Internal Use Only (非公開)

TR-I-0204

階層的クラスタリング手法の訳語選択への応用

Application of the Hierarchial Clustering Method to the Translated Word Selection

井ノ上　直己　竹沢　寿幸
Naomi INOUE　Toshiyuki TAKEZAWA

1991.3

概要

単語を意味的に分類することは自然言語処理の分野では一般に行われている。しかし、人手で行うことは非常に困難であり、しかもコストの掛かる作業となる。そのため、単語を自動的に分類することが望まれる。本稿では、当研究所で構築中の対話データベース(ADD)から名詞と動詞の共起データを抽出し、そのデータを用いて名詞を階層的に分類した。また、その分類結果を訳語選択処理に適応したのでその結果を報告する。

©ATR Interpreting Telephony Research Laboratories
ATR 自動翻訳電話研究所

1. はじめに

類似概念を持つ単語に、同じ意味カテゴリ(シソーラスあるいは意味素性)を付与して意味的に分類することは、自然言語処理の分野で一般に行われている。しかし、意味カテゴリを作成することは容易なことではなく、また、たとえ意味カテゴリが作成できたとしても、単語にこれらの意味カテゴリを安定して付与することはほとんど不可能である。

そのため、自動的に単語を分類する手法がいくつか提案されている。その手法は利用するデータの違いにより以下の2通りに分類できる。

(1)国語辞典等の語釈文に記述された内容から自動的にシソーラスを作成する方法[1][2]

(2)コーパスに出現した共起データ(例えば、係り受けデータ)から自動的に単語を分類する方法[3][4][5][6]

前者については、そもそも国語辞典は上位-下位関係を考慮して作成されていないため、体系だったシソーラスが作成されず数多くのサブ体系に分断されるという問題がある[1]。後者については、通常意味的な遠さを表す距離が用いられるが、すでに指摘されているように、距離を用いて単語を分類する場合、複数の概念を持つ単語はその概念を事前に分離しなければならない[5]。つまり、実際には共起し得ないにも関わらず、概念を事前に分離しないために共起し得ると見なされ、実際には意味的に類似していない単語が意味的に近いと見なされるからである。

そこで、本稿では当研究所で構築中の対話データベース(ADD)から対訳情報を抽出し、日本語と英語とのペア(以下、対訳ペアという)で記述して事前に概念を分離することとした。このことは、日本語と英語の持つ概念の共通部分を抽出し、対訳ペアという客観的な記述で表していることに相当する。さらに、本稿では、階層的クラスタリング手法として知られているセントロイド法を用いて名詞を階層的に分類し、その分類結果を機械翻訳における問題の一つである訳語選択処理に応用したのでその結果を報告する。

以下、第2章に単語を意味的に分類する手法の考え方を示し、第3章でその際の問題点について具体的に示す。また、第4章には具体的な自動分類手法について示し、第5章以下でその結果を考察する。

2. 意味分類手法の考え方

意味的に単語を分類するためには、意味とは何かについてまず示す必要がある。意味とは何かという問題に対して種々の研究が行われてきたが、ここで用いるのは言語学的意味論の一つである「分布」という考えを導入した意味論である。EDRでは「分布」を以下のように紹介している。

ある言語要素の分布(distribution)とは、その言語要素の置かれている言語的な環境(environment)の総和である。つまり、ある言語形式が文中で使われる時、その前後にどのような語または語類がくるかということである。すなわち、文において、ある単語に注目しその単語の前後に現れる単語の集合のことである[7]。

また、「分布」を用いて意味論を展開した学者の1人であるHarrisは以下のように述べている[7]。

●分布が同じなら意味も同じである。
●分布が部分的に同じであるなら、その程度に応じて意味も類似している。

従って、「分布」という考えに基づいて名詞を意味的に分類する場合、その名詞と共起する動詞の集合を最もプリミティブな「環境」として用いることが可能と考えられる。そこで、本稿では当研究所で構築中の対話データベースから名詞と動詞の格関係を抽出し、Harrisの述べる意味論に基づいて名詞を意味的に分類する。具体的には以下のような手順になる。

(1)大規模なコーパスから名詞と動詞の格関係を抽出する。
(2)単語間の意味的な遠さを表す距離尺度を導入し、距離を求める。
(3)距離が近い単語同士をカテゴライズする。

3. 自動分類手法の問題点

コーパスから抽出した共起データを用いて単語を自動的に分類するためには、通常意味的な遠さを示す距離が用いられる。このような距離を用いて単語を分類する場合、複数の概念を持つ単語は事前にその概念を分離しておく必要がある。例えば、以下の例文(1)-(4)を考える。

(1)入場料を出す。
(2)本を出す。
(3)入場料を払う。
(4)本を読む。

ここで、「入場料」、「本」の共起テーブルを求めると以下のようになる。これから、「入場料」と「本」は「出す」という述語を共有する。

	出す	支払う	読む
入場料	＋	＋	－
本	＋	－	＋

ここで、＋は共起することを表し、－は共起しないことを表す

従って、「出す」の持つ2つの概念を事前に分離せず、上記の共起テーブルを用いて名詞を分類すると「入場料」と「本」は「出す」を共有しているため、それだけ類似度が高くなる。よって、「入場料」と「本」は同じカテゴリに分類される可能性が高くなる。しかし、ここで用いられた「出す」は「支払う」という意味(「出す-1」)と「出版する」という意味(「出す-2」)で異なった概念で用いられているため、実際の共起テーブルは以下のようになり、「入場料」と「本」は依然述語を共有しない。従って、「出す」の持つ2つの意味を事前に分離すると、「入場料」と「本」は同じカテゴリに分類される可能性が低くなる。

	出す-1	出す-2	支払う	読む
入場料	＋	－	＋	－
本	－	＋	－	＋

ここで、＋は共起することを表し、－は共起しないことを表す

以上に示した例から分かるように、「分布」という考えに基づいて単語を意味的に分類するためには、複数の概念を持つ単語は事前にその概念を分離して共起データを集めなければならない。

そこで、本稿ではこの問題を解決するため、当研究所で構築中の対話データベース(ADD)に付与された日英対訳対応情報から自動的に日本語と英語の対訳ペアを得て、事前にその概念を分離することとした。それにより、上記の「出す-1」は「出す-pay」と、「出す-2」は「出す-publish」と事前に分離できる。このように対訳ペアを用いることは、日本語と英語の持つ概念の共通部分を抽出し、対訳ペアという形式で客観的に記述したことに相当する。

4. 名詞の意味分類方法

4.1 抽出データ

当研究所で構築中の対話データベース(ADD)には、単語間の係り受け情報[7]および日英対訳対応情報[8]が付与されている。今回、ADDから名詞と動詞の係り受け情報とその単語に対する対訳情報を抽出した。抽出データの例を図1に示す。図1では"|"をdelimiterとして左から係り受けデータが出現した会話番号、修飾語、修飾語の品詞、修飾語の訳語、被修飾語、被修飾語の品詞、被修飾語の訳語および意味関係を示している。本稿の目的はここで抽出した名詞を意味的に自動分類することである。なお、ここで抽出した単語は出現頻度が高い程、共起関係が多岐に亘って得られる可能性があるため、国際会議というドメインで頻出した単語の上位から約500位内の単語を選んだ。国際会議というドメインに於て、これらの単語のカバレッジは約90%であった。

1|会社|名詞|company|出す|動詞|pay|動作主
15|要約|名詞|summary|出す|動詞|send|対象
157|プロシーディング|名詞|proceeding|出す|動詞|issue|対象
4|会議|名詞|conference|有る|動詞|be held|対象
8|質問|名詞|question|有る|動詞|have|対象
3|バス|名詞|bus|乗る|動詞|take|対象
180|新聞|名詞|newspaper|乗る|動詞|see|場所

図1　抽出データ

名詞を意味的に分類するために用いたデータは図2に示すデータであり、図1のデータから動詞として日本語と英語との対訳ペアを用い、異なりの係り受けパターンをその頻度とともに抜き出したデータである。

登録費|出す-pay|対象|1
要約|出すsend|対象|2
プロシーディング|出す-issue|対象|2
会議|有る-be held|対象|6
質問|有る-have|対象|7
バス|乗る-take|対象|1
新聞|乗る-see|場所|1

図2　名詞の分類に用いたデータ

4.2. 距離およびクラスタリングアルゴリズム

単語を自動的に分類するために、クラスタ分析を用いた。クラスタ分析とは多変量解析手法の一つであり、複数個の特性によって決定された個体間の類似度を基に、それらの個体をいくつかのグループに分類する手法のことである。そのため、クラスタ分析手法が名詞を自動的に分類するのに応用できると考えられる。また、この手法で用いる個体間の類似度は利用者の目的に依存して決定することができる。

ここで、第2章で示したHarrisの考えに基づいて単語を意味的に分類する場合、既に白井[3]らによって提案されている以下の尺度を用いることができる。なお、白井[3]らによって提案されたのは類似度ではなく距離である。

$$d(a,b) = 1 - \frac{\sum_{v \in V, r \in R}\sum \Phi(M(a,v,r), M(b,v,r))}{\sum_{v \in V, r \in R}\sum (M(a,v,r) + M(b,v,r))}$$

ここで、 a,b : 単語(名詞) $(a,b \in N)$
r : 意味関係
v : 動詞
N : 名詞の集合
V : 動詞の集合
R : 意味関係の集合
M(a,b,r) : 名詞aが動詞vに意味関係rで係る頻度

$$\Phi(x,y) = \begin{cases} x + y & (x > 0 \text{ かつ } y > 0) \\ 0 & (x = 0 \text{ または } y = 0) \end{cases}$$

また、クラスタ分析手法には様々な手法が提案されているが、今回実験に用いた手法は鎖効果を避けることができる手法として知られている重心法を用いた。以下に重心法について簡単に示す。

重心法

2つのクラスタを併合するとき、各クラスタの重心をもってそのクラスタを代表する点とし、その重心間の距離をクラスタ間の距離とする手法である。

5. 名詞の分類結果

ADDから抽出したすべての名詞を階層的に分類した。その結果の一部を図3に示す。ここで、横軸は距離を表している。図3は「リスト」と「用紙」が同じカテゴリに属し、「リスト」と「用紙」が属しているカテゴリと「資料」が同じカテゴリに属していることを表している。さらに、「リスト」、「用紙」と「資料」が属しているカテゴリと「希望」が同じカテゴリに属し、「リスト」、「用紙」、「資料」と「希望」が属しているカテゴリと「書類」が同じカテゴリに属していることを表している。同様に、「アブストラクト」と「プログラム」や、「決定」、「発表」、「スピーチ」および「話」が階層的に分類されている。このようにして、最終的にはすべての名詞が一つのカテゴリに属するように分類される。以下では、この得られた結果を階層クラスタと呼ぶ。

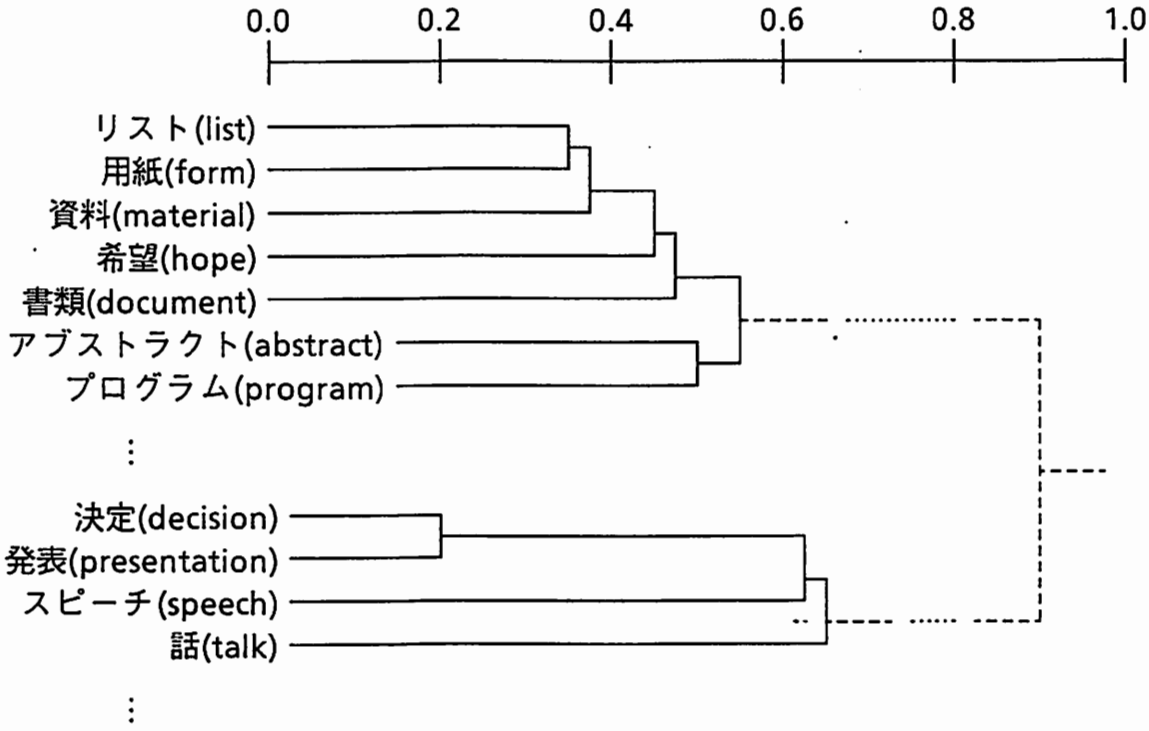

図3　名詞の自動分類結果の一部

6. 評価実験

階層クラスタを評価するために、機械翻訳処理の訳語選択の問題(特に、動詞の訳語選択)に適応した。以下に評価内容について詳しく説明する。

6.1 実験に用いたデータ

評価実験に以下の3種類のデータを用いた。また、実験は国際会議というドメインに関する電話会話約180会話を対象とし、階層クラスタを作成するのに用いた会話(学習会話)とそれ以外の会話(テスト会話)とに分けて行った。

(1)学習データ
階層クラスタを作成するために共起データを抽出した会話(学習会話)から図4に示すような名詞、動詞、動詞の対訳および意味関係の異なりのパターンをその頻度とともに得た。

(2)テストデータ
学習データを抽出した会話以外の会話(テスト会話)から図4と同様のデータを得た。

(3)動詞対訳データ
学習会話から図5に示すような動詞と対訳のパターンをその頻度とともに抽出したデータを得た。

なお、図4および図5の最後の数字は頻度を表す。

6.2 評価手順

登録費|出す|pay|対象|1
要約|出す|send|対象|2
プロシーディング|出す|issue|対象|2
会議|有る|be held|対象|6
質問|有る|have|対象|7
バス|乗る|take|対象|1
新聞|乗る|see|場所|1

図4　学習データの例

出す|pay|1
出す|send|2
出す|issue|2
有る|be held|6
有る|have|7
乗る|take|1
乗る|see|1

図5　動詞対訳データの例

階層クラスタを評価するために、動詞の訳語選択に応用した。実験は各テストデータから名詞、動詞、意味関係を得て、これらをキーとして学習データから動詞の訳語を得ることで行う。また、評価は以下の手順で得た訳語がテスト会話で実際に発話された訳語と一致するかを調べることにより行う。以下に動詞の訳語選択手順を示す。

(1)テストデータと学習データとが名詞、動詞、意味関係でマッチすれば、その学習データから訳語を得る。
　(1-1)訳語が複数得られる場合は、学習データの頻度が最も高い訳語を得る。
　(1-2)(1-1)で訳語が複数得られる場合は、その複数の訳語の内動詞対訳データから頻度が最も高い訳語を得る。
　(1-3)(1-2)でも訳語が複数得られる場合は、複数の訳語の内1つを得る。
(2)学習データとが動詞、意味関係でマッチし、名詞だけがマッチしない場合は、テストデータの名詞と学習データの名詞とが、作成した階層クラスタ内でLeast Upper Boundの関係になっている学習データから訳語を得る。
　(2-1)(1-1)に同じ
　(2-2)(1-2)に同じ
　(2-3)(1-3)に同じ
(3)テストデータと学習データとが動詞でしかマッチしなければ、動詞対訳データから頻度が最も高い訳語を得る。
　(3-1)(1-3)に同じ
(4)テストデータと学習データとが動詞もマッチしなければ、エラー表示する。

6.3 評価結果

6.3.1 階層クラスタと既存のシソーラスとの比較

国際会議というドメインに関する電話会話約180会話を4等分して、その内の1つをテスト会話、3つを学習会話として4通りの実験を行った。前節に示した訳語選択手順(1)、(2)、(3)のそれぞれの過程で訳語を得たデータ数と訳語が一致したデータ数を表1に示す。なお、手順(4)については事例がなかったため、記述していない。ここで、手順(1)は名詞、動詞、意味関係が学習データのものとexactに一致した場合であり、手順(2)は動詞、意味関係が学習データのものと一致したが名詞は階層クラスタを辿ることにより一致した場合である。手順(3)は動詞のみが一致した場合である。また、表1内の各数字は4通りの実験の平均値を示している。なお、比較のために既存のシソーラス(「類語新辞典」角川書店)[9]を用いて同様の実験を行った結果も示した。

表1　実験結果(a)

訳語選択処理過程	階層クラスタ		類語新辞典	
	処理されたテストデータ数	訳語が一致したデータ数	処理されたテストデータ数	訳語が一致したデータ数
手順(1)	209.25	70.5(33.7)	209.25	70.5(33.7)
手順(2)	152.5	43.25(28.4)	121	34.25(28.3)
手順(3)	76.25	25.25(33.1)	107.75	37.75(35.0)
total	438	139(31.7)	438	142.5(32.5)

()内の数字は(訳語が一致したデータ数)/(処理されたテストデータ数)×100を表す。

表1には訳語が一致したデータ数と処理されたテストデータ数の比率(以下、訳語選択率と呼ぶ)も併せて示してある。その結果、階層クラスタを辿ることによる訳語選択率は、階層クラスタを用いる場合も類語新辞典を用いる場合も大差(約28%)ないことが分かる。しかし、この評価は学習データから得た動詞の訳語を実際に発話された訳語と比較しているため、正確な評価とは言い難い。また、訳語が一致したデータの中には動詞の訳語を決定する要素と考えられない意味関係(例えば、時間格)にも関わらず、訳語が一致した場合も含まれている。

そこで、表2に示した22種類の動詞と意味関係について、意味的に正しい訳語が得られたかを人手で評価した。その結果を表3に示す。ただし、動詞と意味関係が一致したものを対象としているため、表3には訳語選択手順(1)、(2)の過程で訳語が得られた結果を示した。また、表2に示した22種類のデータはテストデータの中で動詞の必須格と思われる意味関係とその動詞を全てをピックアップしたものである。

その結果、類語新辞典を用いる方が階層クラスタを用いるよりも約6%訳語選択率が上回っているのが分かる。その結果を詳しく検討すると、実験に用いたデータには「事(こと)」や「物(もの)」等の意味的には重要でない名詞が含まれており、これらの単語の分類に階層クラスタと類語新辞典とで差があるためであった。つまり、類語新辞典では同じカテゴリに属しているが、階層クラスタでは

これらの単語は必ずしも近くに分類されなかった。そのため、動詞の訳語を選択する処理で「事(こと)」や「物(もの)」等が動詞の訳語を決定する重要な名詞と階層クラスタ内でマッチしたため誤って訳語が選択されたからであった。例えば、「書類を作る」の「作る」の訳語を得る場合、類語新辞典を用いると学習データ内の「原稿を作る」とマッチして訳語(“make”)を選択したが、階層クラスタでは「原稿を作る」ではなく「物を作る」から訳語(“form”)を選択し誤った結果を得た。

従って、階層クラスタを訳語選択に応用する場合、これらの意味的には重要でなく抽象度の高い名詞をいかに処理するかが問題である。

表2　調査動詞一覧表

動詞	意味関係	動詞	意味関係
書く	対象	着く	対象
送る	対象	待つ	対象
作る	対象	教える	対象
支払う	対象	取る	対象
乗る	対象	聞く	対象
持つ	対象	伺う	対象
見る	対象	伝える	対象
入れる	対象	おっしゃる	対象
入る	対象	有る	対象
出す	対象	成る	状態変化の終点
出る	対象	出来る	対象

表3　実験結果(b)

訳語選択処理過程	階層クラスタ		類語新辞典	
	処理されたテストデータ数	訳語が一致したデータ数	処理されたテストデータ数	訳語が一致したデータ数
手順(1)	66.25	53.25(80.3)	66.25	53.25(80.3)
手順(2)	38.75	23(59.4)	35.75	23.5(65.7)

()内の数字は(訳語が一致したデータ数)/(処理されたテストデータ数)×100を表す。

6.3.2 対訳ペアを用いる効果

距離を用いて単語を自動的に分類する場合、複数の概念を持つ単語はその概念を事前に分離する必要があることが指摘されている。特に、階層的に単語を分類するのではなく、類似概念を持つ単語をグルーピングする場合には対訳ペアを用いる効果が示されている[10]。しかし、類似概念を持つ単語をグルーピングする場合、閾値にもよるが分類される単語の数が少ないという問題があり、実用的ではない。そのため、階層的に単語を分類する方が実用的であり、階層的に単語を分類する場合に対訳ペアを用いることがどれくらい効果があるかを示す必要がある。そこで、対訳ペアを用いて名詞を分類した場合と、日本語だけを用いて名詞を分類した場合とを6.3.1に示した実験(b)で比較した。その結果を表4に示す。

表4　実験結果

訳語選択処理過程	対訳ペアを用いた場合		対訳ペアを用いない場合	
	処理されたテストデータ数	訳語が一致したデータ数	処理されたテストデータ数	訳語が一致したデータ数
手順(1)	66.25	53.25(80.3)	66.25	53.25(80.3)
手順(2)	38.75	23(59.4)	38.75	22.75(58.7)

()内の数字は(訳語が一致したデータ数)/(処理されたテストデータ数)×100を表す。

表4から、階層クラスタを辿った場合の訳語選択率は対訳ペアを用いて名詞を分類した場合と、日本語だけを用いて名詞を分類した場合とではほぼ同じであった。従って、名詞を階層的に分類してその結果を動詞の訳語選択処理に応用する場合、今回の実験データでは対訳ペアを用いる効果はほとんど無かった。階層的に単語を分類する場合は、単語を単にグルーピングする場合と異なり、単語間の階層クラスタ内における相対的な位置関係が問題となる。従って、対訳ペアを用いる効果はほとんど無かった理由は、今回実験に用いたデータに対しては名詞間の相対的な位置関係を決定するには対訳ペアはあまり影響を与えないからだと考えられる。

6.3.3 学習データ量と訳語選択率との関係

階層クラスタを訳語選択に応用するためには、学習データ量と訳語選択率との関係について検討する必要がある。表2に示した22種類のデータについて学習データと訳語選択率との関係を図6に示す。ただし、図6には階層クラスタを辿った場合の訳語選択率を示している。

図6から、学習データ量が少ない時は訳語選択率が高く、学習データ量を増やすに従って一旦訳語選択率は悪くなるが、その後徐々に訳語選択率は上昇しているのが分かる。学習データ量が少ない時に訳語選択率が高いのは、「リスト」や「アブストラクト」のような意味的に類似していることが比較的容易に分かる語同士が階層クラスタ内でマッチし、正しい訳語が選択されていたからである。その後、訳語選択率が一旦低くなるのは、「リスト」や「アブストラクト」のような語についてはexactにマッチするものが増え、「名前」や「内容」などの抽象的な名詞が階層クラスタ内でマッチして訳語を得る頻度が増したからであ

る。つまり、これらの抽象的な名詞を分類するには学習データ量が不十分であったためだと思われる。また、その後学習データ量を増やすと訳語選択率が上昇するのは、抽象的な名詞を分類するのにデータが揃ってきたからだと思われる。

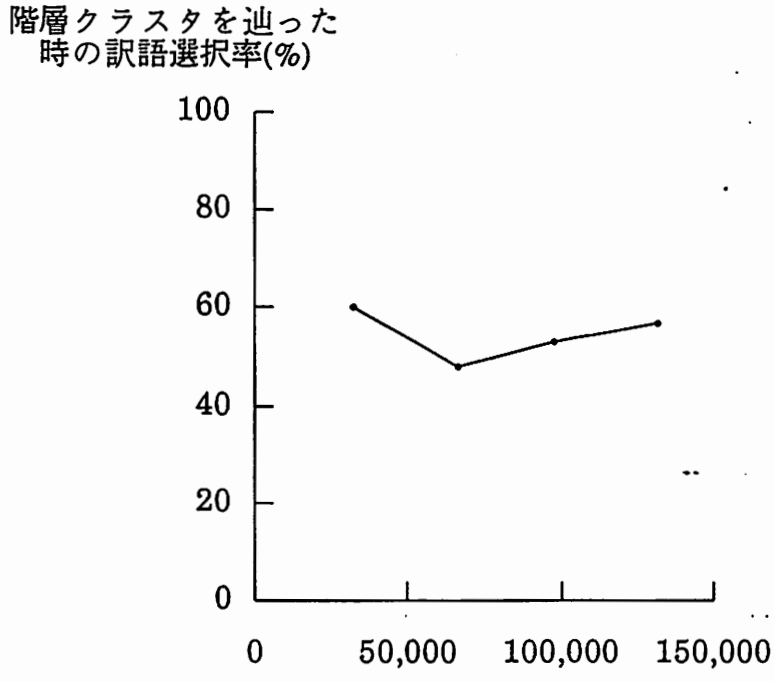

図6　学習会話の語数

7. まとめ

本報告では、当研究所で構築中の対話データベースから名詞と動詞の格関係を抽出して意味的に自動分類し、その結果得た階層クラスタを訳語選択に応用することを提案した。その結果、以下のことが分かった。

●既存のシソーラスを用いるよりも僅かに訳語選択率が悪いが、それは「事(こと)」や「物(もの)」等の意味的にあまり重要でない単語の分類に差があったからである。
●階層クラスタを用いて訳語選択を行う場合は、今回の実験に用いたデータに対しては対訳ペアを用いる効果はあまりなかった。
●学習データ量を増やすに従って、一旦訳語選択率は悪くなるが、その後徐々に訳語選択率が良くなる。

従って、訳語選択率をあげるためには、学習データを増やすとともに「事(こと)」や「物(もの)」等の意味的に重要でない単語の取扱いを検討する必要がある。

謝辞

本研究の機会を与えてくださるとともに適切な助言を述べられたATR自動翻訳電話研究所 樽松　明 社長、に感謝します。また熱心に討論して下さった江原暉将、鈴木雅実両氏他データ処理研究室諸氏および工藤育男氏他言語処理研究室諸氏に感謝する。

参考文献

[1]鶴丸、兵頭他「語義を考慮した単語間の階層構造の抽出について」、情処学会NL研究会64-2(1987)
[2]中村、酒井、長尾「英々辞典を用いた名詞の意味関係の分析」電子情報通信学会NLC研究会86-23(1986)
[3]白井、林他「係り受け解析のための辞書の構成とその学習機能」、情報処理学会論文誌Vol26,No4,706-714(1986)
[4]藤原「共起パターン分布に基づく単語間類似度を用いた動詞、名詞のクラスタリング手法」、人工知能学会全国大会8-3(1988)
[5]Hindle, D. 「Noun Classification from Predicate-Argument Structure 」、Proceedings of the 28th Annual Meeting of ACL(1990)
[6]日本電子化辞書研究所「概念辞書(第2版)」、EDRテクニカルレポートTR-012(1989)
[7]井ノ上、小倉他「言語データベース用格·係り受け意味体系」、ATR自動翻訳電話研究所テクニカルレポートTR-I-0029
[8]幸山、衛藤「日英対訳対応データの仕様」、ATR自動翻訳電話研究所テクニカルレポートTR-I-0152
[9]大野、浜西「類語新辞典」、角川書店
[10]鈴木、長谷川、井ノ上「会話文の翻訳における訳語選択とその規則の抽出」電子情報通信学会春季全国大会(1991)

<付録>

図1に示す係り受けデータ等は以下のDirectoryに存在する。なお、以下のDirectoryに存在するデータの説明はreadme参照のこと。

atr-dp:/data1/INOUE_THESAURUS/DATA/