7F－3

意味属性に基づくテキストベース検索処理

内野一　松尾比呂志

ＮＴＴ情報通信処理研究所

１．はじめに

テキストベース検索は、蓄積されたテキスト情報（文の集合）から目的とする文の検索を行う。一般に検索は、蓄積された各々の文に文の内容を表すキーワードを付けておき、それをユーザの入力した文に含まれるキーワードと照合することで行われる。しかし、このキーワードによる検索法では、ユーザが検索キーを知らないと、検索することができない。そのため類義語辞書などを使ってキーワードを置き換えて検索が行われるが、この場合でも「頭が痛い」という入力で「頭痛がする」といった文を検索することはできない。ユーザは蓄積されている文がどの様な表現を使って書かれているかを知らないので、入力文に同じ表現が使われるとは限らない。それに対応するためには、入力に合致した文が無い場合でも意味内容が類似した文を検索する方法が必要となる。

本稿では単語の持つ意味属性を利用することによって、２つの文の間の意味的な類似度を算出し、類似度の高い文を検索する処理を提案する。また、この処理を用いたテキストベース検索システムの例について報告する。

２．文間の類似度の算出法

蓄積された文から、ユーザの入力した文に意味の近い文を抽出するため、入力文を構成する各単語の意味属性を抽出した後、以下のような形で文間の類似性[1]の評価値を算出し、類似度の高い文を抽出する。

２．１　単語の意味属性

単語には一般的な意味属性[2]を定義した汎用辞書と、その対象分野特有の意味属性を定義した専門辞書によって意味属性を与える。ただし、”する”、”ある”のような、単独では具体的な意味を表さない単語については、意味属性を与えない。

単語の持つ意味属性の例を図１に示す。この例では、”頭痛”という単語は、［頭］、［痛い］という２つの意味属性を持っている。

単語	意味属性
頭	［頭］
痛い	［痛い］
頭痛	［頭］［痛い］
嘔吐	［吐く］
しもやけ	［凍傷］
凍傷	［凍傷］

図１　単語の意味属性

２．２　類似性の評価

以下の観点により、入力文と蓄積された文の類似性の評価値を算出している。

(1)入力文の持つ意味属性が、多く含まれる文ほど、類似性が高い。
(2)入力文の単語と同じ表記（用言に対してはその終止形の表記）を持つ単語が文の中に存在すれば、より類似性が高い。

(2)の評価値は、”しもやけ”、”凍傷”のように意味属性が同じに付けられていて、評価値に差が生じないものに差をつけ、同じ意味属性を持っている中でも同じ表記を持った単語を優先するためであるが(1)の評価値より低い得点を与える。

図２は、２つの文の間の類似度の評価値算出法を示している。ここでは、入力文の各単語に対して基本点α、さらに表記が一致した場合は得点βを加えている。１つの単語に対して、複数の意味属性が与えられている場合は、基本点をその単語の意味属性の数で割った値を、その意味属性が一致した場合の得点とする。入力文の中の”頭痛”は［頭］［痛い］という２つの意味属性を持っているので、それぞれの意味属性が一致するごとに$\frac{1}{2}\alpha$の得点が与えられる。

文１は［頭］［痛い］の意味属性でそれぞれ得点$\frac{1}{2}\alpha$［吐く］という意味属性で得点αが与えられる。文２に対しては［頭］［吐く］の意味属性による得点$2\times\frac{1}{2}\alpha$に加えて、”頭痛”という表記が一致していることによる得点β（$\beta<\alpha$）が与えられ、文１の類似度は2α、文２の類似度は$\alpha+\beta$となり、より意味の類似した文が高い得点となる。これは、類義語辞書を用いた検索法ではできないことである。

入力文　頭痛がして、嘔吐もある。
文１　頭が痛くて、吐いた。　（６０点）
文２　頭痛がする。　（３５点）
（$\alpha=30$、$\beta=5$とした場合）

入力文			文１		文２	
単語	意味属性	得点	意味	表記	意味	表記
頭痛	［頭］	$\frac{1}{2}\alpha$	○	×	○	○
	［痛い］	$\frac{1}{2}\alpha$	○	×	○	
嘔吐	［吐く］	α	○	×	×	×
		合計得点	2α		$\alpha+\beta$	

図２　評価値の算出

Text type database retrieval based on semantic category
Hajime UCHINO, Hiroshi MATSUO
NTT Communications and Information Processing Laboratories

2．3　文の照合処理

検索対象となる文の蓄積時に文の形態素解析で得られた各意味属性から、文へのインデックステーブルを作成しておき、このインデックステーブルを使用して、照合処理[3)4)]を行う。インデックステーブルを使用することにより、検索処理は高速化され、蓄積される文が増加しても実用に耐える時間での処理が行える。

照合処理の流れを以下に示す。

1）ユーザが入力した文の形態素解析処理を行い、各単語の意味属性と表記を抽出する。
2）単語ごとに各意味属性からインデックステーブルを用いて、蓄積された文を抽出し、その文に対する評価値を加算する。（意味属性による評価）
3）さらに表記が一致した単語があれば、表記に対する得点を加算する。
4）評価値の合計の高い文を抽出する。

4．システムの実行結果

テキストベース検索システムの実行イメージを図3に示す。テキストは検索対象となる登録文とその内容を詳しく記した本文から成り立っている。システムは入力文と蓄積されたテキストの登録文との類似度を算出して類似性の高い登録文から順に表示する。ユーザは表示された文の中から自分の意図に最も合致した文を選択することにより、そのテキストの詳しい内容を記述した本文を見ることができる。

このシステムの評価のため、テキストベース「応急手当の方法」（約１５０文登録）に対して実験を行った。実験は登録文を見せ、それに対しての違った表現方法での入力文を入れてもらうことで行った。従来の類義語による方法では検索率が約８０％（４位以内）であったのが９０％となった。この向上分は、「頭が痛い」、「頭がガンガンする」という入力文から登録文「頭痛がする」が検索できるようにする本方法の効果である。

この方法を適用しても検索できなかった例として、「腹の調子がおかしい」と入力して下痢に関する文を検索しようとした入力例があった。「腹の調子がおかしい」という文は間接的な言い回しを使っているためであり、このような文に対しては、意味属性を用いる方法によっても対処することができない。これを解決するためには、間接的表現で表される概念と具体的な事象を表す概念とを結び付けるための知識が必要となる。

また、対象分野に関しての用語が数多く使われるため、汎用の辞書による意味属性だけで評価すると、分類が不十分で検索したい文以外にも多くの文が検索されることがあった。これに対しては用語に対する意味属性をつけるための辞書を強化していくことで対処できるが、一般のユーザにもこのような辞書が作れるようにツールを用意する必要がある。

5．おわりに

本稿では、単語の意味属性によって、テキストベースの探索を行う手法について提案した。この手法を用いることで、ユーザの多様な表現に対応したテキストベース探索を行うことができる。また、文法的な処理などと組み合わせることでさらに多様な表現を許容することができるであろう。

今後は、それらの処理との組み合わせを考えるとともに、今回の実験で必要性が明らかになった、対象分野に関する意味属性をつけるための方法を検討していく予定である。

[参考文献]
(1) 松尾他「日本語対話処理のためのユーザ入力支援」情処３８回全国大会pp. 400-pp. 401 1989
(2) 池原他「言語における話者の認識と多段階翻訳方式」情処論　28, 12, pp. 1269-1279
(3) 松尾他「特徴要素別にカテゴリ選択を行う高速パターン照合法」信学論　j70-d, 12, pp. 2503-2519 1987
(4) 松尾他「連想統合型照合による単語あいまい検索法」情処３４回全国大会 pp. 1845-pp. 1846 1987

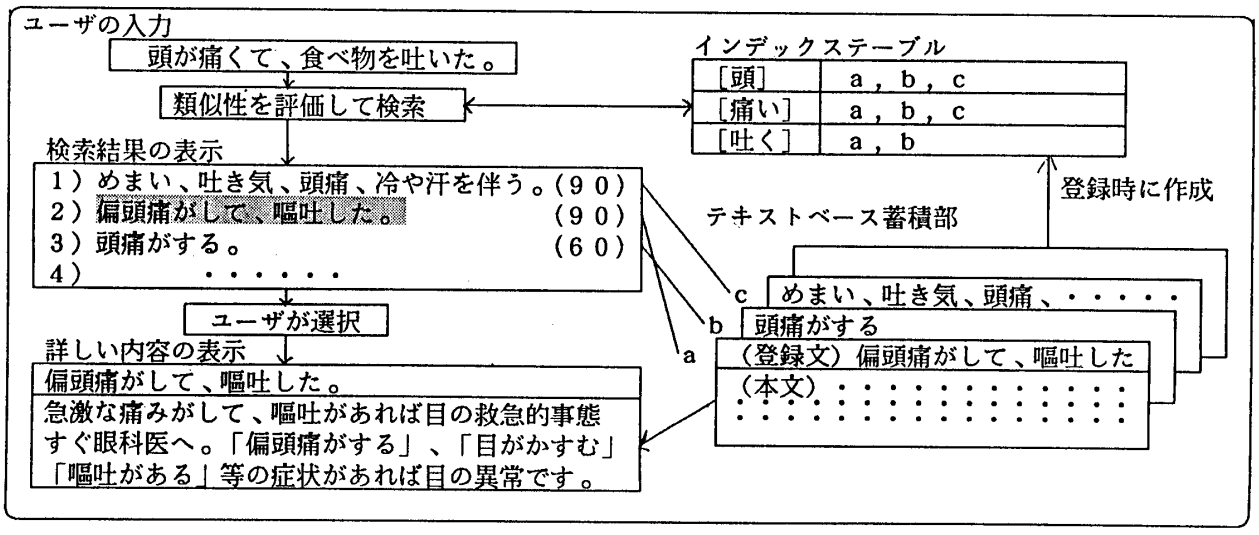

図3　システム実行イメージ