キーワード抽出方式についての検討

3Y－1

畑本 直樹　　岩瀬 成人

NTT 情報通信研究所

1　はじめに

企業やお店の名前、電話番号、住所、営業内容等の情報を案内するイエローページサービスにおいては、ユーザインタフェースの向上が重要である。情報検索が容易な入力方式として、自然言語処理技術と推論処理技術を適用した自然言語インタフェースの情報案内システムを既に開発している[1,2]。このシステムでは、例えば「渋谷で酒が飲める店」といった、日常使う自然言語テキストを入力するだけで渋谷の居酒屋を検索案内することができる。

本システムを実現するにあたっては、お店で扱う商品をシソーラスの形で登録する必要があり、省力化の面から登録作業の自動化が望まれていた。シソーラスの自動生成のためには、第1段階として、お店の案内文を解析し商品をあらわすキーワードを自動的に抽出する必要がある。

従来、文献や新聞記事等で自動的にキーワードを抽出する方式としては統計的手法による頻度情報を元にした方式が用いられていた[3]。しかし、お店の案内文は文の長さが100字程度なので、重要なキーワードが出現頻度のみでは特定できないこと、複合語の形でキーワードが出現するので、単純な単語の頻度情報ではキーワードが特定できないという問題点があった。

そこで、本稿では、文の係り受け関係を解析し、文型の上で重要なキーワードを特定する文型解析、及びシソーラス上で近い意味を持つキーワードの頻度をカウントすることにより重要な商品キーワードを特定するキーワード抽出方式を提案する。

2　構成

図1に商品キーワード自動登録の全体構成を示す。

案内文を解析し、単語の意味及び係り受け関係を求める自然言語解析部、係り受け関係を元に案内文に出現した位置で商品が重要か否かを判定する文型解析部、商品を表すキーワードを抽出し、出現頻度で重要度を決めるキーワード抽出部、抽出した商品キーワードを各種シソーラスに登録するシソーラス登録部からなる。

以下に、検索方式の中心となる文型解析とキーワード抽出について述べる。

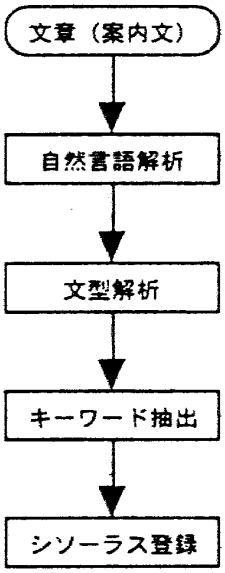

図1．商品キーワード自動登録の全体構成

3　処理概要

3.1　文型解析

同じ商品が案内文に出現しても出現した位置で重要度が異なる。

例えば、「テレビ、ビデオ、各種電気製品を扱っております」という文中における「テレビ」は重要な商品キーワードであるが、「テレビで紹介されたお店です」という文における「テレビ」は商品キーワードとしては重要ではない。そこで、キーワードとして重要な単語、あるいは重要でない単語をルールとして記述し、ルールを参照しながら、解析をすすめる。その際、単語の意味、係り受け関係が記述できるようにした。評価は頻度を増減するような形で出力するようにした。例えば、評価“0”は案内文に出現しなかった場合と同じであり、“1”は1回出現した場合と同じであり、“2”は2回出現した場合と同じ重要度である。

表1にルール例を、図2に解析例を示す。

A Study of Keyword Extraction Method for a Yellow Page Service
Naoki Hatamoto, Shigehito Iwase
NTT Information and Communication Systems Laboratories
3-9-11 Midori-cho, Musashino city, Tokyo 180, Japan

表1．文型解析に用いるルール例

文型	評価
([名詞節1]:〈物〉) (名詞節2:〈店代名詞〉% 連体修飾: 名詞節1) (動詞節1:"有名だ"% 理由格: 名詞節2)	2
([名詞節1]:〈物〉) (動詞節1:"紹介される"% 手段格: 名詞節2)	0

[ ] ：評価の対象とするキーワード
〈 〉 ：意味
" " ：単語実体
% 以降 ：係り受け関係

例文：「和食の老舗で有名だ」

				評価
名詞節	和食（料理）	の	→	2
名詞節	老舗（店代名詞）	で	→	1
動詞節	有名だ		→	1

図2．解析例

3.2 キーワード抽出

キーワード抽出では案内文に含まれる商品キーワードを重要な順に抽出する。案内文では商品を箇条書きで羅列し、同じキーワードは1回しか出現しない場合が多い(例: テレビ、ビデオ、オーディオ、電話、FAX各種取扱い)。このような場合でも案内文で重要なキーワードを特定するために、シソーラスを用いて、近い意味のキーワードはまとめて頻度をカウントすることで重要なキーワードの判別を行なう。

ここで、どこまでの商品を近いと見なすかが問題になる。ここでは、分野に結び付く[4]最上位の商品を代表商品分類と定義し、同じ代表商品分類をもつ商品は「近い」商品であると定義する。例えば、上記の例の場合、分野は電気店なので、代表商品分類は「家電」となる。「テレビ」「ビデオ」「オーディオ」は「家電」なので、頻度は3になる。

もう1つの問題点は案内文では商品が複合語で表されている場合が多いことである。例えば、「電気製品販売」「競馬情報」のように、末尾にサ変名詞や語尾がつく場合や、「貸しおしぼり」のように先頭につく場合もある。この場合のキーワードはそれぞれ「電気製品」「競馬」「おしぼり」なので、末尾のサ変名詞や先頭の「貸し」などを自動的に削除して商品キーワードとする処理を組み込んだ。さらに、削除したサ変名詞にも動詞としての意味があるので、商品キーワードとペアでサ変名詞や語尾と抽出するようにした。「電気製品販売」の場合は「電気製品」と「販売」がキーワードとして出力される。

以上を考慮した処理の流れを図3に示す。

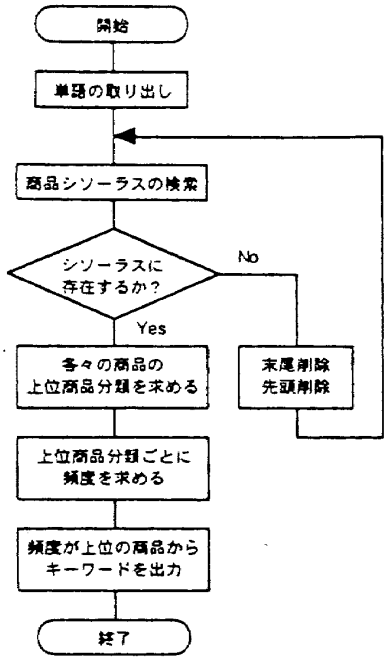

図3．キーワード抽出部の処理の流れ

4 今後の課題

本稿では、イエローサービスに用いられる案内文から、検索に用いられるシソーラスを補完拡充するためのキーワードを抽出するシステムにおいて、特定のパターンで文中に出現する単語の重要度を求める文型解析、及びキーワード抽出方式について提唱した。今後、実装を行ない実際のデータに対し適用することにより評価を行ない、またルールの拡充を図っていきたい。

参考文献

1. Miharu Tobe et al., An Intelligent Directory System with an Inference Function, GLOBECOM'96, 1996

2. インターネットハローダイヤル, http://hello.nttls.co.jp/

3. 長尾 真編, 自然言語処理, 岩波書店, pp419

4. 岩瀬・大山, 自然言語処理技術を用いた職業別電話帳検索の高度化, 電子情報通信学会論文 Vol. J74-D-2, pp1255-1263, 1991