向性の主張をサポートする意見は見られない．

代りに，生まれてきたのが，語の意味は，予め決めておけるものではなく，文脈（談話）の中で決まるものであるという主張である．この主張に基づく語の意味記述の理論が（Asher, Alex, 2005）で提案されているSegmented Discourse Representation Theory（以下SDRTと略す）である．SDRTでは，Dynamic Semanticsという考え方を採用している．つまり，予め語の意味を全て決定しておくのではなく，文脈が変化すると，それに応じて談話全体としての意味が決まってくる，というものである．つまり，決められない意味は未定義のまま変項として残しておくのである．そして，文の意味を生成する段階，更には，談話の意味を生成する段階において，逐一，論理演算を行い，変項に値を当てはめようとするのである．その際，変数が必ずしも決定されるという訳ではない．つまり，決まらない意味は，未定のままにしておくことが出来るのである．後から出現する文により曖昧さの解消が行われる可能性があるのである．なお，SDRTは，談話の取り扱いの理論であり，語の意味記述としては，(Pustejovsky, 1995) のGenerative Lexicon（生成語彙論，以下GLと略す）を採用している．

日本語に対して SDRTを適用しようという試みが存在している（菊池，白井，2004）．そこでは，接続詞「の」が持つ曖昧さが，談話において解決されていく様子を見ることが出来る．

SDRTやGLは，まだ完成された理論ではない．しかし，本研究が目指す，回答文の間の類似度の計算は，絶対的なものではないので，相対的な値を示すことが出来れば十分である．この観点から判断して，SDRTの「決められない意味は変数のまま残しておく」という考え方は適したものであろう．

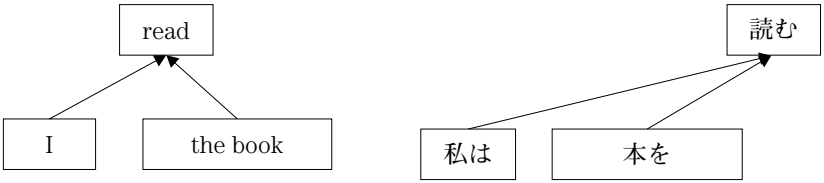
5．GL を使った意味表現

本章では，GLの概略を紹介しておく．

5.1 自然言語から GL への変換

GLでは，自然言語からGLを生成する手続きについては規定されていない．GLが想定しているのは，構文解析の結果が「節（句）に該当するものが依存

構造（係り受け関係）を成している」ということである．つまり，形態素解析と構文解析は任意のものを使用して節（句）間の依存構造を出力し，それを元にGLを作り出すことになる．次の図 2 に，英語と日本語の依存構造の例を示す．



ここで，→の根本の節（句）から先の節（句）に向かって依存が成り立っていることを示している．

図 2　英語と日本語の依存構造の例

5.2 GL を使った意味表現の例

個々の節の意味表現は，次の 4 つの項目から成っている．

Argument structure（項構造）：　統語的に実現される項構造を記述する．

Event structure（事象構造）：　語が表す事象とその時間的な関係を記述する．

Qualia structure（特質構造）：　語の意味を表し，後に述べる 4 つの役割からなる．

Lexical Inheritance Structure（継承構造）：　語が表す概念間の階層構造を記述する．

例えば，動詞 build と名詞 book の意味記述は次の通りになる．

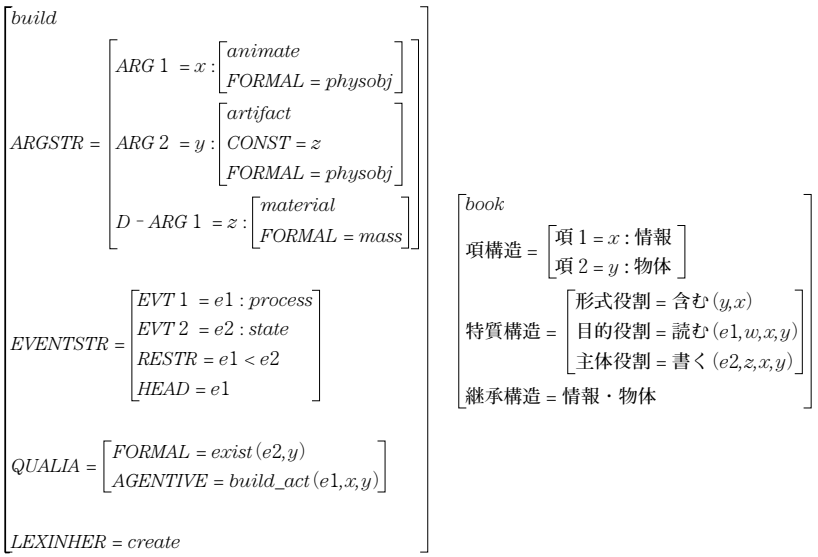


図 3　GLによる語の意味記述例

項構造は次の 3 種類に分かれる.

1） true argument

意味的にも統語的（英語の場合） にも必須なもの.

意味記述においては ARG 1 , ARG 2 , ... ARGn にて表す.

2） default argument（暗黙項）

意味的には必要だが統語的には必須でないもの.

意味記述においてはD-ARG 1 , D-ARG 2 , ... D-ARGnにて表す.

3） shadow argument（影項）

語の中に意味的に組み込まれてしまっているもので，特に情報を追加する場合以外は言及不要なもの

意味記述においてはS-ARG 1 , S-ARG 2 , ... S-ARGnにて表す.

事象構造は，出来事の時間的な流れの見方である．基本的には次の 3 種類がある.

表 2　事象構造

事象	説明	例
state	定常状態を表す	live，住む
process	一般的な動作	walk，歩く
transition	最終点のみ	fall，たたく

更には，幾つかの部分事象を持てるようになっている．そして，部分事象間の関係をRESTR（制約条件）にて表し，部分事象間の関係の主辞関係をHEADにより表す．RESTRは，大きく次の 4 種類を規定している．

表 3　事象間の関係

関係	図示記号
partial order	$\preceq$
strict partial order	$\prec$
overlap	$\circ$
inclusion	$\sqsubseteq$

なお，単語の意味記述の際には，単語が持っている時間的な流れを記載するのであるが，文として意味構造が組み合わせられた場合は，完了形などの文法的を用いて演算を行うことにより，事象が変化することもある．

6．SDRT を使った談話の意味表現

本章では，SDRTの概略について具体例を示しながら述べることにする．

まず，SDRTは，一階述語論理に基づいているので，定項と変項を使って記述を行う．なお，2つの項は，単一化（unification）という操作によって同一視される．つまり，一般的な変数への値の代入操作に該当することを，一階述語論理では単一化と呼ぶのである．特に，一旦，変項と定項との単一化が成立すれば，それ以降は，一貫して，その変項は，その定項として扱われるようになる．

6.1 SDRT による意味表現の例

次の 2 つの文があるとする．

John drives a car.　　（1）

It is red.　　（2）

まず，（1）の文のSDRT表記は次の通りになる．

$\pi 1$：

x, y, e1
john(x), car(y), drive(e1, x, y)

（3）

ここで，$\pi 1$ は命題（1）を表し，e 1 は事象を表す変項である．またjohn, car, driveは，各々，GL的な意味記述を行う必要がある．

次に（2）のSDRT表記は次の通りになる．

$\pi 2$：

z, i
red(z), ＝(z, i), it(i)

（4）

ここで，$\pi 2$ は命題（2）を表している．ここで，red, it は，各々，GL的な意味記述を行う必要がある．

そして（1）と（2）から構成される談話としてのSDRT表記は次の通りになる．

$\pi 1, \pi 2$
$\pi 1$： x, y, e1 / john(x), car(y), drive(e1, x, y)　　$\pi 2$： z, i / red(z), ＝(z, i), it(i)
?($\pi 1, \pi 2$)

（5）

なお，ここで ?（$\pi 1$, $\pi 2$）というのは，$\pi 1$ と $\pi 2$ の間に何か関係があることを表している．SDRTでは，考えられる関係として次が挙げられている．

表 4　命題間の関係

英語表記	説明
Elaboration	詳細
Narration	語り
Explanation	説明
Parallel	並列
Contrast	対比
Background	背景
Result	結果
Evidence	証拠

6.2 Dynamic Semantics

さて，（5）では ?（$\pi 1$, $\pi 2$）の ? は決まっていないし，変項 i が指し示すものも決まっていない．SDRTではDynamic Semanticsという考え方に基づき，世界知識や推論規則に基づく演算を行う．この一連の意味処理の様子を図 3 に示す．

自然言語処理においては，演算結果が唯一になるのではなく，複数の可能性が生じることが多い．これは一般には多義（曖昧さ）と呼ばれる現象で，多義の中から，可能性が高いものを選ぶメカニズムが必要になる．SDRTでは，このために，MDC（Maximize Discourse Coherence）と呼ばれる原理を採用している．基本的には，複数の可能性がある場合，変項の数が少ない方を採用する，ということである．具体的には，修辞関係により単一化の演算を行ったり，2 つの文の間の関係を考慮しながら談話としての意味生成を行う段階において単一化の演算を行ったりすることになる．

（5）では，変項 i が指し示すものの可能性として，johnとcarが考えられる．一般に，これは指示代名詞の曖昧さと呼ばれる．この場合，itのGL的な意味記述において，Argument structure（項構造）としてanimateではなくartifactを要求するように記述しておけば，変項 i は，変項 y と単一化することになる．一方，$\pi 1$ と $\pi 2$ の間の関係は，接続詞が明記されていない状態で

は定まらないが，世界知識から，Background又はNarrationであろうと推測される．その結果として，（6）を経て（7）の意味表現が得られる．

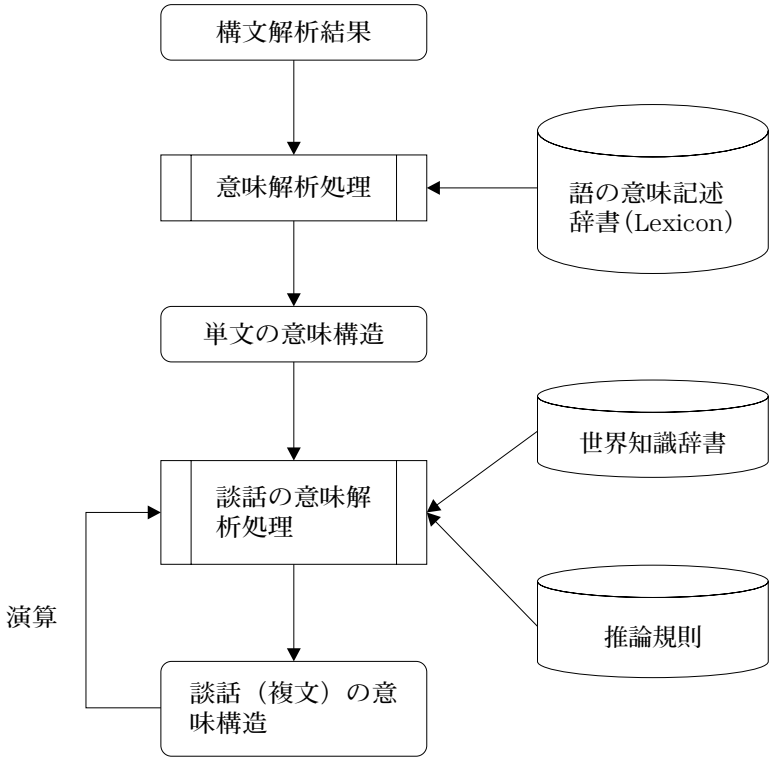


図 3　SDRT における意味処理の流れ

x, y, e1, z
john(x), car(y), drive(e1, x, y) red(z), =(z, y)

（6）

x, y, e1
john(x), car(y), drive(e1, x, y), red(y)

（7）

なお，もしも，第3文目や第4文目が追加される可能性があるのなら，この段階で無理して $\pi 1$ と $\pi 2$ の間の関係を決めておく必要はない．この段階では未定義のまま残しておいて，第3文目以降の意味追加時の演算処理に行われる推論処理に解決を委ねるのである．

6.3 簡略表記

なお，SDRTの表記は図示になっているので，論文中に記載するのは手間がかかる．そこで，本論文では，今後，以下のように簡略化して記すことにする．

$\pi 1$： [x, y, e1 | [john(x) ∧ car(y) ∧ drive(e1, x, y)]] （3）’

$\pi 1$, $\pi 2$ | $\pi 1$：[x, y, e1 | [john(x) ∧ car(y) ∧ drive(e1, x, y)]]
$\pi 2$：[z, i|[red(z) ∧ =(z, i) ∧ it(i)]]
?($\pi 1$, $\pi 2$) （5）’

x, y, e1 || [john(x) ∧ car(y) ∧ drive(e1, x, y) ∧ red(y)] （7）’

7．回答文の SDRT による意味記述

SDRT は，まだ発展途上であり，更に，SDRTを使用した日本語の意味記述の研究は（菊池，白井，2004）ぐらいしか見られない．本研究では，第3章で予備分析したe-Learning回答文について，実際にSDRTによる意味記述を試みる．但し，語のGLとしての意味記述は，後で必要に応じて行うことにして，基本的には割愛することにした．なお，使用したデータは，前出の学生からの

回答文から任意に抽出した２文である．

7.1 回答文の構文解析

SDRTにより意味構造を作成するためには，まず回答文を形態素解析し，次に構文解析して依存構造を生成する必要がある．前稿（吉武，2002）においては語を計算機処理して選び出す為に形態素解析のみがあれば十分と判断して，形態素処理プログラムChasenを使用した．今回の目的には，形態素解析だけでは不足するので，形態素解析の後の構文解析を行う必要がある．日本語の構文解析プログラムとしては，CaboCha／南瓜やKNP（黒橋，2000）がある．KNPは構文解析プログラムであり，形態素解析器JUMAN（中村，黒橋，’・” o，1994）と組み合わせて使用するように設計されている．今回の分析の目的には，どちらを使用しても大差ないと思われたが，前稿（吉武，2002）とは異なる解析プログラムを経験することも必要だろうと判断し，本解析ではKNPを使用することにした．なお，JUMANとKNPは，本学の情報処理センターのサーバ機solomonにインストールしたものを使用した．

なお，学生の回答はHTML文章になっているので，解析処理のためにHTMLコードを除去した．また，回答文には全て異なる固有番号３桁が付いている．これを回答文番号と呼ぶことにする．

7.2 サンプル文１

（原文のままの回答文番号039）

> ASCIIコードにカタカナをつけたものをJISX0201と言う，JISによる日本の漢字コードをJIS X 0208と言い，実用的なコードは三つシフトJIS，EUC-JP，ISO-２０２２-JPが存在する．JISX0201は半角カタカナを格納するASCIIコード部分にヨーロッパでは他の固有の文字が入っており，これが文字化けを起こす．

この回答文は第１文の「と言う」という述語の後に読点「,」が入っているが，これは正しくは句点「.」が入るべきであろう．これは単に回答者の入力

ミスであろうと判断し，句点「.」に変更した．また，全角数字と半角数字が混在しているので，数字を全て半角数字に変更した．また，漢数字の三が使われている個所を算用数字の 3 に変更した．更には，不要な空白を除去した．最終的に分析に使用した回答文は次である．句点で区切られた 3 つの文から構成された談話である．

（分析に使用した回答文番号039）

（039-1）ASCIIコードにカタカナをつけたものをJISX0201と言う．
（039-2）JISによる日本の漢字コードをJISX0208と言い，実用的なコードは 3 つシフトJIS，EUC-JP，ISO-2022-JPが存在する．
（039-3）JISX0201は半角カタカナを格納するASCIIコード部分にヨーロッパでは他の固有の文字が入っており，これが文字化けを起こす．

これらの文のSDRTによる意味記述は次の通りになる．

◎（039-1）の意味記述

π039-1：[x, y, e1, z | ASCIIコード(x) ∧ カタカナ(y) ∧ つける(e1, x, y, z) ∧ もの(z) ∧ JISX0201 (q) ∧ 言う(z, q)]

ここで，言う（z, q）は，発言するという意味ではなく，そのように呼んでいるという意味で使われているので，＝（z, q）に置き直すことにする．その結果，次が得られる．

π039-1：[x, y, e1, z, q | ASCIIコード(x) ∧ カタカナ(y) ∧ つける(e1, x, y, z) ∧ もの(z) ∧ JISX0201 (q) ∧ ＝(z, q)]

◎（039-2）の意味記述

助詞「の」については，（菊池，白井，2004）の考え方に従って，「の」の両側の名詞の間に何かの関係があるものと考え，関係 R（y, z）において R を推論することとした．

助詞「による」については，少し荒いかもしれないが，助詞「の」と同様に，

関係 R 2 (x, z) の R を推論することとした.

「3つシフトJIS，EUC-JP，ISO-2022-JP」という個所は，「3つ」つまりは「コード」の詳細化が行われていると判断し，a-part-ofという関係にて，「シフトJIS」と「コード」の包含関係，「EUC-JP」と「コード」の包含関係，「ISO-2022-JP」と「コード」の包含関係を表すことにした．その上で，「シフトJIS」と「EUC-JP」と「ISO-2022-JP」の合計が 3 であると解釈した.

その結果，(039-2) の意味記述は，次の通りになる.

π039-2：[m, n, o,q, a, b, c, h, i, jlJIS(m) ∧ R2 (m, o) ∧ 日本(n) ∧ 漢字コード(o) ∧ R(n, o) ∧ JISX0208 (q) ∧ ＝(o, q) ∧
実用的(a) ∧ コード(b) ∧ R3 (a, b) ∧ 3つ(c) ∧ ＝(b, c) ∧
SJIS(h) ∧ EUC-JP(i) ∧ ISO-2022-JP(j) ∧
a-part-of(h, b) ∧ a-part-of(i, b) ∧a-part-of(j, b) ∧
＝(sum(h, i, j), 3)]

◎ (039-3) の意味記述

(039-3) は重文になっており，「おり」という接続助詞が用いられている．「おり」の解釈としては，表 4 における命題間の関係のResultに該当すると思われるが，ここでは，単に論理積∧で結合することにする．その結果，(039-3) の意味記述は，次の通りになる.

π039-3：[r, s, e3, t, u, e4, v, w, e5 l 半角カタカナ(r) ∧ ASCIIコード(s) ∧ 格納する(e3, r, s) ∧ 他の(t) ∧ R(t, u) ∧ 固有名詞(u) ∧ 入っている(e4, u, s) ∧ 文字化け(v) ∧ JISX0201 (w) ∧ 起こす(e5, w, v)]

◎ (039) 全体としての意味記述

最終的に，この 3 つの文を組み合わせて 1 つの談話としての意味解析を行うことになる．その際の文の間の関係は 6 . 1 節で述べたように，複数の可能性がある．しかし，今回，実験対象とした回答文では，事実の説明という内容で

あるので，Narration（語り）という関係であると仮定し，論理積∧で結合することにした．結局，回答文番号039の意味記述として得られたのは次である．

π039：[x, y, e1, z, q, m, n, o, q, a, b, c, h, i, j, r, s, e3, t, u, e4, v, w, e5
| ASCIIコード(x) ∧ カタカナ(y) ∧ つける(e1, x, y, z) ∧ もの(z) ∧ JISX0201 (q) ∧ =(z, q) ∧
JIS(m) ∧ R2 (m, o) ∧ 日本(n) ∧ 漢字コード(o) ∧ R(n, o) ∧ JISX0208 (q) ∧ =(o, q) ∧
実用的(a) ∧ コード(b) ∧ R3 (a, b) ∧ 3つ(c) ∧ =(b, c) ∧
SJIS(h) ∧ EUC-JP(i) ∧ ISO-2022-JP(j) ∧
a-part-of(h, b) ∧ a-part-of(i, b) ∧a-part-of(j, b) ∧
=(sum(h, i, j), 3)
半角カタカナ(r) ∧ ASCIIコード(s) ∧ 格納する(e3, r, s) ∧ 他の(t) ∧ R(t, u) ∧ 固有名詞(u) ∧ 入っている(e4, u, s) ∧ 文字化け(v) ∧ JISX0201 (w) ∧ 起こす(e5, w, v)]

7.3 サンプル文2

（原文のままの回答文番号183）

◆日本の文字コードの種類◆
シフトJIS；日本のパソコン用の文字コードとして作られた．
EUC-JP；UNIX用
ISO-2022-JP；電子メール用（最上位ビットは0，つまり7 bitでおさまる）
◆半角カタカナと呼ばれるコードの問題点◆
電子メールに，JISX0201の8単位符号表の右半分が入っていると，表示の際に文字化けを起こすものが多い．
※インターネットの世界では使用禁止．

まず，記号◆と※を除去した．そして，箇条書きの1行が1つの文であると仮定した．全角数字と半角数字が混在しているので，数字を全て半角数字に変

更した．また，セミコロン ； は，主語を示す助詞の代りに使われていると仮定した．括弧で囲まれた文は，括弧から取り出して単独な文として取り扱った．その結果，分析に使用した回答文は，次の 8 つの単文から構成された談話となる．

（分析に使用した回答文番号183）

（183-1）日本の文字コードの種類
（183-2）シフトJIS；日本のパソコン用の文字コードとして作られた．
（183-3）EUC-JP；UNIX用
（183-4）ISO-2022-JP；電子メール用
（183-5）最上位ビットは 0 ，つまり 7 bitでおさまる．
（183-6）半角カタカナと呼ばれるコードの問題点
（183-7）電子メールに，JISX0201の 8 単位符号表の右半分が入っていると，表示の際に文字化けを起こすものが多い．
（183-8）インターネットの世界では使用禁止．

これらの文のSDRTによる意味記述は次の通りになる．

◎（183-1）の意味記述

π183-1：[x, y, z | 日本(x) ∧ 文字コード(y) ∧ R1 (x, y) ∧ 種類(z) ∧ R2 (y, z)]

なお，日本語の構文的には，［日本の］が［文字コード］に係るのか，それとも，［種類］に係るのか，という曖昧さがあるのだが，「読点が入っていない場合は直近の語に係る」という暗黙の了解を適用することにした．つまり，［日本の］は［文字コード］に係るとした．

ここで，助詞「の」の意味については何人もの日本語文法学者が研究をしており，数多くの解釈が存在することが知られている．（菊池，白井，2004）においては，助詞「の」の意味として次の 7 つの可能性を示している．

表 5　「A の B」の解釈

解釈	説明	例
名詞 B で表される関係	名詞 A がその項	太郎の兄
名詞 B で表される事象	名詞 A がその項	言葉の理解
名詞 B に固有的な関係・事象	名詞 A がその項	トヨタの車
名詞 A と名詞 B が独立に個体，事象のタイプを限定	文脈から名詞 B が個体と解釈される場合，名詞 A は付加的情報をもたらす	美人の母，日課の散歩
所有・所属	名詞 A が個体や組織の固有名で，名詞 B は人間・事物	花子の本
属性的	名詞 A が時間・場所・順序	東京のビル
文脈によって決まる任意の関係		太郎の車

但し，計算機処理の立場では（183-1）の文を計算機処理している段階では，この 7 つの内のどれであるかを決めるのは難しい．談話の意味を生成する段階におけるMDC処理に委ねることにする．

◎（183-2）の意味記述

（183-2）では，「A用のB」という日本語の意味解釈を行わなければならない．ここでは，表 5 の所有・所属という解釈と同様に，下記の解釈を行う．

表 6　「A 用の B」の解釈

解釈	説明	例
所属	名詞 A が個体や組織の固有名で，名詞 B は事物	パソコン用の文字コード

これに従って，(183-2) の意味記述は，次の通りになる．

π 183-2：[f, g, h, i | シフトJIS (f) ∧ 日本(g) ∧ パソコン(h) ∧ R3 (g, h) ∧ 文

字コード(i) ∧ 用の(h, i) ∧ 作る(e1, f, i)]

◎ (183-3) の意味記述

(183-3) では,「UNIX用の？」つまり,「の？」部分が省略されていると考え，変項 ？ を導入する．その結果，(183-3) の意味記述は，次の通りになる．

π183- 3：[j, k, m | EUC―JP(j) ∧ UNIX(k) ∧ ？ 1 (m) ∧ 用の(k, m)]

◎ (183-4) の意味記述

(183-4) の意味記述も，同様に，次の通りになる．

π183-4：[n, o, p | ISO-2022-JP(n) ∧電子メール(o) ∧ ？ 2 (p) ∧ 用の(o, p)]

◎ (183-5) の意味記述

(183-5) は重文になっており,「つまり」という接続助詞が用いられている．「つまり」の解釈としては，表 4 における命題間の関係のElaboration又はExplanationに該当すると思われるが，ここでは，単に論理積∧で結合することにする．その結果，(183-5) の意味記述は，次の通りになる．

π183-5：[q, r, s, e2 | 最上位ビット(q) ∧ 0 (r) ∧ ＝(q, r)∧ 7 bit(s) ∧ おさまる(e2, q, s)]

◎ (183-6) の意味記述

(183-6) では「呼ばれる」つまり「呼ぶ」の受身形の意味記述を考える必要がある．

π183-6：[t, u, v, e3 | 半角カタカナ(t) ∧ コード(u) ∧ 呼ぶ(e3, u, t) ∧ 問題点(v) ∧ R4 (u, v)]

ここで，呼ぶ (e3, u, t) は＝(u, t) と同じ解釈が可能だと仮定すると，最終的には次の意味記述が得られる．

π183-6：[t, u, v, e3 | 半角カタカナ(t) ∧ 問題点(v) ∧ R4 (t, v)]

◎ (183-7) の意味記述

(183-7) は重文になっており,「と」という接続助詞が用いられている．「と」

の解釈としては，表 4 における命題間の関係のResultに該当すると思われるが，ここでは，単に論理積∧で結合することにする．また，量子化の問題は取り扱わないことにしているので，「ものが多い」の意味記述は無視する．その結果，(183-7) の意味記述は，次の通りになる．

π183-7：[a, b, c, d, e4, ff | 電子メール(a) ∧ JISX0201 (b) ∧ 8 単位符号表(c) ∧右半分(d) ∧ R5 (b, c) ∧ R6 (c, d) ∧ 入っている(e4, d, a)∧ 起こす(e 4 , a, ff) ∧ 文字化け(ff)]

◎ (183-8) の意味記述

(183-8) では，助詞「では」の意味記述が必要になる．ここでは，単に論理積∧で結合することにする．その結果，(183-8) の意味記述は，次の通りになる．

π183-8：[a1, b1, c1, e5 | インターネット(a1) ∧世界(b1) ∧ R 7 (a1, b1)∧ 使用(c1) ∧ 禁止する(e5, d1, c1)]

◎ (183) 全体としての意味記述

最終的に，8 つの文を組み合わせて談話の意味記述を生成する場合，まず，(183-1) の日本(x)と (183-2) の日本(g)がunifyするので，x と g は同一であると判断される．同様に，文字コード(y)と文字コード(i)がunifyする．文の間の関係は (039) と同様にNarration（語り）という関係であると仮定し，論理積∧で結合した．結局，回答文番号183の意味記述として得られたのは次である．

π183：[x, y, z, f, h, j, k, m, n, o, p, q, r, s, t, v, a, b, c, d, ff, a1, b1, c1, e1, e2, e4, e5 | 日本(x) ∧ 文字コード(y) ∧ R1 (x, y) ∧ 種類(z) ∧ R2 (y, z)] ∧
シフトJIS (f) ∧ パソコン(h) ∧ R 3 (x, h) ∧ 用の(h, y) ∧ 作る(e1, f, y) ∧
EUC-JP（j） ∧ UNIX（k） ∧ ？1 （m） ∧ 用の （k, m)] ∧
ISO-2022-JP(n) ∧電子メール(o) ∧ ？2 (p) ∧ 用の(o, p)] ∧

最上位ビット(q) ∧ 0 (r) ∧ =(q, r)∧ 7 bit(s) ∧ おさまる(e2, q, s)] ∧
半角カタカナ(t) ∧ 問題点(v) ∧ R 4 (t,v) ∧
電子メール(a) ∧ JISX0201 (b) ∧ 8 単位符号表(c) ∧右半分(d) ∧ R5 (b, c)
∧ R6 (c, d) ∧ 入っている(e4, d, a) ∧ 起こす(e4, a, ff) ∧ 文字化け(ff) ∧
インターネット(a1) ∧世界(b1) ∧ R7 (a1, b1) ∧ 使用(c1) ∧ 禁止する(e5, d1,
c1)]

7.4 類似度を推定する実験

以上の意味記述の結果に対して，妥当性の検討を行うために，複数の回答文の間の意味記述の類似度を推定する実験が必要になると考えている．SDRTとGLは，リスト表記が可能であるので，リスト表記の処理が得意なプログラミング言語（LispかProlog）を用いた類似度推定実験を計画している．以下に，その概要を書いておく．

回答文の類似度は，SDRTにより出力した意味表現の類似度計算が必要になる．SDRTにより出力した意味表現は，項を論理積（AND）でつないだものとなっている．4.2節で述べた類似度計算の考え方を参考にして，SDRT中の項の一致率を類似度として採用することにする．つまり，類似度similarityとしては一致率を下記のように定義する．

$$\text{類似度} = \text{一致率} = \frac{C \times 2}{A + B}$$

ここで，A, B ： 項の数
　　C ： 項が unify した数
もしAとBが同じ数で，更にunifyする場合はsimilarity＝100% になる．

7.5 検討を要する点

SDRT自身が発展途上にあり，全ての自然言語の意味記述が定義されている訳ではない．しかしながら，**MDC**という考え方は，筆者が過去に使用したモ

デルでの問題点を克服する可能性を持っていると思う．更には，**SDRT**による意味記述を計算機上にて実現した研究は，未だ，見あたらないため，今後の計算機上での処理プログラムの実現は意味があることであると思われる．

今後，検討すべき点としては次が考えられる．

・量化子（quantifier）や時制などの取り扱い．
・文の間の関係をNarration（語り）と仮定したのは荒すぎないのか？
・HTMLコードを除去してから意味記述を考えたが，回答者がHTMLコードに含めた意味は存在しないのかということの検討が必要．

更には，類似度計算においては下記の問題が想定される．

・述語の否定は，どう対応するか？
・反対の意味の述語は，どう対応するか？
・意味の包含関係は，どう対応するか？
・類似度の計算式は妥当か？

8．あとがき

本研究は，前項（吉武，2002）に引き続いて，処理する入力対象にプリエディティング（前処理）を行わずに計算機処理を行わせ，どこまで有用な結果が得られるのか，という研究の第 2 段目である．本稿では，まず，実際にe-Learningシステム上で学生が回答した文を解析し，色々な回答形態を処理するためには意味処理が必要であると判断した．その結果を踏まえて，研究の根幹をなすモデルの検討と，そのモデルを用いた意味記述の試みを行った．引き続き，本意味記述の妥当性の検討を行う実験を進めたい．

なお，本研究は，西南学院大学から与えられた2006年度在外研究（A）により滞在したオーストラリアのシドニーにあるMacquarie大学のコンピュータ学科にて行ったものである．在外研究の機会を与えて下さった西南学院大学に感謝すると同時に，私を受け入れてくれたMacquarie大学のProfessor Robert Dale に深く感謝します．

参考論文

Asher Nicholas, Alex Lascarides (2005). Logics of Conversation. Cambridge University Press.

Pustejovsky James (1995). The Generative Lexicon. MIT Press.

黒橋禎夫（2000）．“結構やるな，KNP.”情報処理41巻11号，情報処理学会．

中村俊久，黒橋禎夫，長尾真（1994）．“部分文字列情報の利用による日本語単語の高速検索.”情報処理学会研究報告．自然言語処理研究会．1994-NL-101．

菊池隆典，白井英俊（2004）．“文脈に応じた名詞句の意味解釈の検討.”第 6 回年次国際大会ハンドブック 言語科学会．

横田将生，吉武春光，田町常夫（1986）．“自然言語理解システム IMAGES-I の意味解釈過程について.”電子通信学会論文誌 D J69-D巻 5 号．

吉武春光（2002）．“電子メールからのナレッジ・リサイクルの試み.”西南学院大学商学論集 49 巻 2 号，

荒牧英治，黒橋禎夫，柏岡秀紀，加藤直人（2005）．“確率的用例ベース翻訳の実現.”第11回年次大会．言語処理学会

高橋哲朗，乾健太郎，松本裕治（2002）．“テキストの構文的類似度の評価方法について.”情報処理学会研究報告．自然言語処理研究会．2002-NL-150．

山下達雄，富士秀，大倉清司，潮田明（2003）．“翻訳支援に有効な訳例検索の類似度計算方式と検索結果提示方式.”第 9 回年次大会．言語処理学会．