

## ヘルプデスク支援システムにおける言語事例検索 —類似文照合—

1 P-3

相川 勇之 高山 泰博 鈴木 克志

三菱電機株式会社 情報技術総合研究所

## 1. はじめに

近年、顧客からの問い合わせ相談窓口であるヘルプデスク業務が、顧客満足度向上のための重要な部門として注目されている[島津 98]。

しかし従来のシステムでは、数値化可能な定型的情報はデータマイニング手法により有効活用できるが、自由文記述による非定型的情報については全文検索程度の非常に限定された活用しかできない。

我々は、ヘルプデスク支援システムにおいて蓄積された非定型的情報を有効活用するために、自然言語処理技術およびオントロジーを用いた言語事例検索システムを試作した。本稿では、同システムの類似文照合処理について報告する。

## 2. 言語事例検索

我々は、豊富な情報をもつオントロジーを対象領域ごとに構築し、これを類似事例検索に利用する事例検索システムを試作した[相川 99]。同システムではオントロジーを用いた推論処理により、文の表層的な解析だけでは得られない情報をも検索可能としている。

しかし、上記試作システムでは解析結果の依存構造にしたがって照合する類似文照合方式だったため、オントロジーの有無による検索精度への影響が大きく、オントロジー構築コストが大きな業務に対する段階的適用が困難であるという問題があった。

そこで、オントロジーなしでも従来の全文検索より木目細かな検索が可能となるよう照合方式を改良し、オントロジーを徐々に増やすことで段階的に検索性能を向上していくことが可能な類似事例検索ソフトウェアを試作した。

## 3. 係り受け情報を用いた類似文照合

図1に、試作した類似事例検索処理の概要を示す。入力文を解析して依存構造を生成する日本語文解析処理と、依存構造を比較して類似度を計算

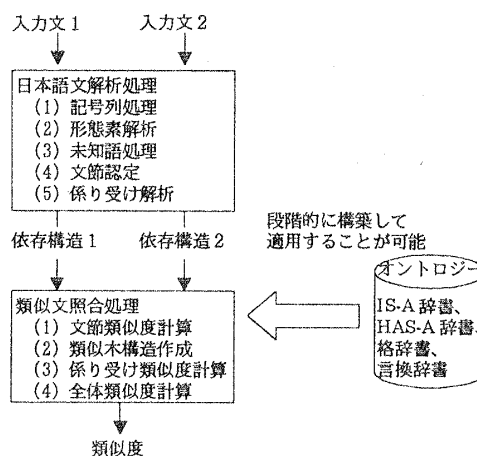


図1 類似事例検索処理

する類似文照合処理から構成され、類似文照合ではオントロジーを参照して類似度を計算する。

## 3. 1. 日本語文解析

依存構造は、事例に含まれる日本語文を形態素解析および係り受け解析することによって得る。ヘルプデスクの応対記録は、電話中のメモ書きなどが元になっており、次の問合せを受け付けるために一刻が争われるので、文法的な厳密さのない文が多い。格助詞の省略なども多いので、係り受け解析は、連用修飾および連体修飾を中心とした緩やかな規則に基づいて行なう。解析処理は下記5段階で行なう。

## (1) 記号列処理

エラーメッセージなどの長単位の記号列をひとつの形態素として処理できるよう前処理する。

（短単位の未知語として処理すると、類似文照合処理における組合せ爆発の原因となるため。）

## (2) 形態素解析

コスト最小法に基づき形態素解析を行なう。

## (3) 未知語処理

カタカナ形態素の連続、1文字形態素の連続などから未知語範囲を推定する処理を行なう。

（例：「情/シ」→「情シ(組織名略語)」）

## (4) 文節処理

形態素解析結果を係り受け解析の単位となる文節にまとめる。ここで接辞処理および様相表現の処理を行ない「～中」と「～ている」などの表現を照合処理で同定できるよう属性に変換する。

Natural Language Case-based Retrieval in Helpdesk Support Systems —Similar Sentence Matching—

Takeyuk AIKAWA, Yasuhiro TAKAYAMA, Katsushi SUZUKI  
Mitsubishi Electric Corporation.

5-1-1 Ofuna, Kamakura, Kanagawa 247-8501, JAPAN

(5) 係り受け解析

係り受け解析には曖昧性の問題および省略表現による解析失敗の問題があるが、類似文照合で用いる必要最低限の単語間の関係をとらえるため、非交差で近い方にかけてという原則により一意に依存構造を求める。また、「Aする。その後、Bする。」という文章と「AしてからBする。」という文を照合可能とするため、複数の文をひとつの依存構造にまとめるように解析する。

3. 2. 類似文照合

類似文照合処理では、日本語文解析処理の出力である依存構造を入力として、2つの文の類似度を計算する。処理は下記4段階で行なう。

(1) 文節類似度計算

依存構造の各ノードである文節間の類似度を計算する。入力された依存構造1と依存構造2のすべての文節の類似度を総当りで計算する。その際、オントロジが存在すれば、IS-A辞書(上位下位関係)や言換辞書(意味的に等価な表現を記述した辞書)を参照して類似度を計算する。

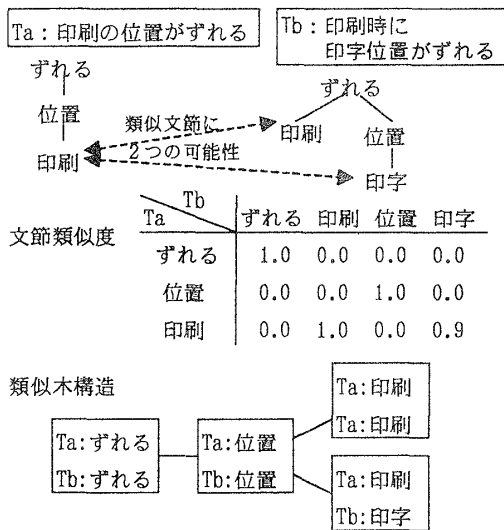


図2 文節類似度と類似木構造

(2) 類似木構造作成 (組合せの可能性展開)

入力される文がある程度長くなると、類似性の強い文節の組合せが複数存在する場合がある。とくに、ヘルプデスクの応対記録のように専門性が高い文章においてその傾向は顕著になる。そこで、まず類似する文節の組合せ可能性を表現する類似木構造を作成する(図2)。

(3) 係り受け類似度計算

次に、図2の類似木構造の各リーフ (=類似文節の組合せ可能性) に対して、係り受けの類似度を計算する。係り受けの類似度とは、類似する2つの文節組合せが共通の係り受けをもつか否か

により計算する。図2の例では、Taにおける「印刷→位置」とTbにおける「印字→位置」が共通の係り受け関係を持つので類似度が高くなる。

(4) 全体類似度計算

異なる部分が多いほど類似していないという方針にしたがい文節類似度および係り受け類似度を計算する。(式(A1)、式(A2))

$$\text{文節類似度} = \frac{\text{類似文節の類似度和}}{\text{合計文節数}} \dots \text{式(A1)}$$

$$\text{係り受け類似度} = \frac{\text{共通係り受け部分の類似度和}}{\text{類似文節ペア数}} \dots \text{式(A2)}$$

全体の類似度は、文節類似度および係り受け類似度に所定の重み付けをした加算平均により計算する。図2の類似木構造において、各リーフについて類似度を計算し、このうち最も良い値を全体の類似度として出力する。

4. 全文検索との比較

類似文照合処理は、計算負荷が大きい。われわれは、あらかじめキーワードの出現頻度に基づいて事例を適切な大きさをクラスタリングしておくことにより、照合処理対象を絞り込んで速度性能を確保している。

現在、実験および評価に用いている計算機関連の障害問合せ記録において、「プリンタ」「印刷」「出る」という3つのキーワードで特定されるクラスタには、下記のような事例が存在する。

事例1: 「プリンタにてはがき印刷時、印字はするが、出てこない」

事例2: 「プリンタから印刷物がしきりに出てくる。既に50枚ほど印刷されている。」

類似文照合の適用により、全文検索とは異なる次のような類似事例検索が可能となる。例えば、「文字化けした印刷がどんどん出てきた」が検索入力だとすると、「印刷がどんどん出てきた」という表現と「印刷物がしきりに出てくる」という表現が構造的に類似しているため、事例1よりも事例2が類似する例として検索される。

5. まとめ

係り受け情報を用いた類似文照合処理について述べた。今後は、オントロジー構築ツールを充実させるとともに、実業務への適用を試み、評価を行なう予定である。

参考文献

[島津 98] 島津, 伊藤, 「ヘルプデスク支援システムの最新動向」, 情報処理 Vol.39 No.9, 1998.  
 [相川 99] 相川, 高山, 伊藤, 鈴木「ヘルプデスク支援システムの最新動向」, 情報処理学会第58回全国大会(2U-1), 1999.