自然言語処理 122－21
（1997. 11. 21）

ＳＤ式モデルによる マルチメディアデータ検索システムの試作

脇山 正博† 河口 英二‡

†北九州工業高等専門学校
wakiyama@kct.ac.jp

‡九州工業大学工学部
kawaguch@kawa.comp.kyutech.ac.jp

ＳＤ式モデルは，著者らによって開発された自然言語の意味処理を行うための枠組みである．本稿は自然言語のマルチメディアデータベースシステム（MDSNL）について記述する． 本システムはデータベースとして HTML（ハイパーテキストマークアップ言語）を使用している．ソースデータとしては映画の映像，音声，場面の意味の記述，台詞としてのテキストデータとその意味記述としてのＳＤ式がある．著者らは場面データを検索する試作システムを開発した．

A Prototype of Multimedia Data Retrieval System by the SD-Form Semantics Model

Masahiro Wakiyama† Eiji Kawaguchi††

†Kitakyushu National College of Technology

††Kyushu Institute of Technology

The SD-Form Semantics Model, developed by the authors, is a framework to deal with semantic processing of natural language. The present paper describes its extension to a Multimedia Database System of Natural Language (MDSNL). The system is using HTML (Hyper Text Markup Language) for the multimedia database. The source data includes, motion picture, sound signal, semantic description of drama scenes, text data, and the SD-Form as their meaning. We made experimental system to retrieve scenes.

1．はじめに

インターネットの普及に伴い，世界中の情報が瞬時に入手可能となっている．さらに計算機処理能力の向上によってマルチメディアデータが扱えるようになった．しかし映像や音声といったマルチメディアデータの意味検索はまだ実現されるに至っていない．本研究は，ストーリーを持つ映像に着目し，シーン単位で映像検索が可能となるシステムの開発を目的としている．

著者らは今まで英会話文とその和文データ，さらにそれらの意味データを統合するシステムについて研究を行ってきた．そのシステムは会話テキストの意味検索を主目的とするものであった[1]．そしてそのシステムの最大の特徴は，ＳＤ式モデルに基づく２つの自然言語表現の間の意味距離が計算できることであった．

これまで多くの研究者が自然言語の意味論に取り組んできており，中でもフレームやネットワークモデルは広範囲にわたって研究されてきた[2]．また第一階述語論理に基づくモデルも典型的なアプローチである．しかしながらこの種のアプローチは文の論理的側面の解析には良いが，人間のコミュニケーションで重要な，喜び，怒り，悲しみ，驚き，失望，激励などの説明には不向きである．

著者らは新しい意味記述言語（ＳＤ式）を提案し，意味処理におけるＳＤ式モデルを開発した[3]．そのモデルの特徴として著者らが主張している点は次の点である．

(A) 世界知識の意味のみならず自然言語の意味を記述できる

(B) 自然言語によって表現された２つの既知の概念間に意味の差が計算できる

ＳＤ式は形式的には，或る一つの文脈自由言語である．

本稿では，ソースデータとしてテキストベースに限らず，動画像（映画の場面）や会話音声を含めたマルチメディアデータまでも含むシステムに拡張する試みを提案する．これらのデータベースのリソースは物語性を持つある一連の動画像からなるビデオ映像である．

これを「自然言語を用いたマルチメディアデータベースシステム(MDSNL)」と呼ぶ．この試みで最も重要な点は，それぞれのビデオ場面（対話が含まれる所）がＳＤ式によって意味記述されるということである．すなわち，それぞれの場面を記述した２つの場面の類似性がシステム知識を利用して計算される（場面知識，一般知識など）．

２章ではＳＤ式モデルの概要を述べ，意味差計算の基本概念を示す．３章ではデータベースについて記述する．映像データについては，「エピソード」「場面」，「ショット」の階層構造を定義し，ＳＤ式による場面の意味記述について論じる．場面検索の概要については４章で記述する．最後に５章では本研究を要約して，今後の課題について述べる．

2．ＳＤ式意味モデルの概要

2.1 ＳＤ式の型

ＳＤ式（Semantics-Structure Description Form）は，形式的に言えば，ＳＤＧと命名された文脈自由文法によって生成された記号列である．ＳＤＧの生成規則によればＳＤ式 は以下の８つの構文的な型に分類することができる．

2.2 ＳＤ式の型

(1)「変数ラベル」

・X, Y, Z, X1, X2, Y1, ･･･

これらの変数ラベル「何か」，「或る」などの意味記述のために導入したものであり，主にルール知識の記述にしばしば使用する．

(2)「単純ラベル」

・男，学校，トム，日本，動く，書く

これらは，いわゆる，概念ラベルである．

著者らは，このようなラベルとして，日本語（または英語）を借用している．

(3)「 パラメータラベル」

・度(12) (12度), DAVIS(教授) (Davis教授)

それぞれ括弧の中のラベルが括弧の外のラベルの意味を補足する.

(4)「修飾ＳＤ式」

・オレンジ/大きい (大きいオレンジ)

修飾形式では, それぞれの「/」の左側の主ＳＤ式を右側の従ＳＤ式が修飾する.

(5)「 規定子付きＳＤ式」

・*nega*(散歩) (散歩しない),

規定子として, それぞれの方法で括弧内の概念を規定する.

(6) 「結合子ＳＤ式」

・(動物)*incl*(人間) (動物は人間を含む)

結合子は概念を結合する (連結する) ために使用され, 結合されてまとまった概念を示す. ＳＤ式モデル自身は結合子の種類や表記法を制限しておらず, モデルのユーザが必要に応じて結合子を定義することもできる.

(7) 「陳述ＳＤ式」

・[s(空), v(青い)」 (空は青い.)

陳述ＳＤ式はＳＤ式のリストである. 各リストの要素は 6 種類の機能項目を持つ.

[s, *v*, *i*, *o*, *c*, *b*].

(8)「感情ＳＤ式」

・[r([s(自分),v(答える/優柔不断),o(相手)])]

(そうね.)

「挨拶」, 「呼びかけや応答」, 「感情表現」はこれらの感情ＳＤ式によって記述する. 感情ＳＤ式の機能項目は3種である [*a*, *r*, *e*].

2．3 ＳＤ式の意味情報量

任意のＳＤ式 d はそれ自身で, ある一定の意味量を表現している. これを ***si***(d) と表記し, d の意味スコアと呼ぶ. このスコアの単位は「semit」と定めている. ***si***(d) は d の構造によって一意に定まる.

次の事例は著者らの実験システム **SDENV-2** での ***si***(d) である.

d_1=本/英語 (English book.)

d_2=[s(トム),v(報告する/過去/昨日),o(メアリー),
o([s((ジョー)*plus*(X)),v(結婚する/過去)])]

(トムはメアリーに「ジョーと誰かが結婚した」と昨日報告した.)

この場合, ***si***(d_1)=21 で ***si***(d_2)=93 (semit).となる.

2．5 二つの概念間の詳述関係

ＳＤ式意味モデルにおける詳述関係は伝統的な「IS_A」, 「PART_OF」あるいは「IF-THEN」関係を包括するものである. ある概念 d_1 と別の概念 d_2 の間に詳述関係が成立するということは, d_2 が d_1 をより一層特定した概念であるか, あるいは詳細な概念であることを意味する.

詳述関係には２つのタイプがある. 一つは構文的詳述関係 ($elab_{syn}(d_1, d_2)=n_s$) であり, 他は知識データによる詳述関係 ($elab_{know}(d_1, d_2)=n_k$) である. 詳述の程度は詳述量 ($n_s$ あるいは n_k) として定量的に扱う. また詳述量は, 真の d_1 から d_2 をアブダクションする際の不確実性の尺度としても捉えることができる.

この２つの詳述関係を１つの表現にまとめて

$elab(d_1,d_2,n)$ または $elab(d_1,d_2)=n$

と表記する. ただし, ここで n は

$$n=min\{elab_{syn}(d_1,d_2), elab_{know}(d_1,d_2)\}.$$

である.

d_1 は d_2 の先祖と呼ぶ. 次の例での詳述量は **SDENV-2** で定めているスコアである.

＜例＞

$elab_{syn}$(本, 本/英語)＝11

(「英語の本」は 「本」を 11 semit 詳述した概念である.)

$elab_{syn}$(o([s(自分),v(会う),o(友人)]),
o([s(自分),v(会う/時/朝),o((友人)*equa*(ジョー))])
=33

(「お早う, ジョー (Joe) .」は「やあ」よりも 33 semit によって詳しい概念である.)

＜例＞

まず, 以下のような知識データがシステムに与えられているとする.

・(東京)*equa*(首都/日本)

(東京は日本の首都である.)

・(少年)*kdof*(男性) (少年は男性の一種である.)

・[*s*(ジョー),*v*(である),*c*(タクシー運転手)]
(ジョーはタクシーの運転手です.)

・(*assu*([*s*(X),*v*(である),*c*(タクシー運転手)]))
caus([*s*(X),*v*(運転),*o*(車)])
(もし誰かがタクシーの運転手であれば,
車を運転する.)

このとき,このシステムでは次のような詳述関係が成立する.

$elab_{know}$(東京, 首都/日本)=0
$elab_{know}$(男性, 少年)=3
$elab_{know}$([*s*(ジョー),*v*(運転する),*o*(車)],
[*s*(ジョー),*v*(である),*c*(タクシー運転手)])=2

2.6 最近共通先祖とその意味差の尺度

d_1 と d_2 の最近共通先祖 (d_0) とは全ての共通先祖の中から d_1 と d_2 両方に最も近い共通先祖のことである.このことを以下のように記述する.

$$ncoa(d_1,d_0,d_2,n_0)$$
$$n_0=\{elab(d_0,d_1)+elab(d_0,d_2)\}$$
$$=min_d\{elab(d,d_1)+elab(d,d_2)\}$$

SD式モデルでは d_1 と d_2 との意味差を以下のように定義する.

$$diff(d_1,d_2)=n_0.$$

<例>

この例は,与えられた知識を利用して,ある d_1 と d_2 間の意味差を求めるものである.

知識 :

(メアリー)*kdof*(女性), (王子)*kdof*(女性),
(白雪姫)*kdof*(王女/美しい)

概念 :

d_1=[*s*(白雪姫),*v*(好む/過去),*o*(キャンディ/甘い)]
(白雪姫は甘いキャンディを好んだ.)

d_2=[*s*(メアリー),*v*(好む),*o*(キャンディ/新しい)]
(メアリーは新しいキャンディが好きです.)

d_0=[*s*(女性/SOME),*v*(好む),*o*(キャンディ/X)]
(ある女性はあるキャンディを好む.)

$diff(d_1,d_2)=35+10=45$ (semit).

SD式モデルにおける意味差は構文構造によるものと知識データによるものを組み合わせて求めるという点が重要である.

3 マルチメディアデータベースの試作[4]

3.1 映像データ

本研究における映像データとは動画像と音声が同期したもので,さらにストーリーを構成しているものとした.通常の映画やドラマ等その例になる.このような映像データは,内容をもとにストーリー,サブストーリーと階層化して捉えることができる (図3.1).

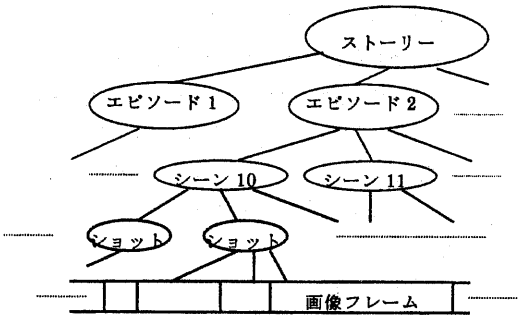

図3.1 映像データの階層構造

映像データの最小単位は「画像フレーム」と呼ばれる静止画像である.この静止画像を集めた「ショット」は映像データとして最小の意味単位となる.ショットが集まって一連の場面をあらわす区間を「シーン (後述)」と呼ぶ.更にシーンが集まってサブストーリーとなり,「エピソード」を形成する.エピソードが集まった全体を「ストーリー」と呼ぶ.例えば,全10話のドラマがあるとすると,全10話がストーリーであり一話一話がエピソードである.

今回の実験に使用した映像データは,或る一つの映画であったが,この映画ついての各階層の構成は以下の通りであった.

放送時間	: 122分
エピソード	: 10
シーン数	: 86

3．2 シーンの定義

本研究では，ほとんどのシーンが何人かの登場人物を含み，登場者らは一つのトピックについて話をしていると考えた．そして一つの話は数分間継続するものとし，これをシーンと定義した．

シーンの検出は，人が実際にビデオ映像を見ながら話題の切れ目と切れ目を見つけ，その区間を一つのシーンとした．話題の切れ目は，“場所”“人物”“時間”が変わるところを目安にした．それらは映像の内容を記述するのに関して，重要なデータである．

本実験で使用した映像データは，以上の定義に基づき，８６のシーンに分割された．

3．3 映像データベース

本研究におけるデータベースは，ＨＴＭＬ言語を利用して作成しWWWブラウザで見られるようにした．WWWブラウザのページ（出力）には，シーンの情報が分かるように以下の三つのデータを用意している．

(1) 映像データ（日本語，英語）
(2) シーンの意味記述
(3) 台詞（日本語，英語，ＳＤ式）

まず，映像データは，各シーンごとにＡＶＩファイルとしてハードディスクに保存した．保存形式は，以下に示す通りである．

・一秒あたりのフレーム数　１５フレーム
・ビデオサイズ　３バイト
　(320×240 ドット，約 1667 万色)
・サンプリングサイズ　１バイト
・チャンネル　モノラル
・周波数　１１ｋＨｚ

１シーンは平均 2 分弱で，データ量は，圧縮技術を使用して約３０メガバイトである．今回は検索システムの有効性を確かめるため，一つのエピソード（９シーン分）のみの映像データを作成した．

次に，シーンの意味記述は人が実際にシーンを見て，ＳＤ式で記述した（３．４節参照）．その時，シーンに関する人物，場所，時間，人物の行動に注目した．

台詞は作成した映像データをもとにシーンごとにファイル化して保存した．

これらは映像データの階層構造を利用して，ＨＴＭＬ言語を用いて作成した（図３．２）．ルートの部分でエピソードの選択を行い，次にシーンの各情報と枝別れしていき，最後の葉の部分に実際のシーンに関する情報が保存してある．

本データベースはＨＴＭＬの利用し，階層ごとにリンクを貼っているため，単体でインターネットでホームページを見るように，次々とシーンの情報を見ることができる．

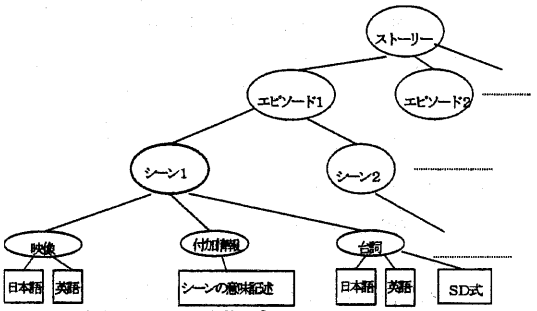

図３．２　映像データベース

データは映像データの階層構造を利用し，その階層構造にあわせてディレクトリを作成し，ＷＥＢのページもリンクしている．

今回，これらのデータは手作業でディレクトリの作成からファイルの作成まで８６シーン分のリンク付けを行った．本研究では後で説明する検索システムの出力としてもネットワーク上で可能である．

3．4　シーンの意味記述

本システムでは意味記述データを通して特定の場面を検索する新しい枠組みをインプリメントしている．例えば，「飛行機を爆破するシーン」を見つけだすことを望むとする．

そのようなシーンを検索可能にするために，MDSNL 内での（シーンｉｄを付加した）各シーンは以下のものとリンクしている．

(1) シーンのキーワード
(2) シーンの構造的情報

(3) ＳＤ式によるシーン記述

(4) ＳＤ式によるシーン固有の知識

カテゴリ(1), (2), (3)での情報はシーンの意味記述である. カテゴリ(4)の情報はシステムがあるシーンを捜す時の知識として動作する.

＜例＞

以下, 映像データのある個別シーンに意味情報を与える方法を示す. このシーンの内容は以下の通りである.

テロリストが７４７の飛行機で逃亡していた. テロリストの指導者はスチュアート大佐である. 警部補マクレーンが飛行機からこぼれているガソリンに火をつけてテロリストの逃亡を阻止した.

シーンID: ID2: EP10: 084

(1) シーンのキーワード

飛行機, コックピット, 爆破, テロリスト

(2) シーンの構造

登場者：警部補, テロリスト

数：３

登場者名：マクレーン, スチュアート

感情：恐怖, 喜び

(3) シーン記述

S-1: [*s*(747($1)),*v*(MOVE/(CONTINUE)*para*(place/RUNWAY))*para*(concur/[*s*($1),*v*(SPILL/(CONTINUE)),*o*(FUEL)])]

(747 が燃料をこぼしながら滑走路を移動している.)

A-1: [*s*(MCCLANE),*v*(LIGHT), *o*(TRAIL/FUEL)]

(マクレーンが燃料の後尾に火をつける.)

(4) シーン固有の知識

一般知識：

((assu([*s*(X),*v*(EXPLODE), *o*(AIRPLANE)])) *caus*([*s*(CREW),*v*(DIE)]))

(飛行機を爆破すれば, 乗組員は死ぬ.)

シーン固有の知識：

[*s*(CREW/747),*v*(BE),*c*(TERRORIST)]

(７４７の乗組員はテロリストである.)

４. シーン検索システム[5]

４．１ 検索システムの概要

本検索システムは, 与えられた条件を満たす映像情報を検索し, 出力するシステムである. このシステムをネットワーク上で実現するために, 本研究ではクライアント／サーバ方式を用いて作成した. このシステムは, 検索プログラムをクライアント側で実行し, サーバにあるデータを利用して検索を行なう, というものである.

ネットーワーク処理のプログラムはＪａｖａ言語で記述を行った. 検索処理プログラムはPrologを用いて行った.

５. まとめ

本研究における実験的な検索プログラムと映像データベースのネットワーク化は実現できた. しかし, 映像データの検索を選択したとき, 映像が送られてくるのに数十秒かかった. 本システムを実際にネットワーク上に置くことに対して, まだその出力には改善の余地がある.

参考文献

[1] Shao, G. et al. :SD-Forms as interlingua and system, J. of Japanese Society for Artificial Intelligence, Vol. 9, No. 5, pp.684-693 (1994)

[2] Sowa,J.F.: *Conceptual Structure: Information Processing in Mind and Machine,* Addison-Wesley, Reading, MA, (1984).

[3] Kawaguchi, E., Wakiyama, M.: The Semantic Metric Computation Scheme in the SD-Form Semantics Model, Proc. PRICAI, pp.623-629 (1993)

[4] Kawaguchi, E. and Wakiyama, M.: Toward Development of Multimedia Database for Conversational Natural Language, Information Modeling and Knowledge Bases VIII, ed. by H.Kangassalo, et al, IOS Press, pp.250-273 (1997)

[5] 占部忠幸, 脇山正博, 河口英二：ネットワーク上での映像データベースの構築とその検索システム, 情報処理学会第５５回全国大会論文集, ５Ｇ－１０(1997)