Vol. 36 No. 2　　情　報　処　理　　Feb. 1995

解　説

最先端の科学技術とスーパコンピューティング

# 8. 人工知能におけるスーパコンピューティング†
## ——言語表現の類似性を利用する自然言語処理技術——

飯　田　　仁††

## 1. はじめに

最近の通信分野を中心としたニューメディア関連記事が新聞紙上を賑わしている．情報スーパハイウェーやマルチメディア通信の構想が打ち出され，米国ではすでに国会図書館の電子化プロジェクトがスタートした．また，近い将来紙資源の不足から電子新聞化は必至とも言われている．

このように大規模な電子化情報が蓄積されるようになると，その利用法が重要になってくる．情報源としては，書物や新聞に代表される未加工の文字列であるテキストを直接対象にすることが多くなろう．本稿では，それらテキスト上の情報検索や情報の加工などの基本的な利用技術になると考えられる言語表現の類似性計量法を中心に説明する．そして，超並列計算機を使った類似検索の現状，ならびにその応用としての翻訳処理の超並列化の現状について説明する．

## 2. 求められる自然言語処理の応用技術

日ごと蓄積されていく電子情報を使いこなす技術として，1)見当による適切な検索，2)テキストの自動分類とフィルタリング，3)事例の検索，が考えられる．

さらに，大規模テキストなどの情報を異なる言語に置き換える翻訳技術も重要になる．現状では，日米間，米中南米間などの情報収集のために不十分ながらも翻訳機が利用されている[5),6)]．同時に，異なる言語間における逐次的な情報検索とコミュニケーションも望まれており[4)]，4)要約を備えた高速言語翻訳，5)近似訳を許す緩やかな同時通訳，も重要な技術となっている．

以上のような情報加工処理は，処理の目的そのものが明確に規定されず，そのときの状況や環境に応じて要・不要の情報が異なってくることもある．そこで，以下に説明する言語表現の意味的な類似性を計量する技術は，加工情報を利用者に提示することが処理の中心であり，数値計算や問題解決の処理とは異なる．

## 3. 言語の類似表現の抽出

### 3.1 構造をもつ表現間の距離と意味概念上の類似表現間の距離

言語表現間の類似度を計量するためには，まず何か基準が必要であり，いろいろな基準が考えられるが，従来から構造をもつ表現間の距離について，つまり記号系列間の距離，図形間の距離，木構造表現間の距離，グラフ間の距離などについては広く研究されてきている[15)]．それに対し，ここで紹介する基準は，意味概念間の常識的な係わり，あるいは連想のしやすさを基準とするものである．この基準は，意味概念の階層的な連想関係で構成される系であり，通常広い意味でシソーラスと呼ばれる．この基準を使うことにより，探索範囲は連想の可能性がある限り拡散していく．その点で，構造をもつ表現間の距離を求める場合よりも，より大きな探索範囲を考慮しなければならず，超並列計算が有効となる．

### 3.2 概念間の距離に基づく単語間の意味的距離

意味概念とそれらの間の階層的な包含関係とからシソーラスを構成し，終端の概念はその概念に包含される単語の集合を備える．最上位の概念と最下層の概念との距離を1と設定し，最も簡単な例として$(n+1)$の階層に対して上下に連接する概念間の距離を$1/n$ずつに設定することにする．

† Super Computing on Artificial Intelligence—Natural Language Processing Technologies Using Similarity between Linguistic Expressions— by Hitoshi IIDA (ATR Interpreting Telecommunications Research Laboratories).

†† ATR 音声翻訳通信研究所

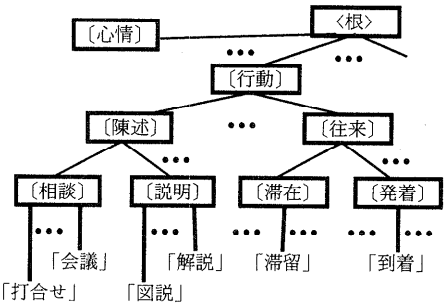
そして，同一概念に属する単語間の距離を0とする．4階層のシソーラスの例と単語間の距離の例を**図-1**に示す．図-1に例示した単語および概念の包含関係はだれもが共通的に了解できる常識として記述したものであり，細部にわたっての見方の違いや相互の係わりの程度が均一でないなどの問題を含む．しかし，基本的には，このような方法でデータ(あるいは記号)の照合を行うことにより，次に説明するようなベスト・マッチ(最適照合)が可能となり，従来のイグザクト・マッチ(完全照合)とまったく異なる処理を実現することになる．

### 3.3 句同士の間の距離

具体的な句の例を取り上げて説明する．日本語のテキストでも，自然の対話においても，「ナポリの会議」などのように助詞の「の」が大変便利に使われていて，その使用頻度も相当高くなっている．いま，二つの句，$P_1$ (例：「京都の会議」)と $P_2$ (例：「東京の滞在」)の距離を計算することを考える．言語表現間の距離を $d$，$P_1$ と $P_2$ を構成する $n$ 個の単語(あるいは構成素)をそれぞれ $c_{1i}$，$c_{2i}(1 \leqq i \leqq n$，例の場合は $n=3$) とするとき，次のように定式化できる．

$$d(P_1, P_2)=\sum_i d(c_{1i}, c_{2i}), \quad (1 \leqq i \leqq n)$$

この「の」を使った表現はいろいろな意味をもつが，それぞれの使われ方は話題や分野ごとに偏りを生ずるのが一般的である*．ここで，図-1の概念を示す表記[・]を使って，たとえば $c_{13}$ と $c_{23}$ に対する最も下位の共通概念である[行動]と $c_{11}$ の概念を固定した表現，「[地名]の[行動]」を考え，これに包含される具体的な句の意味用法と頻度を使うことにより，その分野における用法の偏りを把握する．同様に，$c_{13}$ の概念を固定し，$c_{11}$，$c_{21}$ に対して「[地名]の[相談]」の表現に代表される句の用法に対する偏りをとらえることができる．ある分野で日本語と英語の表現対を集めて，たとえば英訳のタイプ分類に従って各タイプの頻度を求めることができるから，その偏りを次のような重み付けとして計算できる[12]．




注) 単語ならびに概念間の距離を $d$ として，距離の例を示す．
$d$(「打合せ」，「会議」)=0
$d$(「会議」，「到着」)=2/3
$d$([相談]，[説明])=1/3
$d$([相談]，[滞在])=2/3
$d$([心情]，[発着])=1

**図-1** 概念間の距離の例と語の関係

$$w_i=\sum(\text{freq. of patt. on } [c_{1i}]=[c_{2i}])^2$$

(freq. of patt. on $[c_{1i}]=[c_{2i}]$：第 $i$ 項の共通概念を固定し，他項の概念を制約とする全観測例に対する各タイプの割合)

この重み付けは，観測される全用法に対し多次元の表現タイプの空間を作り，各用法頻度に基づいた各次元の値を要素とする表現タイプ空間上のベクトルを設定することになる．そして，言語表現間の距離 $d$ を，

$$d(P_1,P_2)=\sum_i d(c_{1i}, c_{2i})\times w_i, \quad (1 \leqq i \leqq n)$$

とすることができる*．

## 4. 並列連想プロセッサによる類似検索法

### 4.1 概念階層を使った用例検索

類似用例を逐次計算する検索の方法は図-1に示した概念階層に従って順次上位の概念におけるベストマッチを試みることが基本操作となる．つまり，ある特定の表現パタンで構成される用例の数 $N$ に対して，概念を**図-2**に示す3桁のシソーラスコードで表し，概念同士の照合(全3桁の一致)，上位の概念上の照合(上2桁の一致)，さらにより上位の概念上の照合(上1桁の一致)を行う．それに対し，連想される全上位の概念について同時にベストマッチを試み，最小の距離とともにその用例を検索できれば，大規模の用例を高速

* 「京都の会議」という句では，その使われ方を，対応する英訳のタイプで分類してみると，'conference in Kyoto,' 'conference on Kyoto,' 'Kyoto conference,' 'coference in Kyoto's style,' などが現れる．計算機科学に関する国際会議について，その事務局と質問者との話しには「京都で開催される会議」という意味で「京都の会議」が使われても，「京都に関する会議」という意味で使われることはまずない．

* ここに示した重みの与え方は一例であり，より良い重み付けの方法が議論されているが，本稿で扱う言語表現間の類似度計算はこの重みを使って計算したものである．

| | 入力 | 用例 | 距離 |
|---|---|---|---|
| (1) | 123 | 023 | 1 |
| (2) | 123 | 113 | 2/3 |
| (3) | 123 | 122 | 1/3 |
| (4) | 123 | 123 | 0 |
| ⌇ | ⌇ | ⌇ | ⌇ |
| (N) | 123 | 962 | 1 |

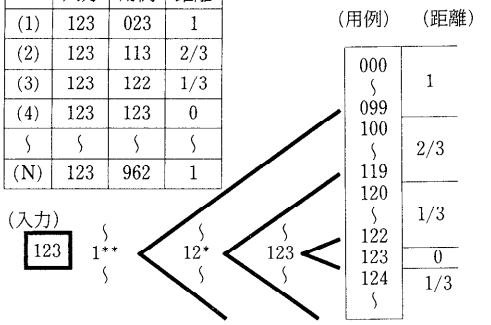



図-2　概念階層を使った基本的な単語間の距離計算(文献14)).ただし，入力と用例の数字は4階層のシソーラスコードを表す

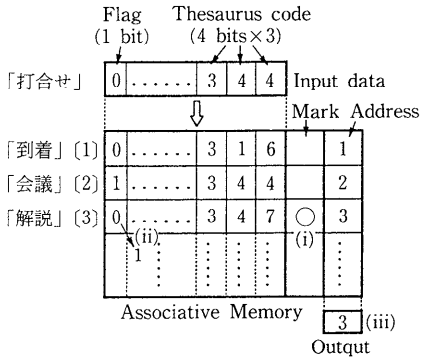



図-3　連想メモリによる意味距離計算の概念図(全3桁の照合後，上位2桁の照合を行ったところ)

に検索できることになる．つまり，連想できる概念について並列に検索する方法が可能となる．

### 4.2　連想プロセッサを使った用例検索

類似用例を並列計算するために，連想メモリを利用する方法が具体化している．そのような連想メモリを駆動する連想プロセッサ(AP*)として，並列連想プロセッサIXM2[2),3)]を構成するAPが使える．IXM2は知識表現の一つである意味ネットワークを処理することを中心に開発されたプロセッサであり，64台のAPからなり，256K語×40ビットの連想メモリを備える．

連想メモリは，トランスピュータのアドレス空間にメモリマップされる．そして，通常のイグザクトマッチとともに，並列書き込み機能や部分書き込み機能などを有している．1台のAPは，4K語の連想メモリを有するので，最大並列度4Kの並列処理が可能となる．

意味距離を計算するには，シソーラスコードを十進$n$桁で表し(図-2では3桁)，各単語にこのシソーラスコードを付与し，照合を一斉に行う[9)]．連想メモリを使った意味距離計算法の概要を**図-3**に示す．

## 5.　意味的な類似性を使った言語の翻訳

### 5.1　用例主導翻訳手法

用例主導翻訳手法(Example-Based Machine Translation)は，形態的な言語用法ごとに大量の翻訳例を用意して，入力される翻訳対象の言語表現と最も似かよった表現例を捜し出し，その対訳を直接，または一部変更して対訳を得るものである．言語規則に従って翻訳する従来の手法と比べ，用例主導翻訳手法は次のような特徴をもつ．第一は，類似検索を使うことにより最も似かよった対訳用例を捜し出すことができる．このとき同時に，翻訳システムが保有している対訳用例の量と質の程度に応じた，意味的に等価または近い訳を常に提供することができる．さらに，新たな対訳用例を既存の対訳用例データに追加していくことにより，翻訳システムの能力増進が可能である．

この手法を使うことにより，助詞の用法の典型例である「名詞＋の＋名詞」に対して約80%の自動翻訳率が得られている[11)]．また，他の助詞によって構成される2項関係(たとえば「XがY」/「XにY」/「XをY」「XでY」など)にも効果がある[12)]．

### 5.2　対訳用例を使った文の翻訳
### —変換主導翻訳手法—

用例主導翻訳により得られる句の翻訳結果を組み合わせることにより，通常の文が翻訳できる．さらに，文や句の典型的なパターンや特定品詞のパターンなどに着目した文の類似性をとらえた翻訳へも拡張可能となる．典型的な依頼表現の「お願いします」や，主題表現の「ダ文」(「僕はきつねダ」などに代表される文)についても，対話の話題を特定しておけば主動詞を補完した訳を作り出すことにも効果がある．IXM2上にこのような手法を実装し，高速な処理を実現している．さらに，変換主導翻訳手法**TDMT**と呼ぶ新しい翻訳

* APの主な仕様は，INMOS T801トランスピュータ(25 MHz)，4KBオンチップRAM(40 ns)，32K×32 bit SRAM(80 ns)，シリアル転送リンク4本(10 Mbits/sec)，4K×40 bit連想メモリ(375 ns)である．

手法[1]が提案され，日英・英日両方向による対話文の翻訳システムが作られている*．そこではいろいろな用例を利用し，可能な近似訳を即座に作り出すことが試みられている．

## 6. 超並列計算を使った自然言語処理応用

### 6.1 超並列計算機を使った従来の言語翻訳関連の研究

超並列計算機を使う言語翻訳の試みが開始されていて，並列連想プロセッサIXM2を使った解析システムASTRAL[7]をあげることができる．また，マーカ・パッシング技術を実行する意味ネットワーク用のアレイ・プロセッサSNAPを使った翻訳システムDmSNAP[8]が簡単な対話を翻訳している．それらはいずれも検証の段階にあり，拡張性などに関して課題を残している．そのほか特定の言語表現に関する翻訳実験として，nCUBE2を使った翻訳システムMBT3[10]がある．このシステムは漢字熟語などで記述される専門用語の名詞句を翻訳するもので，用途を限定すれば利用価値が高い．

### 6.2 類似検索を使った応用

連想プロセッサを使った用例検索について**4.2**で述べたが，その処理手法を基本として，SIMD型超並列計算機MasPar-mpp12000(8,000プロセッサ)上で，新聞記事から取り出した「の」を使った複合型名詞句10万用例をデータに使った類似検索実験がある．入力の句と類似している順に全データをソートするのに要する時間は，0.98ミリ秒であり[13]，**5.2**で述べた**TDMT**などを実時間処理するための基本操作として十分価値がある．つまり，特定のタスクならびに領域において収録した約13,000文をみると，平均文長約14語であり，また10秒前後の手短な発話であれば平均約20語と考えられることから，助詞を介した2語の関係や，文のパターンなどの例がそれぞれ10万件あったとしても，文全体の翻訳を100ミリ秒以内で十分処理できる．また，100万冊規模の電子図書館における類似タイトルの検索もほぼ実時間で実現できる見通しがあり，類似度に応じたフィルターを使うことによりインタラクティブな検索が可能となる．

### 6.3 音声翻訳手法

音声翻訳が当面扱っていく対話の範囲では，意図や話題や特定語に係わる派生的，あるいは類似的な表現が多用され，言い回しが豊富であることから，近似照合による対訳用例を利用する翻訳手法のほうが生産性，翻訳率，翻訳精度などの点で優れていると言えそうである．

このように，翻訳に関する経験的な知識は多様で相互の係わり合いがあるために，それらの多様な情報が適宜必要に応じて相互の情報を補完しながら，一つの解釈内容，または翻訳結果を導き出していける処理が必要であろうと考えられる．このような手法の実現を目指して，TDMTではSIMD型とMIMD型の並列計算機を組み合わせた対話文の翻訳システムの拡充と評価を進めている．

## 7. おわりに

本稿では，大規模テキストの高速類似検索，高速の対話翻訳などを対象とする並列計算機の利用法と効果などを中心に述べ，電子図書館利用技術の可能性についても考えてみた．大規模な電子化情報の構築を効率的・経済的に進める方法について今後活発な議論が期待される．

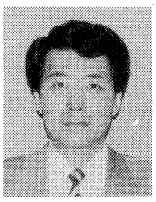
## 参考文献


1) 古瀬，隅田，飯田：経験的知識を活用する変換主導型機械翻訳，情報処理学会論文誌，Vol. 35, No. 3, pp. 414-425 (1994).
2) 樋口，古谷：人工知能への応用，特集—機能メモリのアーキテクチャとその並列計算への応用，情報処理，Vol. 32, No. 12, pp. 1287-1300 (1991).
3) Higuchi, T. et al.: The IXM2 Parallel Associative Processor for AI, COMPUTER, IEEE Computer Society, Vol. 27, No. 11, pp. 53-63 (1994).
4) 飯田 仁：対話翻訳と高度自然言語処理，人工知能学会誌，Vol. 6, No. 3, pp. 328-337 (1991).
5) Carbonell, J. et al.: JTEC Panel Peport on Machine Translation In Japan, Japanese Technology Evaluation Center, U. S. Department of Commerce (1992).
6) Kay, M. et al.: Verbmobil—A Translation System for Face-to-Face Dialog—, CSLI Lecture Notes No. 33, Center for the Study of Language and Information (1994).
7) Kitano, H. and Higuchi, T.: Massively Parallel Memory-Based Parsing, Proc. of IJCAI'91, pp. 918-924 (1991).


* 「国際会議の参加登録に関する問合せ」という話題に対話を限定しているものの，被験者5人による自由入力による実験で約71%という高い自動翻訳率を得ている．

8) Kitano, H.; Speech-to-Speech Translation: A Massively Parallel Memory-Based Approach, Kluwer Academic Publishers (1994).
9) Oi, K. et al.: Toward Massively Parallel Spoken Language Translation, Chapter 14 in Parallel Processing for AI 2, North-Holland (1994).
10) Sato, S.: MIMD Implementation of MBT3, Proc. of the Workshop on Parallel Processing for AI, IJCAI'93, pp. 28-35 (1993).
11) Sumita, E. and Iida, H.: Experiments and Prospects of Example-Based Machine Translation, Proc. of 29th ACL Annual Meeting, pp. 185-192 (1991).
12) 隅田，飯田：Example-Based Transfer of Japanese Adnominal Particles into English，電子情報通信学会論文誌 INF. & SYST., Vol. E-75-D, No. 4, pp. 585-594 (1992).
13) Sumita, E., Nishiyama, N. and Iida, H.: The Relationship between Architectures and Example-Retrieval Times, Proc. of AAAI-94, pp. 478-483 (1994).
14) 隅田英一郎：用例を利用した自然言語処理，「コーパスに基づく自然言語処理」講習会，電子情報通信学会，pp. 25-37 (1994).
15) 田中栄一：構造をもつものの距離と類似度，情報処理，Vol. 31, No. 9, pp. 1270-1279 (1990).

(平成6年11月29日受付)



**飯田 仁（正会員）**

1972年早稲田大学理工学部数学科卒業．74年同大学院修士課程（数学専攻）修了．同年日本電信電話公社武蔵野電気通信研究所入社．日本電信電話(株)基礎研究所を経て86年4月よりエイ・ティ・アール自動翻訳電話研究所に出向．93年3月よりエイ・ティ・アール音声翻訳通信研究所に再出向．現在，音声対話の理解・翻訳の研究に従事．言語処理学会，電子情報通信学会，人工知能学会，日本認知科学会，ACL各会員．