SD式意味モデルの意味差を用いた機械翻訳システムの検討

吉原 将太・脇山 正博・河口 英二

Examination of a Translation System Using Semantic Difference Measure of SD-Form Semantics Model

Shota YOSHIHARA, Masahiro WAKIYAMA and Eiji KAWAGUCHI

Abstract

The SD-Form Semantics Model, developed by the authors, is a framework to deal with the meaning of natural language in a quantitative way. Under this model, we are constructing a conversational text-database in English and Japanese language together with its meaning data by way of SD-Form. The merit of this model is that the system can calculate the semantic difference between two language expressions.

This paper describes the idea of the conversational English/Japanese texts semantic translation system using SD-Form semantics model.

1. はじめに

機械翻訳については、これまで様々なシステムやアプリケーションソフトが研究開発されてきた。例えば、科学技術振興事業団（http://pr.jst.go.jp）が開発した「JICST日英機械翻訳システム」やロゴヴィスタ社の「LogoVista X」(http://www.logovista.co.jp)、富士通ミドルウェア社の「ATLAS V9」（http://www.fmw.co.jp）など、実用化されているものも数多い。しかし、日英・英日への機械翻訳においても多くの技術的課題が残されており、たとえ8万円を超える高価格翻訳ソフトであっても、後編集が不必要な高品質の訳文を提供できるレベルには至っていない。ゆえに、これまで開発されたシステムは、翻訳者が大量のマニュアルなどを翻訳する際の下訳用とする場合、あるいは外国のWWWの文章を読む場合のように不完全な訳であっても利用者が概要を把握できるレベルで良いと割り切る場合に使用されるにとどまっている[1]。

多くの機械翻訳ソフトは、用例をデータベース化しており、翻訳対象の語がデータベースのものと一致する場合には、その用例を用いる。一致するものがない場合には、形態素解析をして単語レベルで訳する。このような手法の場合、次のような問題点がある。英日翻訳の場合には、直訳した翻訳では不自然な日本語になることが多く、だからといって全ての用例をデータベースに

登録するには、際限がない。また、その言葉を使っている状況によって意味が異なることがある。例えば、「They」を訳す場合、「彼ら」と訳すか「それら」と訳すかは前述の文の状況による．日本語では主語や目的語が省略された文が多いが、そのような文を日英翻訳すると適切な英語文にならないことが多い。また、長い文を翻訳することも不得意である。日英翻訳にしても英日翻訳にしても、会話文の翻訳は不得意と言えよう。

近年、著者等の研究グループは自然言語概念の意味表現形式の一つとしてSD式(Semantic-structure Description Form)を提案している[3,4,5,6,7]。SD式は、自然言語における個々の概念、陳述表現、感情表現、あるいはシステムに与える知識データ等を記述するための一種の中間言語であり、その構文は、曖昧さの無い文脈自由文法で規定されている[4]。自然言語概念をSD式として捉え、その記述データを基にして意味処理を行おうとするモデルをSD式意味モデルと言う。SD式意味モデルの最大の特徴は、従来からの意味記述モデルでは扱い難かった2つの概念間の意味的な近さを定量的に扱える点である。

また、著者等は、15～20年にわたって「NHKラジオ英会話テキスト」を基にした英会話例文の「英日会話文データベース」の構築を行っている。これは、主に会話文（英語と日本語）とそれに対応するSD式で記述した意味データ、および会話が行われている背景を述べたデータ（背景データ）からなる。本研究の目的は、この英日会話文データベースを用いて、入力した文に対して「意味的に近い文」を複数個出力する翻訳システムをWeb上に実現することである[8,9]。このシステムを「英日会話文翻訳システム（以下、本システム）」と呼ぶ。意味的にできるだけ近いものを結果とするために、処理の過程において、SD式意味モデルにおける意味差の尺度を利用する。本論文では、このようなシステムを作成するための試みを述べる。そこで第１段階として、本システムによる実験の便宜上、入力については自然言語の文を入力するのではなく、SD式で記述した１文を対象とする（自然言語文から自動的にSD式を生成するシステムは、現在開発中であるが、まだ実用レベルに達していない)。また、意味差の計算には、多くの計算時間を要するため、データベース中の全ての会話文との意味差を計算するのは非効率的である。また、SD式には多様な概念ラベルを用いることが可能であるため、異なる概念ラベルが用いてあったとしても、意味的には同じような内容であることがある。この場合、入力のSD式に用いられているラベルと一致するものだけを意味差計算の対象としたのでは、意味的に近くても対象から漏れてしまうことになる。そこで、概念ラベル間の類似性に関しては、国立国語研究所の分類語彙表[10,11]を用い、意味的に近そうなデータが漏れないようにして対象のデータを絞り込むこととした。

以下、２章ではSD式の記述例とSD式意味モデルにおける意味差の尺度について述べる。３章では、英日会話文データベースについて、４章では、英日会話文翻訳システムについて述べる。５章では、本システムの処理手順を追いながら実験例を示す。最後に６章で、まとめと今後の課題を述べる。

2．ＳＤ式意味モデルの概要

SD式意味モデル（Semantic-structure Description Form Semantics Model)は、自然言語の意味を定量的に分析するための枠組みである。このモデルに従った意味記述をSD式と呼ぶ。

2.1 会話文のＳＤ式記述例

会話文は、①発話意図を含む陳述形式のSD式（発話意図SD式)、または②感情SD式 で記述する。

発話意図とは会話において話し手が聞き手に伝達したい意図のことである。発話意図のラベルとしては、質問・告げる・依頼・命令・許可・提案・陳謝・感謝・確認などを設定している。発話内容は陳述SD式で記述する。発話意図SD式の構文は次の通りである。

[*s*(発話者),*v*(発話意図),*o*(対話相手),*c*(発話内容)]

英語で５Ｗ１Ｈ型と呼ばれている疑問文は、「発話意図」の部分を“質問”とし、「発話内容」の部分を５Ｗ１Ｈ型のSD式ラベル（何時、何処、誰、何、何故）を用いた陳述SD式で記述する。HOW疑問詞については、How many、How farなどいろいろな種類の疑問文があり、一意に決めることは難しい。発話意図SD式の例を以下に示す。

＜例１＞

・[*s*(自分),*v*(命令),*o*(相手),*c*([*s*(相手),*v*(閉める),*o*(ドア)])]

：ドアを閉めなさい。

・[*s*(自分),*v*(質問),*o*(相手),*c*([*s*(相手),*v*(行く/(未来)*para*(時/明日)*para*(場所/何処))])]

：あなたは明日何処に行くのですか？

感情SD式は、自然言語による感情的な発声や発話の状況の記述に用いる。感情SD式では、呼びかけ・応答・感嘆の３種類を定義している。感情SD式の例を以下に示す。

＜例２＞

・[*a*(トーマス)]	**：トーマス**	(呼びかけ)
・[*r*(否定)]	**：いいえ**	(応答)
・[*e*([*s*(トマト/指示),*v*(大きい)])]	**：このトマトは、なんて大きいんだ！**	(感嘆)

2.2 ＳＤ式の意味的情報量

SD式では、記号列の構造で何か固有の概念を表現しようとするだけでなく、その概念の意味量の大小も表すこととしている。そのため、SD式意味モデルでは、各SD式記号に意味素量を設定している。任意のSD式をDとするとき、その意味量を

si(D)= *n* 〔*semit*〕

と表す。また、その単位を*semit*としている。SD式意味モデル実験システムSDENV-4[7,12]の場合は、意味素量を以下のように設定している。

(1)	変数ラベル	"X, Y, Z, ..."	: 1〔*semit*〕
(2)	単純ラベル	"馬, CAT, ..."	:10〔*semit*〕
(3)	修飾子	" / "	: 1〔*semit*〕
(4)	規定子	"*nega, only, assu, ...*"	: 2〔*semit*〕
(5)	結合子	"*plus, incl, para, ...*"	: 1〔*semit*〕
(6)	機能項目記号	"*s, v, c, o, ...*"	: 1〔*semit*〕
(7)	区切り記号	"[ ]"	: 1〔*semit*〕
(8)	区切り記号	"( )", " , "	: 0〔*semit*〕

SD式全体の意味量は、それらの総和となる。

SD式意味モデルでは、これらを定義するときの一般的指針は示しているが、意味量の値については、モデルの利用者が独自に定めて良いとしている。

＜例３＞

・*si*(壷/古い) = 21

：古い壷

・*si*([*s*(一郎),*v*(である),*c*(選手/野球)]) = 45

：一郎は、野球選手である。

・*si*([*s*(国民/日本),*v*(納める/mood/義務),*o*(税金)]) = 67

：日本の国民は、税金を納めなければならない。

・*si*([*e*(感嘆/賞賛)]) = 23

：すばらしい。

2.3 ＳＤ式意味モデルにおける意味差の尺度[5,12]

２つの概念間の「詳述関係（elaboration relation）」は、SD式意味モデルにおける最も基本的な枠組みである。２つのSD式概念D_1、D_2に関して、D_1の意味をより具体化したものがD_2であり、逆に、D_2の意味をより抽象化したものがD_1であるとき、D_1とD_2には「詳述関係」があるという。この関係を

$$elab(D_1,D_2) = n$$

と表す。このときのnを「詳述量（単位：*semit*）」と呼ぶ。このようなD_1をD_2の先祖、D_2をD_1の子孫と呼ぶ。ただし、詳述関係が成り立たない場合は、

$$elab(D_1,D_2) = \infty$$

と定義している。

SD式意味モデルでは、概念間の「意味差の尺度」を次のように定義している。２つの概念D_1、D_2の全ての先祖D_{01}、D_{02}、…、D_{0i}、…の中でD_1、D_2に最も近い先祖を「D_1、D_2の最近共通先祖」と呼び、D_0で表す。

$$elab(D_0,D_1) + elab(D_0,D_2)$$
$$= min\{elab(D_{0i},D_1) + elab(D_{0i},D_2)\}$$

$= n_1 + n_2$

$= n_0$

この関係を

$ncoa(D_1, D_0, D_2, n_1, n_0, n_2)$

または

$ncoa(D_1, D_0, D_2) = n_0$

と表す。ことのきのn_0を「D_1とD_2の意味差」といい、

$diff(D_1, D_2) = n_0$

と表す。すなわち、意味差は与えられたD_1、D_2の最近共通先祖を探索することにより求められる。

最近共通先祖の導出及び意味差の例を以下に示す。

＜例4＞

次の概念D_1、D_2の最近共通先祖D_0、及び意味差n_0を求める。

D_1 = [*s*(恵子),*v*(育てる/手段/粉ミルク),*o*(健太)] ：**恵子は健太を粉ミルクで育てる。**

D_2 = [*s*(知美),*v*(育てる/手段/母乳),*o*(裕二)] ：**知美は裕二を母乳で育てる。**

ここでは次の知識データがシステムに与えられているものとする。

F_1 =(女性)*incl*(恵子) ：**女性は恵子を含む**

F_2 =(女性)*incl*(知美) ：**女性は知美を含む**

F_3 =(乳児)*incl*(健太) ：**乳児は健太を含む**

F_4 =(乳児)*incl*(裕二) ：**乳児は裕二を含む**

F_5 =(飲物)*incl*(粉ミルク) ：**飲物は粉ミルクを含む**

F_6 =(飲物)*incl*(母乳) ：**飲物は母乳を含む**

動作を分かり易く示すために、システムはF_1～F_4以外の知識データを持たないものとする。これらの知識データより、概念D_1、D_2の最近共通先祖D_0は、

D_0 = [s(女性/SOME),*v*(育てる/手段/飲物/SOME),*o*(乳児/SOME)]

：ある女性は、ある乳児をある飲物で育てる。

となる。D_0導出における詳述関係を図1に示す。図1における数字は詳述量である。概念D_1とD_2の意味差n_0は、図1における詳述量の総和で、

$n_0 = diff(D_1, D_2)$

$= elab(D_0, D_1) + elab(D_0, D_2)$

$= (1+1+1) + (1+1+1)$

$= 6$

となる。

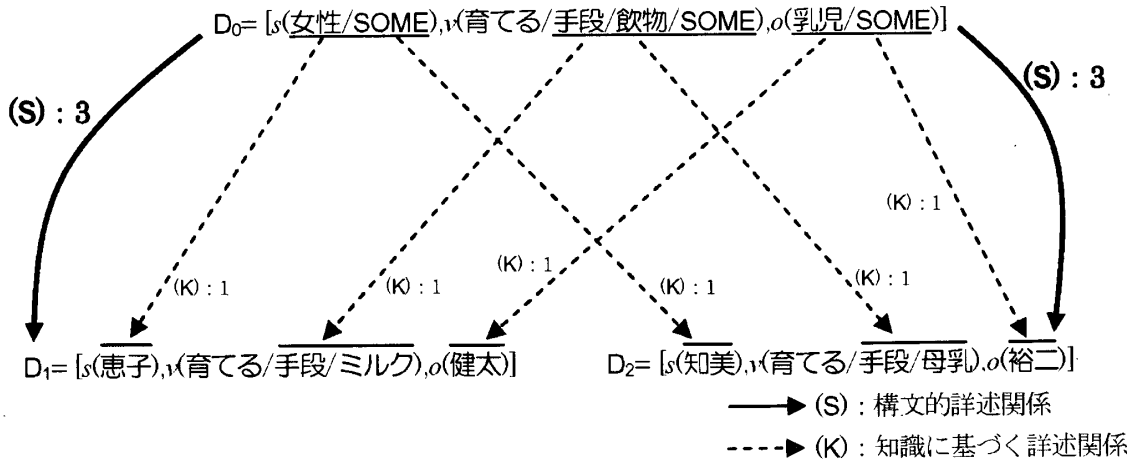

図１　D_1, D_2の最近共通先祖D_0の導出における詳述関係

３．英日会話文データベース[8]

本研究における英日会話文データベースは、原資料を「NHKラジオ英会話テキスト」(1989年4月号～1997年12号)として、「背景データベース」と「会話文データベース」の２つで構成している。現在のデータ件数は、背景データベースが約1000、会話文データベースが約15000件である。ただし、以下で説明するSD式データは、手作業で作成しており、その登録作業は完了していない。

3.1　背景データベース

背景データとは、会話が行われている状況(背景)を記述しているデータであり、「背景データ番号」、「タイトル(英語)」、「タイトル(日本語)」、「タイトル(SD式)」、「背景文(英語)」、「背景文(日本語)」、「背景文(SD式)」、「背景知識」の８種類で構成したものである。背景データベースとは、これらの背景データをデータベース化したものである。

背景データの各項目の説明を以下に述べ、背景データの例を表１に示す。

(1)　背景データ番号

背景データ番号は、次の７桁で表記する。

$SYYMMWN_1$

- S　：原資料の種類(ただし、現在原資料は「NHKラジオ英会話テキスト」のみで、これを“ｒ”と表記している)
- YY：年
- MM：週
- N_1：各週における背景文の連番

＜例5＞

r950442：NHKラジオ英会話テキスト1995年4月号、第4週 第2番目の背景データ。

(2) 背景データ番号

原資料では、各週の単元にタイトル(英語文／日本語文)が付けられている。これを、英語、日本語、それに対応するSD式の3種類の形式で背景データベースに登録している。

(3) 背景文

原資料には、会話が行われている背景が日本語および英語で記述されている。この日本語および英語、それに対応するSD式を登録している。

(4) 背景知識

会話が行われている場面に関連した情報を、知識データとして登録している。

表1 背景データの例

種 類	デ ー タ
背景データ番号	r950442
タイトル(英語)	Good Morning
タイトル(日本語)	おはよう
タイトル(SD式)	[*a*(挨拶/時/朝)]
背景文(英語)	At that moment, a small woman in a bathrobe emerges from the bathroom. She is drying her hair.
背景文(日本語)	そのとき、バスローブを着た小柄の女性が浴室から出てくる。女性は髪を乾かしている。
背景文(SD式)	[*s*(女性/(小柄)*para*([*s*(女性),*v*(着る/状態),*o*(バスローブ)])),*v*(出てくる/場所/浴室)]+ +[*s*(女性),*v*(乾かす/進行),*o*(髪/所有/女性)]
背景知識	(男)*incl*(ヒロキ) (女)*incl*(リサ) (女)*incl*(リサ) (食事)*incl*(昼食)

3.2 会話文データベース

会話文データとは、個々の会話文を記述したデータであり、「背景データ番号」、「会話文データ番号」、「話者(英語)」、「話者(日本語)」、「会話文(英語)」、「会話文(日本語)」、「会話文(SD式)」、「重要文指定」の8種類で構成したものである。会和文データベースとは、これらの会話文データをデータベース化したものである。

会話文データの各項目の説明を以下に述べ、会話文データの例を表2に示す。

(1) 背景データ番号

その会話文が使用されている場面に対応する背景データの番号を、**3.1**節(1)の形式で登録している。

(2) 会話文データ番号

会話文データ番号は、背景データ番号に3桁の会話文の連番を結合したものであり、次の10桁で表す。

SYYMMWN$_1$TTN$_2$

TT：話者の連番(ただし、同一人物が数度登場し発話しても、それぞれ１回の話者とする)

N$_2$：１回の話者による会話文の連番。

＜例６＞

r971023070：NHKラジオ英会話テキスト1997年10月号、第2週 第3番目の場面において、第７番目の話者による0番目(最初)の会話文データ。

(3) 話者

その会話文の話者を英語と日本語で登録している。

(4) 会話文

英語文１文を基準として、英語文、日本語文、およびそれに対応するSD式を登録している。

(5) 重要文指定

原資料において、「会話に役立つ重要表現」としてマークされている文を重要文としている。重要文の場合１を、重要文でない場合は0を登録している。

表２ 会話文データの例

種　類	データ
背景データ番号	r950442
会話文データ番号	r950442062
話者(英語)	Hiroki
話者(日本語)	ヒロキ
会話文(英語)	I′ll go directly to the office!
会話文(日本語)	直接オフィスに行くから！
会話文(SD式)	[*s*(ヒロキ),*v*(告げる),*o*(リサ), *c*([*s*(自分),*v*(行く/(直接)*para*(場所/オフィス))])]
重要文指定	0

4．英日会話文翻訳システム

本研究の目的は、英日会話文データベース(3節参照)を用いて、入力した文に対して「意味的に近い文」を複数個出力する翻訳システムをWeb上に実現することである。ただし、自然言語文からSD式データへの自動変換システムは現在開発中で未完成のため、現段階の本システムでは、入力文はSD式で記述したものを入力することとした。本システムの特徴は、完全に一致するものがなくても、意味的にできるだけ近いものを翻訳結果とするものである。そのために、次の処理を行う。

(1)SD式に使用されているラベル間の類似性に関して、分類語彙表[10,11]を用いて意味差を計算する対象データを絞り込む(**4.1**節参照)

(2)意味差の計算処理(**4.2**節参照)

翻訳結果は、入力文に意味的に近いと判断された会話文データを出力する。

4.1　分類語彙表による検索対象の絞り込み処理

分類語彙表は、単語を「体の類(名詞の仲間)」、「用の類(動詞の仲間)」、「相の類(形容詞の仲間)」、「その他」の4つに大分類したシソーラスである[10,11]。

SD式の概念ラベルは、既成の単語(英語や日本語)を使用する。そのため、多様な概念ラベルを用いることが可能であるが、異なる概念ラベルが用いてあったとしても、意味的には同じような内容であることがある。そこで、SD式に用いられているラベル間の類似性に関しては分類語彙表を用い、意味的に近そうなデータが漏れないように検索対象のデータを絞り込むようにしている。

使用した「分類語彙表[フロッピー版] [10]」には、語彙表と索引が収録されている。語彙表には、語を意味的に分類し、分類番号と見出し語が与えられている。本システムでは、語彙表のみを用いた。語彙表の一部を図2に示す。

1.1911　値・額
1　値　～値　価　価値　最高値　最低値　平均値　偏差値
2　値段　単価　頒価　建値　株価
3　栄養価　結合価　原子価
4　額　価額　金額　額面　帳面　実額　総額　差額　残額　満額　高額　低額 同額　倍額　半額　年額　月額　定額　税額　控除額　割り当て額　配分額 分け前　高　出来高　金高　金額　売れ高　売り上げ高　上がり高　生産高 漁獲高　産額　月産　日産　石高　禄高

図2　語彙表のデータ例

分類語彙表を用いた絞込み処理手順を以下に述べる。ただし、入力文はSD式とするが、そのSD式が「発話意図SD式」の場合と「感情SD式」の場合がある(**2.1**節参照)。

(1) 発話意図SD式の場合

① 入力文の意味を表すSD式の「発話内容」の部分からラベルを抽出し、リスト1とする。

リスト1：U_1、U_2、・・・、U_n

ただし、“自分”、“相手”、“当該”などのようなSD式記述ルールにおいて、一意に決められたラベルは除外する。

② リスト1のラベルU_nの同義語を分類語彙表から抽出し、リスト2に加える。このときU_nも加える。

リスト2：U_1、C_{U11}、C_{U12}、・・・、U_n、CU_{n1}、C_{Un2}、C_{Unm}

③ 発話内容にリスト2のラベルを1つでも持つ発話意図SD式を、英日会話文データベースから検索する。その後、入力文との意味差を計算することになる。

(2)感情SD式の場合

① 入力文の意味を表すSD式から全てのラベルと抽出し、リスト3とする。

リスト3：E_1、E_2、・・・、E_n

② リスト3のラベルE_nの同義語を分類語彙表から抽出し、リスト4に加える。このときE_nも加える。

リスト4：E_1、CE_{11}、CE_{12}、・・・、E_n、C_{En1}、C_{En2}、C_{Enm}

③ リスト4のラベルを1つでも持つ感情SD式を、英日会話文データベースから検索する。その後、入力文との意味差を計算することになる。

ラベルの同意語とは、分類語彙表(語彙表)において、同一の分類番号と段落番号をもつものである。例を以下に示す。

＜例7＞

ラベル：金額

同義語：額、価額、金額、額面、帳面、実額、総額、・・・

(分類番号：1.1911、段落番号：4)

SD式では、区切り記号“＋＋”を用いて複数のSD式を連結して1つの意味を表すことができる。英日会話文データベースには、このようなSD式で記述したデータも存在する。この場合には、“＋＋”の前後のSD式についてリスト2またはリスト4のラベルを用いているかを調べ、次のパターンで対象データを絞り込む。ここで、D_1、D_2は任意のSD式としたとき、対象データの形式がD_1＋＋D_2であるとする。

① D_1のみにリスト2またはリスト4のラベルが用いてある場合は、D_1のみを意味差の計算対象とする。

② D_2のみにリスト2またはリスト4のラベルが用いてある場合は、D_2のみを意味差の計算対象とする。

③ D_1およびD_2ともにリスト2またはリスト4のラベルが用いてある場合は、$D_1 + + D_2$を意味差の計算対象とする。

4.2 意味差の計算処理

分類語彙表を用いた絞り込み処理によって得られたSD式をD_1、D_2、・・・、D_iとする。入力SD式D_{IN}と各D_iとの意味差を求める(**2.3**項参照)。この処理においても、入力SD式D_{IN}が発話意図SD式である場合と、感情SD式である場合に分けて処理する。

(1) 発話意図SD式の場合

D_{IN}の発話内容とD_iの発話内容についての意味差を求める。

(2) 感情SD式の場合

D_{IN}と各D_iとの意味差を求める。

4.3 システムの概要

英日会話文翻訳システムは、Perl Ver.5.6.1でプログラミングし、Web上で動作するようにしている。翻訳処理の過程で意味差を計算しているが、その処理は、SD式意味モデル実験システムSDENV-4[7,12]の機能を利用している。SDENV-4も、Perl Ver.5.6.1でプログラミングしており、SD式意味モデルの応用システムでも利用できるように、それぞれの処理をサブルーチンとして記述している。他のシステムでSDENV-4を利用するには、Perlのrequire関数を用いてサブルーチン群を記述したファイル“sdenv4.pl”を取り込む。

英日会話文翻訳システムの入力ページを図3に示す。

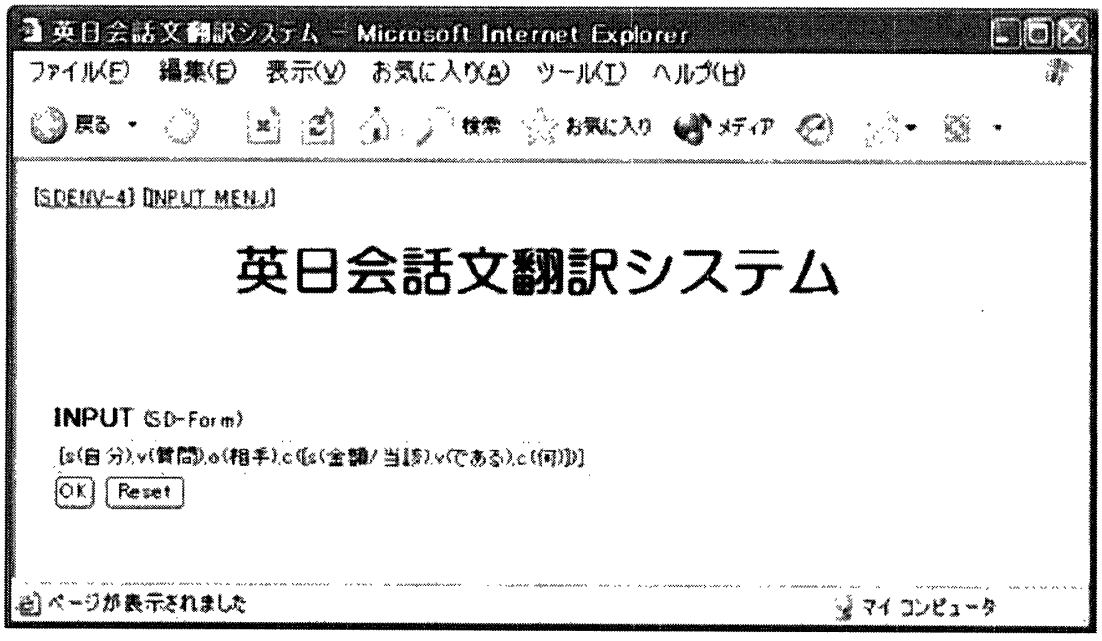

図3 英日会話文翻訳システムの入力ページ

英日会話文翻訳システムで実験する際には、図３において、“INPUT”の欄に翻訳したい文をSD式で記述する。ただし、現在の本システムでは、１文のみの入力に限定する。

“OK”ボタンを押すと翻訳処理が実行され、結果が表示される。結果は、「会話文(SD式)」、「会話文(英語)」、「会話文(日本語)」、「意味差」の情報を表示する。このとき、意味差が小さいものから順にリスト表示される。翻訳結果のページの例を図４に示す。

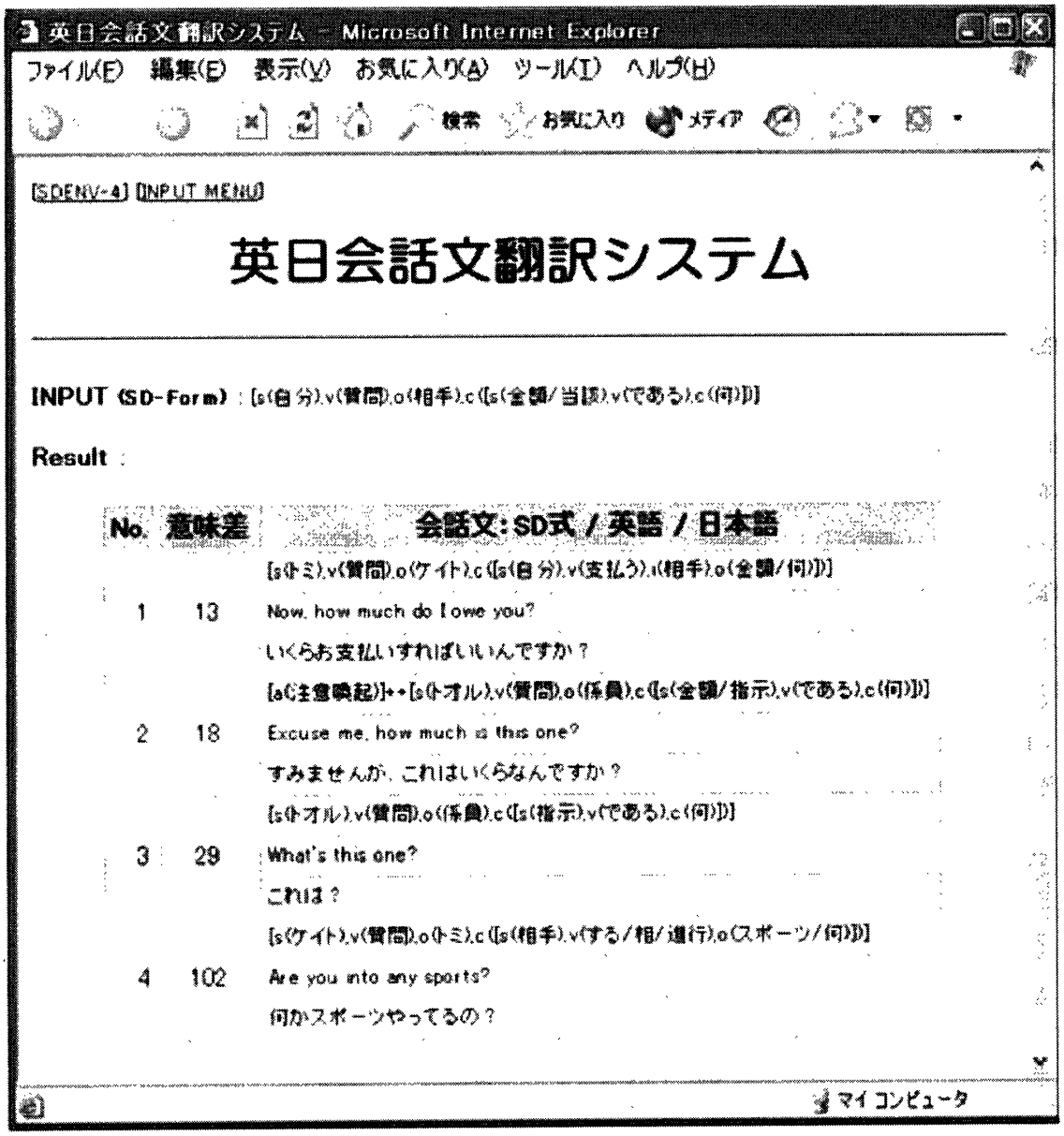

図４　翻訳結果のページ(英日会話文の表示)

５．実　験　例

本研究における英日会話文データベースは、「背景データベース」と「会話文データベース」の２つで構成しており、現在のデータ件数は、背景データベースが約1000、会話文データベースが約15000件である(３節参照)。しかし、SD式データは、手作業で作成しており、その登録作業は完

了していない。そこで、今回はシステムの動作検証を主とした基礎実験として、会話文データの件数を30件に限定して翻訳実験を行った。

以下に、英日会話文翻訳システムを用いた実験例を、処理手順を追いながら示す。ただし、自然言語文からSD式への自動変換プログラムは、現在開発中であるため、入力文はSD式とする。

入力SD式を次のD_{IN}とする。

D_{IN}:[*s*(自分),*v*(質問),*o*(相手),*c*([*s*(金額/当該),*v*(である),*c*(何)])]

(それはいくらですか？)

このD_{IN}は、発話意図SD式である（**2.1節**参照)。

(1) D_{IN}は、発話意図SD式であるので、発話内容の部分からラベルを取り出す。

リスト1：金額、である、何

(2) リスト1のラベルの同義語を分類語彙表から抽出し、リスト2とする。

リスト2：金額、値、額、値段、･･･、である、何、いくら、数、円、ドル、･･･

(3) 発話内容にリスト2のラベルを1つでも持つSD式を、英日会話文データベースから検索する。そのSD式をD_1、D_2、･･･、D_iとする。この例の場合、次のSD式が検索結果となる。

D_1：[*s*(ケイト),*v*(質問),*o*(トミ),*c*([*s*(相手),*v*(する/相/進行),*o*(スポーツ/何)])]

(何かスポーツやっているの？)

D_2：[*s*(トミ),*v*(質問),*o*(ケイト),*c*([*s*(自分),*v*(支払う),*i*(相手),*o*(金額/何)])]

(いくら支払いすればいいんですか？)

D_3：[*a*(注意喚起)]＋＋[*s*(トオル),*v*(質問),*o*(係員),*c*([s(金額/指示),*v*(である),*c*(何)])]

(すみませんが、これはいくらなんですか？)

D_4：[*s*(トオル),*v*(質問),*o*(係員),*c*([*s*(指示),*v*(である),*c*(何)])]

(これは？)

(4) 入力SD式D_{IN}と絞り込み処理で得られたSD式D_i（i =1、2、3、4）との意味差n_{0i}を求める。ただし、D_{IN}の発話内容の部分をD_{IN_U}、D_iの発話内容の部分をD_{i_U}とする。

D_{IN_U}:[*s*(金額/当該),*v*(である),*c*(何)]

①D_{IN_U}とD_1の発話内容D_{1_U}との意味差計算

D_{1_U}:[*s*(相手),*v*(する/相/進行),*o*(スポーツ/何)]

D_{IN_U}とD_{1_U}の最近共通先祖は、

[*s*(X),*v*(X)]

となり、意味差n_{01}は

$n_{01} = diff(D_{IN\_U}, D_{1\_U}) = 102$

である。

②D_{IN_U}とD_2の発話内容D_{2_U}との意味差計算

D_{2_U}:[*s*(自分),*v*(支払う),*i*(相手),*o*(金額/何)]

このとき、次の知識データが背景知識に登録されているとする。

(*asuu*([*s*(金額),*v*(である),*c*(何)]))*caus*([*s*(X),*v*(支払う),*i*(Y),*o*(金額/何)])

（もし金額がいくらかであれば、XはYにいくらかの金額を支払う。）

D_{IN_U}とD_{2_U}の最近共通先祖は、

[s(自分),v(支払う),i(相手),o(金額/何)]

となり、意味差n_{02}は

$n_{02} = diff(D_{IN\_U}, D_{2\_U}) = 13$

である。

③D_{IN_U}とD_3の発話内容D_{3_U}との意味差計算

D_{3_U}:[s(金額/指示),v(である),c(何)]

D_{IN_U}とD_{3_U}の最近共通先祖は、

[s(金額/X),v(である),c(何)]

となり、意味差n_{03}は

$n_{03} = diff(D_{IN\_U}, D_{3\_U}) = 18$

である。

④D_{IN_U}とD_4の発話内容D_{4_U}との意味差計算

D_{4_U}:[s(指示),v(である),c(何)]

D_{IN_U}とD_{4_U}の最近共通先祖は、

[s(X),v(である),c(何)]

となり、意味差n_{04}は

$n_{04} = diff(D_{IN\_U}, D_{4\_U}) = 29$

である。

(5) ①～④より、意味差が小さい順に結果を表示する。

このとき最小となる意味差とSD式は、

$n_{02} = 13$

D_2：[s(トミ),v(質問),o(ケイト),c([s(自分),v(支払う),i(相手),o(金額/何)])]

である。このSD式に対応する会話文は次の通りである。

英　語：Excuse me, how much is this one?

日本語：**すみませんが、これはいくらなんですか？**

4.3節の図３、図４は、この実験例の入力画面と結果画面を示したものである。

6．おわりに

英日会話文データベース（**3節**参照）を用いて、入力した文に対して「意味的に近い文」を複数個出力する翻訳システムをWeb上に実現することについて検討した。このシステムはPerl Ver.5.6.1で開発し、「英日会話文翻訳システム」と呼んでいる。翻訳処理の過程で意味差を計算しているが、

その処理は、SD式意味モデル実験システムSDENV-4[7,12]の機能を利用している。ただし、意味差の計算には、多くの計算時間を要するため、データベース中の全ての会話文との意味差を計算するのは非効率的である。また、SD式には多様な概念ラベルを用いることが可能であるため、異なる概念ラベルが用いてあったとしても、意味的には同じような内容であることがある。そこで、概念ラベル間の類似性に関しては、国立国語研究所の分類語彙表[10,11]を用い、意味的に近そうなデータが漏れないようにして対象のデータを絞り込む処理を行っている。

システムの有効性を確認するために実験を行った。ただし、現段階のシステムにおいては、自然言語文(日本語や英語)からSD式を自動的に生成するシステムが未完成であるため、入力はSD式とした。5節で例示した実験では、入力文の意味に近いものが検索結果として得られた。

今回の実験では、システムの動作検証を主とした基礎実験として、会話文データの件数を30件に限定して翻訳実験を行った。それゆえに、その中に含まれないような意味の文を入力した場合には、無関係な意味の文が検索結果となる。しかし、含まれているものについては良い結果を得られたことから、英日会話文データベースの約15000件の会話文データ全てにSD式を登録して実験すれば、ある程度解決できると思われる。他にも次のような問題点がある。意味差の計算は時間がかかる処理であるため分類語彙表を用いたデータの絞り込み処理を行っているが、たとえ絞り込まれた結果が100件であったとしても、それらと入力文との意味差を求める処理は瞬時には終了しない。また、主観的に意味が最も近いと思われる文が、意味差が最小となるとは限らない。これは、意味差はシステムに登録されている知識データにも依存しており、どのような知識データを登録しておくべきかを決めるのは難しい。それゆえに、複数個の文を翻訳結果として出力する必要性がある。

現段階の本システムには多くの改善するべき問題点があるものの、より自然な文を翻訳結果として得ることができるシステムとして、その有効性が確かめられた。

参考文献

[1] 内野一，白井諭，横尾昭男，大山芳史，古瀬蔵："速報型日英翻訳システムALTFLASH"，電子情報通信学会論文誌 D-II Vol. J84-D-II No.6，pp.1167-1174 (2001)

[2] 目良和也，市村匠，相沢輝昭，山下利之："語の好感度に基づく自然言語発話からの情緒生起手法"，人工知能学会論文誌 17巻3号A、pp.186-195 (2002)

[3] Kawaguchi, E., Wakiyama, M. and Nozaki, K.:"A Semantic Structure Description Model of General Concepts in Natural Language World", Proc. of PRICAI、pp.298-303 (1990)

[4] Kawaguchi, E., Kamata, S. and Wakiyama, M.:"Elaboration Relation and the Nearest Common Ancestor of a Concept Pair in the SD-Form Semantics Model", Proc. of 2nd PRICAI、pp.426-432 (1992)

[5] Kawaguchi, E., Kamata,S. and Wakiyama, M.:"The Semantic Metric Computation Scheme in the SD-Form Semantics Model"、Proc. of 3rd PRICAI, pp.623-629 (1994)

[6] Wakiyama, M., Shao, G. and Kawaguchi, E.:"The Toward Generalization of the Semantics Metric in the SD-Form Semantics Model", Proc. of 4th PRICAI'96 Poster, pp.61-68 (1996)

[7] 吉原将太，峯脇さやか，脇山正博，河口英二："CGIを用いたSD式意味モデル実験システムの試作"，電子情報通信学会技術研究報告 信学技報 Vol.101 No.484、pp.29-36 (2001)

[8] 峯脇さやか，吉原将太，脇山正博，河口英二："SD式と分類語彙表を用いた英日会話文意味検索システ

ム”，電子情報通信学会技術研究報告 信学技報 Vol.101 No.484、pp.45-50（2001）
[9] 脇山正博，峯脇さやか，吉原将太，河口英二：“SD式を用いた映像データ検索システムの試作”，電子情報通信学会技術研究報告 信学技報 Vol.101 No.484，pp.45-50（2001）
[10] 国立国語研究所：“分類語彙表[フロッピー版]”，秀英出版（1994）
[11] 国立国語研究所、中野洋：“「分類語彙表」形式による語彙分類表”，増補版，国立国語研究所（1996）
[12] 吉原将太，脇山正博，河口英二：“SD式意味モデルをWeb上で実験するための試み”，純心人文研究 第8号，pp.57-79（2002）