3L-10

段階的スコアリングによる文章間の類似度計算法

藤原祥隆　菊池英夫　松岡浩司
ＮＴＴ電気通信研究所

1．まえがき

情報処理装置のマンマシンインタフェースの高度化・高速化のためには、文字，単語レベルから一歩進んだ文（或いは文章）レベルの認識技術が今後重要な技術の一つになると考えられる。　一方，マンマシンインタフェースが重視されるデータベース検索等を目的とする質問応答システムでは、その質問文の表現パタンは幾つかの類型に分類できる有限集合を形成すると考えられる。　このような有限な入力文の集合を前提とする場合、システムに予め類型を代表する標準表現パタンを用意しておき、入力された質問文とこれらの標準表現パタンとの類似性を数量的に評価し順序づけを行うことによって、入力された質問文が何かを認識するアプローチが可能と考えられる。

本稿では上記の認識アプローチを可能とする中核となる手法、すなわち意味的類似性により分類した単語知識を基礎に単語レベルから出発して段階的により大きな表現間の類似度を求めることを特徴とする表現間の類似度計算手法を提案する。

（例文）
入力パタンA　：「音節を分析する手段」
標準パタンB　：「音声を認識する装置」

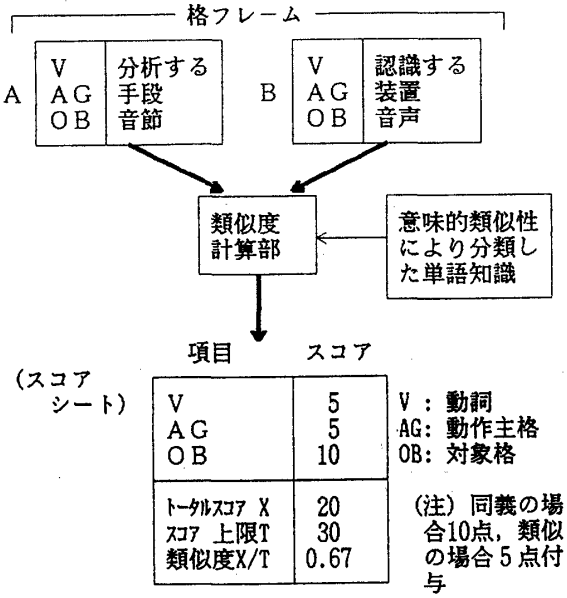

図１　類似度計算法の概念

2．類似度計算法

2．1　類似度計算の基本的考え方

異なる表現間の類似度計算の基本的な考え方を図１の一組の格フレーム間の類似度計算を例にとって説明する。

①質問文パタンおよびシステムに予め用意する標準表現パタンを格フレーム構造により表現する、②標準表現パタン側を基準として相対応する役割（動詞，対象格，等）にある単語相互の類似性を意味的類似性により分類した単語知識を用いて評価し、スコアを付与する、③各々の役割に関して得られた上記スコアを総計しトータルスコアＸを求める（一般には加重加算）。また同時に到達可能なスコアの上限値Ｔを求めこれら二つの値から類似度Ｘ／Ｔを求める。

上記の手順は複数の格フレームよりなる文（文章）間の類似度計算へと以下のように拡張できる。

①標準表現パタンを構成する各々の格フレームについて、比較対象の入力質問文を構成する格フレーム群の中から、それに最も類似する格フレームを上述の類似度計算手順を用いて見つけ、これを類似ペアとする。②各々の類似ペアについて得られたトータルスコアを総計しこれを新たに着目する文（文章）間のトータルスコアとする。同時に各類似ペアについて得られたスコア上限値の総計を求めこれを新たにに着目する文（文章）間のスコア上限値とし、前述の一組の格フレームの場合と同様にこれら二つの値から文（文章）間の類似度を求める。

2．2　ボトムアップ型の単語知識自動構成法

本類似計算法のポイントの一つは、意味的に分類された単語知識をどのように構成するかにある。単語を概念により分類する試みに関しては、人間が分類体系を予め設定し、その枠組みのなかに個々の単語を人間が判断してあてはめていくトップダウン的方法が古くから行われている[1]。一方、日本語文の係り受け解析に関し、実際の言語活動の結果である多数の文例から、同じ係り特性或いは、同じ受け特性を持つ単語を自動的にグループ化しようとするボトムアップ的方法が最近提案されている[2]。

本類似度計算法の適用を想定している質問応答システムでは、個々の応用によって使用される単語の種類や使用頻度、用法上等価な単語グループ等の特性が異なると思われる。こ

Phased scoring method of semantic similarity between sentences
Yoshitaka FUJIWARA , Hideo KIKUCHI , Kouji MATSUOKA
NTT Electrical Communications Laboratories

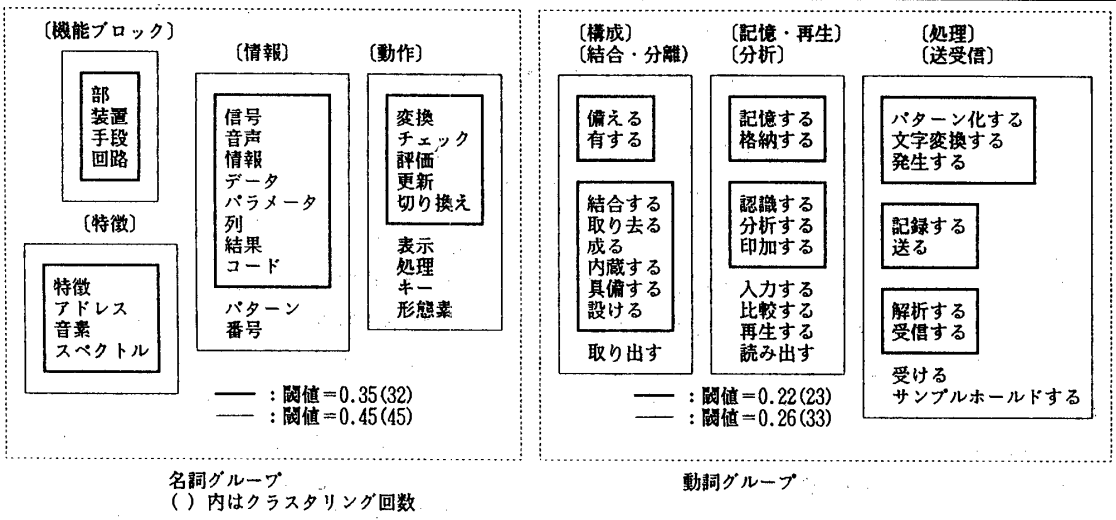

図2　単語グループ化の例（巨大グループ形成開始前）

のため単語知識の構築にあたっては、応用ごとに言語活動の結果である実際の用例を基礎に単語の意味によるグループ化を目指すボトムアップ的アプローチ法が適当であると考え、上述の方法と同様のボトムアップ的方法を試みる。

すなわち、2．1で述べたように、本類似計算法では、どういう動詞が使用され、それに対してどういう名詞がどういう格で関係するかを問題とする。そこで対象とする文例における単語の係り受け関係を、（動詞，格タイプ，格タイプに属する名詞）の3項関係に整理してとらえ、単語（名詞，動詞）間距離を文献（2）と同様に以下のように定義し，クラスタリングアルゴリズムを用いて単語を意味によりグループ化することを試みる。

単語間距離　$d(A,B)=1-N(AB)/(N(A)+N(B))$

ここで,N(A),N(B)；単語A,Bの出現回数、N(AB)；単語A,Bが同じ格タイプで同じ動詞に係る（或いは，同じ名詞を受ける）回数

3．単語知識構成法の評価

データベース検索の例として特許検索を取り上げ、特定の技術フィールドに関する文例を対象に2．2の方法による単語集合のグループ化の可能性を調べた。すなわち、名詞（係り単語）側，動詞（受け単語）側それぞれ独立に、同義語（名詞のグループ化のときは同義動詞，動詞のグループ化のときは同義名詞）を初期条件として与え、グループ形成の過程、グループ化の効率等を調べた。表1に実験に当たっての諸元を示す。これらの実験結果から以下のことが分かった。

1）グループ形成の様態：　概略以下の順序でグループ化が進行する。①独立した小グループの形成、②各グループの成長、③グループの統合と巨大グループの形成、④巨大グループの成長。　巨大グループの形成段階以後は本手法による分類の自動化は難しい。

2）グループの性質：　一つの類似概念からなるグループと2程度の異なる類似概念からなるグループが混在する。またそれぞれのグループには、ノイズと思われる関係の無い単語が含まれる傾向がある。

表1　諸元

項　目	内　　容
文例	特許請求の範囲記載文　67例 (7～10句/ 文，音声認識分野)
格の種類	通産省主管の日英翻訳プロジェクトの分類を基礎に以下のカテゴリに分類 動作主格，対象格，目標格，源泉格，役割格，手段・道具格，方式格，条件格，根拠・理由格
グループ化の対象	名詞;142個（出現回数2回以上，複合語は末尾の名詞を対象），動詞；164 個（出現回数2回以上，複合動詞は先頭の動詞を対象）
初期条件	同義語（名詞；42組　動詞;30 組）

3）グループ化の効率：　グループ化される単語の割合は、初期条件として与える同義語の集合にかなり左右される。例えば、本手法を適用できる臨界点と考えられる巨大グループ形成時点において、グループ化される単語の割合は、同義語を初期条件として与えない場合は全単語数の15％程度であるが、表1の条件のときは40％程度となる。

図2にグループ化の一例を示す。

4．むすび

表現間の類似度を計算する方法およびその基礎となる意味により分類された単語知識構成法の具体例による評価結果を示した。今後は本評価結果に基づく一層効率的な単語知識構成法の確立、類似度計算法の具体例による評価等が課題であると考える。

参考文献

1）石川他，“Muプロジェクトにおける意味マーカの概念と体系”，情報処理学会　自然言語処理研究会55－1，1986

2）白井他，“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，Vol.26,No.4