ユニフィケーションを用いた用例検索システム

7G-2

堤　豊*・隅田英一郎**

*日本アイ・ビー・エム株式会社 東京基礎研究所　**ATR 自動翻訳電話研究所

1．はじめに

翻訳支援システム、あるいは言語教育支援システムの機能として、対訳文のついた用例の検索は非常に有効である。用例検索の方法として、データベース照会言語を使用して、直接キーを指定して引く方法(1)や、KWIC(Key Word In Context)、あるいは、自動的に自立語を検出し、それを検索キーとして認識するもの(2)がある。しかしこれらは、あらかじめ検索キーを熟知していなければならかったり、また、付属語の単語の区切りが分かりづらいなどで、主に自立語の検索にしか用いられない。これに対して、冒頭に述べたような用途を考えると、用例検索としては、自立語の検索よりもむしろ構文の検索のほうが重要となる場合が多い。

筆者らは、構文の類似性をもとに用例を検索するシステム、ETOC (Easy TO Consult) を研究開発してきた(3)。このシステムは、従来の用例検索の枠組みと異なり、入力として日本語文を許しており、この入力文から自動的に検索キーを抽出する。また、検索キーの抽出のための規則がユーザーに開放されており、規則の順序を変更することで、検索の仕方を変えることができる(4)。

本稿では、ETOCを拡張して、構文と意味の両方で類似した文を検索する方法を提案する。この方法は、まず、構文的に類似した文を「緩い」規則で検索し、そのあと意味的な類似性を尺度にして順序付けを行うものである。また、構文的な検索のためにユニフィケーションを使用するという拡張を試みる。

2．類似検索の概要

図 1に、ETOCの枠組みを示す。ここで示されるように、システムは、入力文解析部、一般化部、検索部の三つから構成される。入力文解析部では、与えられた日本語入力文を形態素解析し、品詞認定をする。次に、検索部で一致するものをデータベースから検索する。もし一致するものがなければ、一般化部で規則に従って一般化し、また検索部に戻る。この枠組みの特長として

(1) 検索入力が日本語文であり、検索に特殊な能力や知識を必要としない。

(2) 一般化規則が開放されており、ユーザーはこの順序を変えることで検索のしかたを変更することができる。

(3) データベースの作成は、全く機械的に行われるため、新たなデータベースの作成も容易に可能である。

しかし、徐々にマッチングの条件を緩めていき、条件を満たすデータがあるところでやめるというアルゴリズムのため、データベースに含まれる用例文が増えるほど出力される結果が多くなるという欠点がある。これを防ぐには、

(1) 一般化規則をより細かく設定する。

(2) 出力された結果をさらに順序付けする。

の2 つの方法が考えられる。一般化規則をこれ以上詳細にすることは、構文的な規則だけでは難しく、意味的な規則を導入する必要がある。しかしこの2 つの要素は全く異なった観点のものであり、単一のマッチングの順序に埋め込むには問題がある。また、一般化規則の数が多くなると、ユーザーが順序を変更するのが大変であること、一般化－データベース検索のループを何度も繰り返すため、効率的でないということも考慮して、(2) の方法をとることにした。

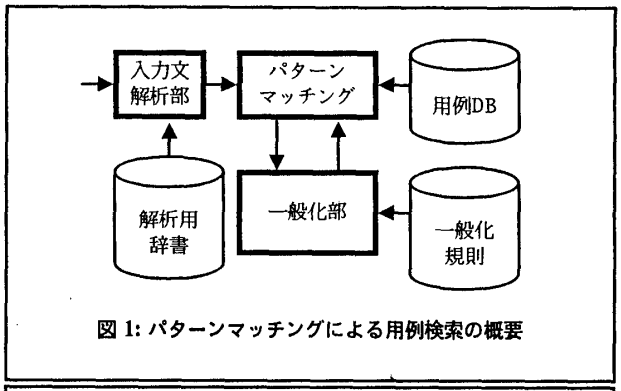

図 1: パターンマッチングによる用例検索の概要

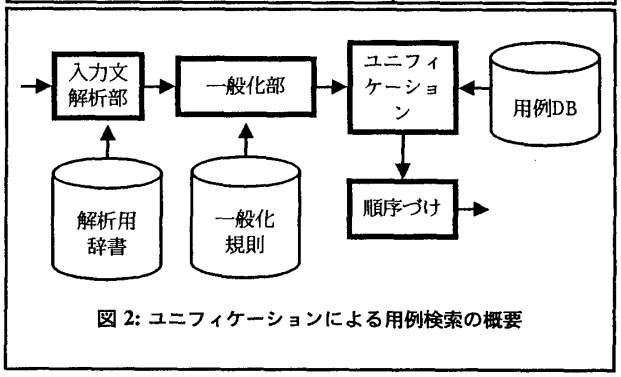

図 2: ユニフィケーションによる用例検索の概要

3．ユニフィケーションによる改善

図 2にユニフィケーションを用いた類似検索の枠組みを示す。ここで、一般化規則は、検索入力文についてのみ適

A Sentence Retrieval System Using the Unification.

Yutaka TSUTSUMI*, Eiichirou SUMITA**

*Tokyo Research Laboratory, IBM Japan, Ltd., **ATR Interpreting Telephony Research Laboratories

用され、最も骨格となる構造まで一般化する（図 3）。このとき、どこまで一般化すればよいかという判断として次の基準を設けた。

・機能語のみのパターンでは、最低機能語が3 つ以上含まれていること。

これにより、「～は～を」のような極めて抽象的なパターンは排除することができる。

次に、これを含むような用例をデータベースから検索する。このとき、マッチングにユニフィケーションを用いる。すなわち、検索入力文の構造を含んでおり、かつ矛盾しないようなデータベース中の例文はすべて検索される。このような操作は繰り返し検索によってユニフィケーションを用いなくとも可能ではあるが、ユニフィケーションを使用することで、規則の記述を簡略化することができる。このようにして得られた用例文の集合を1 次検索結果とする。

代名詞->体言
名詞->体言
体言「の」体言->体言
動詞語尾の削除
動詞->用言
形容詞，形容動詞->用言
修飾語の削除
格助詞の削除
接続助詞の削除

図 3: 一般化規則の例

次に1 次検索結果を順序付けを行う。この順序付けは、検索入力に近いものを先に表示するために行う。

1 次検索結果のそれぞれの文と入力文との類似性の尺度として次のようなものを考える。

（1）入力文中で1 次検索において使用されなかった単語がいくつ検索結果に含まれているか。ただし、「私」が「体言」のように一般化されたものは、1 次検索において使用されたとみなす。（w1 ）

（2）入力文中に含まれている単語以外にいくつ単語が入っているか。ただし、ユニファイされた単語は数えない。（w2 ）

（3）ユニファイされた単語が、もともとの検索入力の単語とシソーラス上で何レベル離れているか。ここでシソーラスとしては、図 4のようなものを考える。（t）

距離＝w2 ＋t －w1

とする。

4 . 検索結果例

検索入力：彼女は買物がうまい。

1 次検索のパターン：［体言］は［体言］がうまい。

彼は水泳がうまい。　（距離：2　t＝2）

彼女は歌がとてもうまい。　（距離：2　t＝1,w2＝1）

山田さんは将棋がうまい。　（距離：3　t＝3）

秋はさんまがうまい。　（距離：5　t＝5）

ステーキがうまい。　（1 次検索で検索されない）

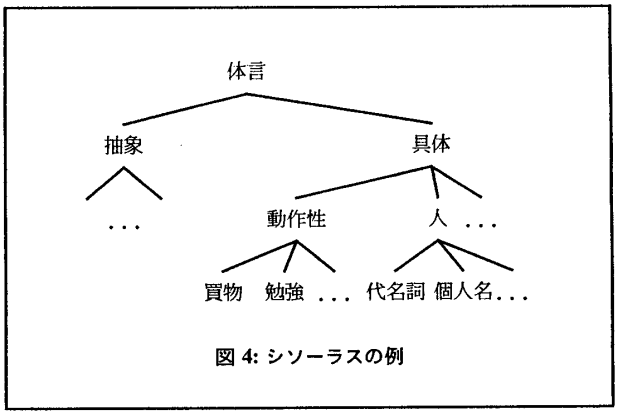

図 4: シソーラスの例

5 . まとめ

ユニフィケーションを用いた類似検索システムについて述べた。この類似検索は、与えられた入力文から自動的に抽出される検索キーによりデータベースを検索し、さらに順序付けを行うという、2 段階の作業から成り立っている。これは、1 次検索では、構文的な類似性を検出し、順序付けは意味的な情報を用いた類似性により行うものである。文の類似度を計算するためには、構文的な情報と意味的な情報とを同一のレベルで比較するという方法もあるが、ここで論じているようなアプリケーションでは、構文的な類似性を重視した方法が望ましいと考えられる。

今後は、順序付けのための規則についてさらに考える必要がある。また、順序付けのための距離の計算についても引き続き考えていきたい。

謝辞

本研究を遂行するに当たり非常に有益なコメントをいただいた、武田浩一氏、横井伸司氏に深謝致します。

文献

1. 野美山: テキストからの知識獲得支援ツール. 情報処理学会第3 7 回全国大会講演論文集 **7B-3** (1988).

2. 中村: 用例検索翻訳支援システム. 情報処理学会第3 8 回全国大会講演論文集 **4E-5** (1989).

3. 隅田他: 構文の照合による柔軟なテキスト検索機能を備えた翻訳支援システム. 情報処理学会第3 7 回全国大会講演論文集 **4B-6** (1988).

4. 堤他: 用例検索による教育支援システム. 情報処理学会コンピュータと教育研究会 **89-CE-4** (1989).