2U－1

ヘルプデスク支援システムにおける言語事例データの類似検索

相川 勇之　高山 泰博　伊藤 山彦　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]。同業務には下記の要求があり、これを支援するためのヘルプデスク支援システムが製品化されている。

(1)過去の事例の検索による迅速、的確な応答。

(2)大量の事例分析による情報抽出および経営、開発へのフィードバック。

しかし従来のシステムでは、数値化できる定型的な情報ならばデータマイニングの手法により有効に活用できるが、自由文で記述された非定型的な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かった。

我々は、非定型的な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類似検索機能を試作したので、本稿で報告する。

2．オントロジー利用の類似検索

インターネット上の検索エンジン等では、類義語辞書を用いたキーワード拡張による検索方式やベクトルモデルに代表される統計的検索方式が実用化されている。しかし、ヘルプデスクにおける応対記録のように専門的な文書では、もともと意味的に近い文書の中からさらに類似性の高い文書を検索する必要がある。そのため、非常に広範囲の文書から大雑把に検索できれば良いインターネット検索とは異なり各文書を単語の使用傾向だけで特徴づけることはできず、文の構造まで意識したより強力な検索方式が必須となる。

2語以上の関係をとらえて非定型情報を活用するための研究として、諸橋らのテキストマイニングに関する研究がある[諸橋 98]。諸橋らは構文解析せず、単語分類を記したカテゴリ辞書のみを知識として用いて、マイニング用の情報抽出を行なっている。一種のパターンマッチにより文内の2語以上の関係を取り出す方法を与えているが、構文解析をしていないため語順の任意性には対処しにくい。またヘルプデスクの応対記録で頻出する省略表記への対処もなされていない。

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、これを類似検索に利用する。入力文を係り受け解析するので、語順の異なる文でも検索でき、さらに、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。図1に、試作したヘルプデスク支援システムの構成図を示す。

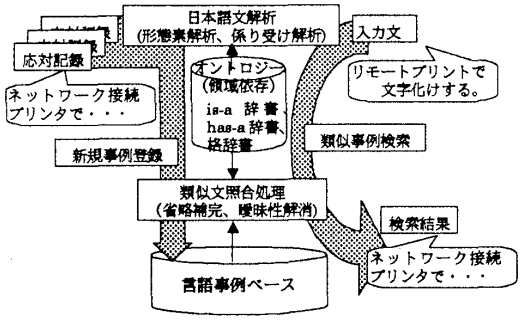

図1　システム構成

このシステムでは、応対記録のような生の事例を文の集合として扱い、個々の文をあらかじめ解析して単語間の関係を木構造で表現した依存構造をつくり、言語事例ベースに登録しておく。検索時には入力文を解析して、言語事例ベースに登録された文と類似文照合を行ない、類似する事例を検索して利用する。本稿では主に類似事例検索の部分について述べる。

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、オントロジーを用いた類似文照合のための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るという原則により一意に依存構造を求めている。こうして得られた依存構造に対して、後述する類似文照合処理において必要に応じて曖昧性

Similar Sentence Retrieval for Natural Language Case Data in Helpdesk Support Systems
Takeyuki AIKAWA, Yasuhiro TAKAYAMA, Takahiro ITO, Katsushi SUZUKI
Mitsubishi Electric Corporation.
5-1-1 Ofuna, Kamakura, Kanagawa 247-8501, JAPAN

を解消し、また省略等を補うというアプローチをとっている。このようなアプローチをとることにより、例えば「リモートプリンタで文字化けする」という入力検索文に対して、

・リモートプリンタ＝ネットワーク接続プリンタ（is-a 知識）
・文字化け＝プリンタ出力の障害（has-a 知識）
・出力が文字化けする（格知識）
・プリンタで出力する（格知識）

などの知識を用いて、「ネットワーク接続したプリンタの出力が文字化けする」という事例を的確に検索できるようになる。

3．類似文照合処理

類似文照合処理は、依存構造の類似性を判定する処理である。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。

3．1．使用するオントロジー

類似文照合処理には、以下のオントロジーを利用する。

(1) IS-A 知識： 同義語や類義語を IS-A 知識として記述する。文の要素である単語間の類似性を計算する際に、IS-A 階層上での距離を使用する。

(2)HAS-A 知識： 部分全体関係を HAS-A 知識として記述する。IS-A 知識と同様に単語間の類似性を計算するのに使用する。また、ある機器が属性としてとり得る状態に関する制約知識としても利用する。

(3)格記述： 格助詞の省略等を補なうために使用する。また依存構造の類似性を計算する際に、格要素の不一致には大きなペナルティを与えるなど、構成要素の類似度に重みづけを与えるためにも使用する。

3．2．類似度の計算

完全に同一の表現同士の類似度を１として、類似度は０から１の値をとるものとする。類似度の計算は、依存構造のルート側（受け側）から順次ノードを対応づけることにより行なう（図２）。ノード間の類似性と構造の類似性と２つの観点から類似度を計算する。類似度は、ノード間および構造間の「異なり度合い」をペナルティ値として表現して、累積していくことにより計算する。

ノード間の異なり度合いは、IS-A 知識や HAS-A 知識などのオントロジー上での距離にしたがって与える。このとき、用言であれば推量表現や否定表現などのモダリティについても考慮した点数づけを行なう。例えば、一方に否定表現があれば高いペナルティ値を与える。

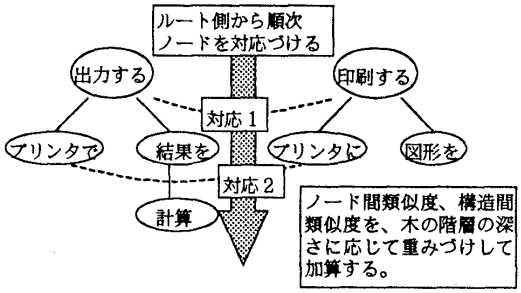

図２ 類似度の計算

構造間の異なり度合いは、対応のとれなかったノード（不一致ノード）へのペナルティ値として与える。その際に、不一致ノードが依存構造に対して果たす役割の強さに応じてペナルティ値の重みを調整する。例えば、格記述において重要な格となっている場合には、不一致ノードのペナルティ値を大きくする。

依存構造のルート側すなわち受け側のノードが、その文の意味的な中心部分をになうと考えるのが自然である。そこで、階層が深くなれば、より詳細な相違を示すと考え、ペナルティ値の重みを小さくする。

接頭語および接尾語については、解析時にノード内に属性として埋め込まれるので、一種のモダリティ情報として扱うことが可能である。例えば、「～中」と「～している間に」という表現が類似の意味内容を示すという計算ができる。

4．まとめ

オントロジーに基づいた類似事例検索について述べた。現在、社内ヘルプデスク部門など数種類の分野について適用を検討中である。現時点ではオントロジー構築には多くの手作業を要しているが、今後は、専門用語の抽出、格辞書の自動獲得など自動化をはかっていく予定である。

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.

[諸橋 98] 諸橋,那須川,長野,「テキストマイニング：膨大な文書データからの知識獲得―意図の認識―」，情報処理学会第 57 回全国大会(5K-3),1998.