

スクリプトの 文章からの帰納的な学習

堀浩一 斎藤忠夫 猪瀬博
(東京大学工学部)

A method of inductive learning of scripts from natural language texts is given in this paper. A script is a stereotyped structure of contents of texts. An experimental system to learn scripts from texts called SEM (Script Extractor Mechanism) is described. The results of experiments of learning scripts from the texts of newspaper and of applying the learned scripts to classify texts are presented to testify the effectiveness of SEM.

1. まえがき

自然言語処理の研究は着実に成果をあげ、限られた分野の、限られた形式(ナイトル、抄録など)の文章については、機械翻訳などの応用が実現されつつある。しかしながら、今後、第5世代計算機の研究成果なども大きくなり、自然言語処理システムを、広く知識情報処理システムの一環としてとらえる時には、まだ残されていける問題が多い。特に、文章の意味の理解の問題は、いずれ避け難くなることのできぬ問題である。機械翻訳においては、表層構造の取扱いに重点を置いても、限られた应用では充分实用に供する事が期待できるが、抄録の作成、知識ベースやデータベースとのインターフェイスなどの应用においては、自然言語処理システムそのものが充分な知識をもつていて、文章の意味を理解できるようにならなければならぬ。意味の取扱いが重要であり、自然言語理解システムに知識を付与することにより、その取扱いが可能となるとするところは従来から重ねて指摘されてきた。それにもかかわらず、期待されるほど研究が進展しないのは、自然言語理解システムに与えた知識作成の困難さに起因していると筆者たちは考えた。筆者たちは自身、以前に、論文抄録の文章が内容の定型的な構造をもつていて、その構造の知識を利用して、抄録の文章からキーワードの抽出を行なうシステムを作成し、実験した。このシステムは、かなり良好な実験結果を示したが、知識の作成、修正、保守の困難のために、大規模な实用システムの作成には至らなかった。

この問題を解決するためには、自然言語理解システムが必要とする知識を文章そのものから学習システムによって学習することを考える必要がある。自然言語理解システムの必要とする知識には、辞書、文法規則、対象分野の専門知識、スクリプトやシナリオなどと呼ばれる文脈の構造に関する知識などが含まれる。これらすべての知識について、作成や保守の手法を考えなくてやかねばならぬ。本論文では特に、文章の内容の定型の知識があるところのスクリプトを、文章からの帰納的に学習する手法について論ずる。スクリプトが自然言語理解システムにとって有効であることはよく知られており、また辞書や文法規則に比べれば変化に乏しいので实用的価値のある学習結果を得ることが可能であると期待されるからである。2章で、スクリプトの表現法と利用法と帰納的学習の方法について論じ、作成した実験システム SEM (Script Extractor Mechanism)について説明する。3章で新聞記事の文章を対象とした実験について述べ、スクリプトの帰納的学習の有効性と限界を示す。

帰納的学習は、実例を与えることにより、実例が共有していける性質あることは原理と学習させようとするものであり、従来、帰納的推論のアルゴリズムの研究、類推の理論の研究³⁾、Meta-DENDRAL⁴⁾やBacon⁵⁾などの特定の分野の学習システムの作成、文法の推論の研究などに行われてきている。従来の研究については文献⁶⁾によくまとめられており、帰納的学習へ一般論に重点を置いたシステムでは実用性に欠け、実用性に重点を置いたシステムでは、問題特有の知識如学習プログラムにうめこされてしまい、実用性に欠けるという傾向をもつ。本論文で提案するSEMは、実用性をめざしつつ、問題特有の知識は分離して記述することにより汎用性、拡張性を確保している。また、学習の目標であるスクリプトは、正解が1つに定まることではなく、スクリプトの使用目的に応じて、別形のスクリプトが必要なことも考えられる。そこで、帰納的学習を行うための規則もプログラムとは分離して記述することにより柔軟性をもつように考慮している。

2. スクリプトの表現法と利用法

2.1 スクリプトの表現法と利用法

スクリプトは、当初、Schankにより提案された、レストランでの食事などの、人間の日常の行為の定型の記述である⁷⁾。同じくSchankにより、スクリプトを自然言語の文章に適用して文章を理解するSAM(Script Applier Mechanism)と称するシステムが作成された⁸⁾。Schank自身はその後、スクリプトはダイナミックに構成されるべきであるとして、その構成要素としてMOPs(Memory Organizing Packets)という考え方を提案している⁹⁾。本論文では、スクリプトという用語は文章の内容の定型の記述という意味で用い、必ずしもSchankの研究にはとらわれないことにする。最終的には、Schankと同様に、MOPsとスクリプトの階層を考へ、複雑な理解システムを構成することを目標とした。しかし、MOPsの学習やスクリプトのダイナミックな構成の研究は未知の分野であり、まずは理解システムの行う仕事をとしては簡単なもので設定して、それに必要な比較的簡単なスクリプトの学習の研究から着手すべきである。そこで、本論文では、スクリプトを用いて文章の分類を行なうこと、自然言語理解システムへ行う仕事をして設定する。分類の次に、キーワードの抽出システム、その次に抄録の作成システムなどを想定している。

システムの全体像は図1のようになる。たとえば、新聞記事の文章を、日米経済摩擦に関する記事とそうでない記事に分類したいとしよう。すると、ユーザは、SEM(Script Extractor Mechanism)に、日米経済摩擦の記事の実例をいくつか与える。SEMが帰納的学習を行って、日米経済摩擦とはいうものだとうスクリプトを出力する。SAM(Script Applier Mechanism)が、そのスクリプトを新しい文章に適用して、日米経済摩擦にあてはまるかどうかを判断する。ここで、SEMの出力するスクリプトは直接人間が見て意味わかるものでなければならぬとする。

スクリプトの表現方法としては、筆者らの提案したH-net(Hierarchical network)と称する知識表現法¹⁰⁾を用いる。H-netは、概念の外延と内包を論理的に表現す

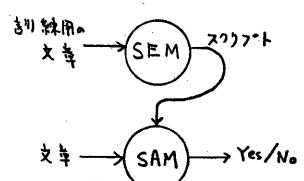


図1：システムの全体像

る知識表現法である。H-net を用ひると、スクリプトからシソーラスまで一貫して論理的表現ができる、表現の意味を明示することができる。たとえば、日本経済摩擦のスクリプトが、「アメリカが日本と非難した、アメリカは**をさだめた。」XXは経済問題である。日本は**の対応とし、た。」と「うものごと、たとすると、このスクリプトは図2のように表現できる。さらに、経済問題とはどういうものか、対応とはどういう動作かと「うシソーラスや、格構造のような動詞に隣りる知識が存在して、やはり H-net で表現されるが、それをはすじめ与えられた「うもの」とす。図2で *text は、文章にバインドされる変数である。(非難 *text アメリカ 日本)は、*text 中に、アメリカが日本を非難したという記述があることを示している。本来、スクリプトは時間的前後関係を含んでいたが、新聞記事の文章では、できごとの時間関係と文の前後関係は一致せず、いざいざな順序で出現するので、図2のスクリプトには、時間の前後関係は含んでいない。時間の前後関係を扱うには、各述語に、調べた残りの文書をバインドする変数を付加して次々に理解するようにすればよい(たとえば、(非難 *text *rest1 アメリカ 日本)(言う *rest1 *rest2 アメリカ XX)というようになります)。

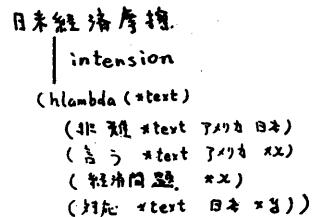


図2 スクリプトの H-net による表現の例

2.2 帰納的学習の方法

スクリプトは次の手順のステップをへて学習される。

Step 1. 文章を愚鈍な理解システムで理解してイベント・ストラクチャと呼ぶ構造を得る。

Step 2. イベント・ストラクチャ相互のマッチングをとる。(3つめ以後の文章については、文章のイベント・ストラクチャとスクリプトのマッチングをとる。)

Step 3. マッチングのための部分とそれなりの部分をルール・セレクタに渡し一般化のためのルールを選択する。

Step 4. 一般化のルールを適用して、構造を一般化する。

Step 5. スクリプト再構成のためのルールを適用してスクリプトを再構成する。

Step 1 から Step 5 をくり返す。

各ステップの詳細を例を用いて説明する。

今、「Aと「う人がBと「う人を殺し、Aと「う人は自殺した」という新聞記事の文章のスクリプトを学習させた」とする。ユーザは図3に示すような実例を入力する。これを愚鈍な理解システム(parserといつてもよい)が解析して、図4のようないベント・ストラクチャを出力する。ここでは、イベント・ストラクチャは、述語、主語、目的語からなる構造としておく。(理解システムが賢く店子ハッカのイベント・ストラクチャも精密にすることは将来の課題とする。)また、照應の取扱いは、愚鈍な理解システムはできないものとする。次に、このイベント・ストラクチャどうしのマッチングを述語名につけてとる。(3つめ以後の文章とスクリプトとのマッチングの場合には文章のイベント・ストラクチャがスクリプトに含まれるかどうかを調べる。)この例の場合、(殺す 調理師 幸枝さん)と(殺す 自分 二郎)

山口県須田郡須田町三丁目九家
電気自動車販売業者某(63)
が二月十四日出勤したばかりの方不
幸な事故で死んでしまった。同日午後四時
前後、彼の母親が電話で来る。内閣府の調査課が「交通事故で死
して須田市に運びだされた」という通報を
受け、着つて調査しているので須田
警察が連絡がある。
同様に彼は死体運送事件
として須田署へ運搬。死体の運
送料金を支払ってから、
お通夜があるなど

(a)

(b)

図3 入力例（日本經濟新聞（1933年）の記事67）

などからマッチする。マッチした部分としなかった部分がまとめて、ルールセレクタと呼ばれる部分に渡される。ルールセレクタは構造・一般化の規則を選択する部分である。一般化の規則には、大きく分けて、general-rules & case-by-case-rules の2種類がある。各々の規則はアーリーアクション規則のようないくつか

さて、やは「H-net を用いて記述されることは。general-rules は、定数と変数にあきかえる、2つの概念を上位の概念にまとめあげる、などの一般的な一般化の方法を記述した規則である。case-by-case rules は、たとえば、「飲む」という述語は睡眠薬または青酸カリを飲んだ場合のみ「自殺」に一般化することができる、などと「う概念ごとの一般化の方法を記述した規則である。ルールセレクタは、出力するスクリプトかどのようなスクリプトか最適かの評価基準によると、複数の適用可能な規則の中から1つの規則を選出することを意図してあるが、現在実装されたシステムでは、case-by-case-rules, general-rules の順にさかれて、最初に見つけた適用可能な規則を選んで出している。現在、実装された general-rules は、定数と変数に変えた規則、述数とタイプ付き変数にあきかえる（たとえば、神戸と大阪を対象にし、(city xx) を付け加える）規則、述語を上位の概念にあきかえる（たとえば、(刺殺 a b) と (殺す c d) を (殺す e f) に一般化する）規則の3種類である。ルールの記述例は次節に示す。上位概念の探索方法が問題であるが、各概念をシソーラス上で一段だけ一般化した概念すべてを、単純に探索していく。ただし、出現する全述語どうしにツリエントを調べると時間がかかるので、述語名ヘマッチした述語の周辺どうしのみについて調べようとしている。次に、ルールセレクタの選出したルールを構造に適用することにより、スクリプトが得られる。上の例の場合、図5のようなスクリプトが生成される。最後に、再構成のステップとして、出力されたスクリプトを検査し、矛盾の除去、簡約化などをを行い、

国5 出力スクリプトの例

a)

(b)

(出勤 江利事務課人 ?)
(行方不明 江利事務課人 ?)
(經營 母親 ?)
(殺手 調理師 ?)
(強盗 調理師 遠吉)

(くつた) 斧頭一枚 ?)
(死 =人 ?)
(見つけた 長女 死)
(馬鹿の子 又人 ?)
(110 倍 又人 ?)

図4 イベントストラクチャ

最終的なスクリプトが得られる。この再構成のステップは現在はまだ実装されていない。

以上の説明から明らかなように、SEM では、スクリプトにおける定例だけを訓練例として与えて学習を行う。あとは定例の例についても教える方法を考えるかと思われる。Winston の「ランニアミスを与えた必要があり」³⁾ 実際の文章からニアミスを選んで与えるのは面倒である。また、人工的なニアミス例をつくるとすると、それは正解と知っているからできることであり、本論文の趣旨に反する。あとは定例だけからの学習を図式的に示すと図6のようにある。Xを定例、実線の円を正解、破線と SEM への出力とする。Xを囲むように、破線の円を少しづつ広げて、正解に近づけようとするのが SEM の学習のやり方であると言える。

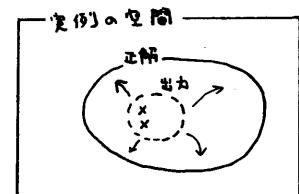


図6 SEMの学習の図式的説明

2.3 実験システム SEM (Script Extractor Mechanism)

前節で述べた学習を実現するための、SEM 1.5 と称する実験システムを、東京大学大型計算機センタ上に、Franz Lisp と筆者らの作成した H-Net (H-net Representation Language)¹⁰⁾ を用いて実装した。

SEM 1.5 の構成を図7 に示す。実際に実装されることは、図中で、実線で囲んだ部分である。dull-understander は現在は実装されておらず、人間がイベント・ストラクチャを作成している。また、reconstruction のステップは素通りしている。四角で囲んだ部分は Lisp による、丸く囲んだ部分は H-Net による記述これである。

matcher は出現した述語名にマーカーをつけることにより、同じ述語名の述語を見つけ、マーカした述語との周辺の述語からなる構造と rule-selector に渡す。

rule-selector は前節に述べたとおり、述語を一般化するためにルールを generalization-rules の中から選択する。

generalization_rule の例
x 1 z, 述語と上位概念の一般化子規則の記述を図8 に示す。「睡眠薬を飲む」を「飲む」一般化する case-by-case-rule を図9 に示す。概念の上下関係の知識(クラス)もまた H-Net で記述されるところ、概念の上下関係には ISA の関係(クラスとサブクラスの関係)と Class-Instance の関係の2種類がある。図10(a)のような外部表現で概念の上下関係を与えて図10(b)のような H-Net 表現に変換されて蓄えられる。

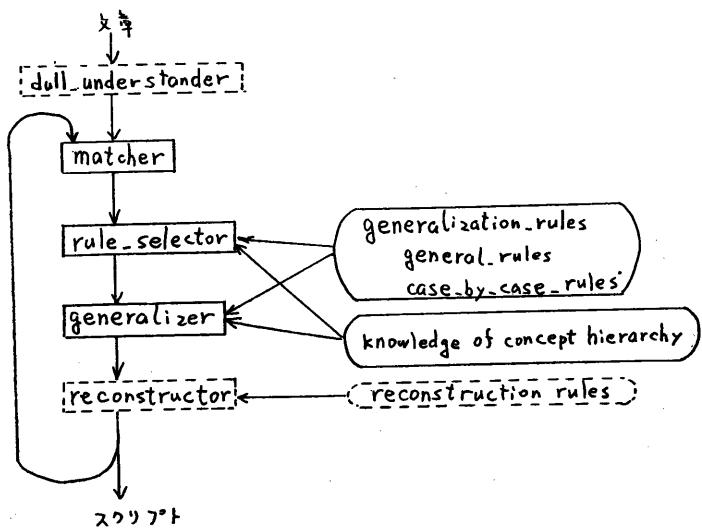


図7 SEM 1.5 の構成

aufheben_pred

```

| intension
( hlambda ((P : xv)(xg : xw))
  (parents *P *PP)
  (parents *g *gg)
  (conjL *PP *gg *r)
  (aufheben_vars xw *w *uv *types)
  (hreturn `((xw , *uv) , *types)) )

```

図8 general_rule → 151

generalizer 13 -

一般化の規則を HNRL

← 関数 deduce に 13'

演繹することによる

一般化を行う。たと

えば、(deduce ((aufheben

-pred (太郎 次郎 次郎)

(刺す 三郎 花子)))) にす

り、(暴力 *x *y)

が得られる。generalizer は定数と変数へ置きかえのリストをもつ、これで、同じ定数は同じ変数へ置きかえられる。deduce は Prolog と同じように top-down, depth-first の推論を行い失敗したらバックトラックする。

SAM は HNRL による容易な実現される。スクリプトを deduce あるだけでよい。実際のシステムでは、図2に示した *text という変数日用“す”，1つの文章のイベント・ストラクチャ中の各述語を一時的に外延として付加しておいて、スクリプトを deduce するという方法をとっている。たとえば、イベント・ストラクチャ中に (刺す 太郎 次郎) と “刺す”述語がある、たとえ、(太郎 次郎) を「刺す」の外延として一時的に宣言して、スクリプトを演繹すれば、スクリプト中の (刺す *x *y) が (暴力 *x *y) が成功し、変数がバインドされる。こうして、スクリプト中のすべての述語の演繹に成功すればスクリプトにおけるまとまること例があり、1つでも失敗すればスクリプトにおけるまとまることの判断がなされる。

3. 実験と評価

スクリプトの帰納的学習の有効性を評価するための実験を行った。実験対象としては、日本経済新聞の1月から5月までの新聞記事を、殺人、自殺、心中、火事、転落事故に関するもの42件を用いた。はじめ、これらの記事を池の6つのグループに分類する。(a)殺人、(b)無理心中、(c)自殺、(d)殺人、(e)焼死、(f)転落死。この分類は、キーワードだけでは選別できないようなお互いに接近した微妙な分類であり、等者か自分でもは、きりした基準をもたずには、それがどの文章を読み直感的に分類を行った。これは、SEM がこの直感的分類と一致した分類を行えるようなスクリプトを出力できるかどうか評価しようとするためである。表1に、使用した記事の見出しと参考のため示す(実験では本文だけを用い、見出しが用いられない。)

SEM の出力したスクリプトを用いた SAM による分類の実験結果を図10に

飲む_rule

| intension

```

( hlambda ((飲む *x 飲眠薬)(xg : xw))
  (parents xg *gg)
  (or (eg xg '白殺)(meng '白殺 *gg))
  (aufheben_vars (xg 飲眠薬) *w *uv *types)
  (hreturn `((白殺 , *uv) , *types)) )

```

図9 case-by-case_rule 151

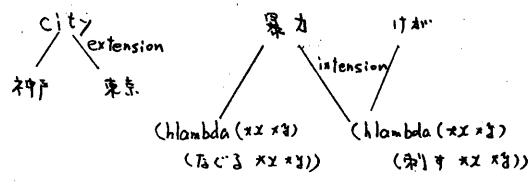


図10 概念の上下関係の表現

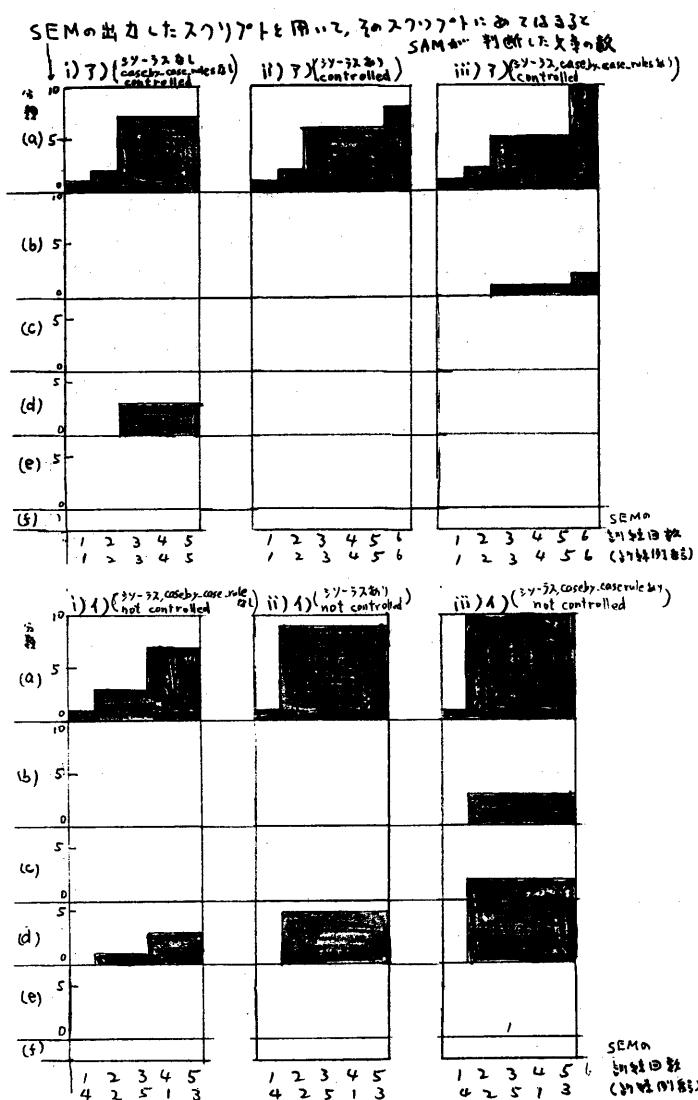


図11 実験結果

分類	番号	見出し
殺人として自殺	1	女優殺し正犯として自殺判決
殺人として自殺	2	尼母とパート殺人
殺人として自殺	3	2児の愛し川へ投入
殺人として自殺	4	借金高、心中自殺
殺人として自殺	5	津で4男と7女
殺人として自殺	6	母親の使用者自殺
無理心中	7	都立高生が母親刺殺
無理心中	8	刀で金吉2児殺し自殺回復
無理心中	9	生徒殺した妻と殺人未遂
無理心中	10	子供殺し(飛行機)
無理心中	11	4歳と母子3人電車転落死
無理心中	12	海へ心中で1死
無理心中	13	3児と母が死んで
無理心中	14	3児とお母が竹火
無理心中	15	始業式前日に悪魔心中
無理心中	16	93重苦、心中ドライ
無理心中	17	愛知で6死
無理心中	18	親子3人が掛合の心中
無理心中	19	無理心中で3人焼死
無理心中	20	2兄弟殺傷二重死
自殺	21	京都府最大級強姦殺
自殺	22	小6少女が自殺
自殺	23	下見の豪傑生自殺
自殺	24	参考人の社長自殺
自殺	25	カソリンをかき自身首自殺
自殺	26	デュトール男自殺
自殺	27	マンホールに男の金棒死体
殺人	28	基地で運転手刺殺
殺人	29	結婚反対で刺殺
殺人	30	訴訟相手を刺殺
殺人	31	隣生、本縁の凶刃
殺人	32	ホーランド人殺人
殺人	33	養育費の殺す
焼死	34	母子3人が焼死
焼死	35	父婦3人、母子5人焼死
焼死	36	中華店焼3人死焼死
焼死	37	一老5人焼死
焼死	38	幼い3姉弟死焼死
焼死	39	母子2人焼死
焼死	40	孫女放火、毛女死焼死
死因	41	伴人暴死、バス転落
死因	42	飲酒の車3軒死

表1 実験に用いた新聞記事の見出し

示す。これは「殺人として自殺」(分類(a)) のスクリプトを準備させた結果である。SEMに「殺人として自殺」の記事を1つずつ与えて訓練を行ふ。そのたびに、出力されるスクリプトと42件のすべての記事をSAMに与えて、SAMが42件のうちのいくつを「殺人として自殺」の記事として分類できたか調べた。横軸にはSEMに与えた記事の個数とそれを与えた順番を表す番号が示している。たて軸には、表1にそれまでの分類の記事の中で、「殺人として自殺」の記事としてSAMにより分類された記事の数を示している。理想的には、(a)の欄のみがぬりつぶされることになる。実験方法としては6種類心比較されてる。i)はシーケンスもcase-by-case-ruleも与えなかつた場合、ii)はシーケンスのみ与えた場合、iii)はシーケンスと

case-by-case の両方を与えた場合である。3)は訓練例を与えた順序と教師がコメントした（典型的と思われる例から順に与えた）場合であり、1)はランダムに与えた場合である。

出力されたスクリプトへ例として1), iii)3)の場合、訓練回数5で出力されたスクリプトは ((boryoku *v1 *v2)(jisatsukonkei *v3 *v4)(ishop *v5 *v6)(korosu *v7 *v8)) である。たゞ、ii), iii) で与えたシソーラスには、図10(a)のような関係が114組入っているおり、iii) で与えた case-by-case_rules の数は5つである。同じ知識と規則を用いて(b) ゲループのスクリプトを学習して分類する実験も行、たゞ、ほぼ同様の結果を得られた（すなわち、(a) ゲループ専用に微調整した知識とルールではない）。(b) ゲループのスクリプトとし1), ((iku *v1 *v2)(shi *v3 *v4)(kurushimi *v5 *v6)) などが出力された。SEMの1回の訓練はCPU timeで約10秒、SAMの1つの文章の判断は CPU timeで約1秒かかる。

図11をみると、訓練例を与えた順番をランダムにすると、スクリプトが一般化されすぎることわかる。シソーラスや case-by-case rules を与えるとかえり、その結果が悪くなる場合もあるのは、SAMのようだ。DRL概念に該当する下位概念が小えたため、理解が成功する率が上から3つ目である。図11のiii)3)をみると、ほぼ満足できる学習が行えたといえる。しかし、上述した出力スクリプトで、本当は自殺を殺人を行、たゞ3とv7が等しいはずだが、dull-understanderが想応の取扱を行、2つ目ため、これが検出されていない。

4. 結論

自然言語理解システムの用いる知識の学習システムの必要性を主張し、スクリプトを文章から帰納的に学習する手法と実験システムを与えて、実験による有効性を示した。本論文で用いた学習の手法自体は、人工知能の分野では特に目新しいものではないが、汎用性と有用性へのねらいに配慮した実験システムを作成し、実際に、自然言語理解システムの用いる知識の学習への可能性を示したことは意義があると考える。今後、もとと複雑なスクリプトを掌握できるシステムを構築するためには、SAMからSEMへのフィードバック、スクリプトの評価基準の導入、評価基準に基づく一般化と制御すること、disjunctionとスクリプト中に含まれること、照応を取り扱う規則とその学習などを考えてゆかねばならぬのである。

参考文献

- 1) 猪瀬常彦:シリオと用いる論文抄録理解・作成技術入門、情報処理学会論文誌、Vol. 24, No. 1 (1983)
- 2) Shapiro,E.Y.:Inductive Inference of Theories from Facts, Research Report, No. 192, Yale Univ. (1981).
- 3) Winston,P.H.:Learning and Reasoning by Analogy, Comm.ACM, Vol. 13, No. 12, pp. 689-703 (1980).
- 4) Buchanan,B.G. et al.:Dendral and Meta-Dendral:Their Application Dimension,A.I., Vol. 11, pp. 5-24 (1978).
- 5) Langley,P.:Data-Driven Discovery of Physical Laws,Cognitive Science, Vol. 5, pp. 31-54 (1981).
- 6) Feigenbaum,E.A. et al.(eds):The Handbook of Artificial Intelligence, Chapter 14, pp. 323-511, Pitman (1981).
- 7) Schank,R.C. et al.:Scripts, Plans, Goals and Understanding, John Wiley and Sons (1977).
- 8) Schank,R.C. et al.:Inside Computer Understanding, Lawrence Erlbaum Associates (1981).
- 9) Schank,R.C.:Language and Memory, Cognitive Science, Vol. 4, pp. 243-284 (1980).
- 10) 施貴彦,猪瀬常彦:帰納的理論・証明と知識表現の一提案、情報処理学会論文誌、Vol. 24, No. 1 (1983)