企業名の検索方式の高度化

2C−6

岩瀬　成人　　高橋　克巳

NTT情報通信網研究所

1. はじめに

企業名をキーにして電話番号を検索する場合、複数単語からなる長い企業名では後方の単語が曖昧になる性質を利用して、前方一致で検索を行っている。しかし、先頭の単語が省略される場合も多数ある。検索時に企業名を解析して検索条件を派生させる方法は文献(1)で述べた。ここでは、データベースに問い合わせる可能性のある表現をあらかじめキーとして登録しておく手法について検討を行った。

まず、企業名の性質について報告する。次にユーザが問い合わせる可能性のある企業名を言語処理技術を用いて作成する方法について検討したので報告する。

2. 企業名の性質

企業名を解析するためには、企業名を単語に分割する必要がある。一方、企業名は以下の性質を持つ。

①他の企業と区別するため、平仮名や片仮名の名称にする（ビューティ・ハラダ）
②略称が多い（日銀）
③単語の途中で字種が変わる場合がある。（つぼ八）
④外国語をカタカナ表記した企業が多い。（ラ・モード・サントノーレ）
⑤当て字の企業名が多い。（亜羅仁：アラジン）

従って、辞書未登録語は必ず含まれることを前提にして解析を進める必要がある。

次に、企業名を構成する単語を12種類の意味に分類した（表1）。企業を認識する上で中心的な単語を固有名と定義する。通常、固有名は姓・企業名から構成されるが、職業や地名の組み合わせの場合もある。一般にユーザからの問い合わせで固有名が省略されることはほとんどない。また職業・住所は表現そのものよりも、意味を記憶している場合が多い。従って、実際の企業名と異なる可能性が高い。

表1　企業名の構成単語の意味分類

意味		例
固有名	姓・企業名	田中、三和
	名	清、和子
職業等	職業	ホテル、レストラン
	職業語尾	店、学院
	商品	法律、電器、英会話
	サ変名詞	販売、派遣
地名等	住所	東京、札幌
	方向	東、西、南、北、上、中
付属語・冠詞		新、ザ
下部語尾		支店、営業所
番号		第1、第2
その他		公認、国立、ニュー

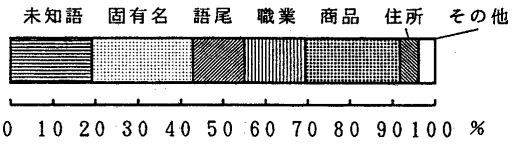

図1　意味別の単語出現頻度

この分類をもとに全国の企業名義（約900万件）を分割した結果を図1に示す。ここで用いた辞書は1万6千単語で、出現頻度が100回以上の単語を登録した。この図より、企業名と商品が20%あり、名義を認識する上で重要な役割を果たしている。また、未知語が20%含まれる。1企業名あたり2．4単語であることが同じ調査で分かっているので、平均40%

A study of retrieval method on Yellow Pages
Shigehito IWASE, Katsumi TAKAHASHI
NTT Network Information Systems Laboratories

(=(1-(1-0.2)$^{2\cdot4}$)*100) の名義に未知語が含まれていることが分かる。このことからも、未知語対策が重要なことであることが分かる。

実際の解析ルールでは未知語文字数、字種の変化の有無、単語間の係り受け関係の有無を組み合わせた評価関数を用いて、関数の値が最良になる解を解析結果とした。

3. 名義キーの作成

電子番号案内システムでは、企業の別称でも検索できるように読み替え名を登録する機能がある。ここに登録されている読み替え名を調査したところ、本来の別称以外に企業名の先頭を除いた部分を登録してある例が多く見受けられた（除かれた部分を冠称名と呼ぶ）。これらを分類したところ表2の結果になった。この中で企業名冠称以外は単語の意味によって冠称部分を判断できる可能性がある。この分析をもとに約20のルールを作成した。主なルールを表3に示す。

表2　読み替えの分類

分類	例
職業冠称	ホテル／はまなす
地名冠称	紋別市／ミンク組合
企業名冠称	昭和／シェル
国立冠称	国立／弟子屈病院
番号冠称	第2／樋口歯科
冠詞冠称	ザ／ジュウタン屋
ニュー冠称	ニュー／シャルマン

／以降が読み替えに登録されている

表3　主な冠称名認識ルール

意味パターン	例（／までが冠称）
職業等＋固有名	カットサロン／パリ
地名等＋固有名	南北海道／保健センター
番号＋固有名	第2／樋口歯科
ニュー＋固有名	ニュー／シャルマン
地名＋料理＋固有名	札幌ラーメン／どさんこ
その他＋語尾＋固有名	バラエティショップ／ヒロタ

4. 評価

表3のルールで実際に冠称名が認識できるか職業分類が美容院である企業3千件を用いて評価を行った。辞書の単語数は1万6千語で、言語はC、解析方法はBUP[2]を用いた。図2にその結果を示す。この図より96%は正しく冠称名を認識している事が分かり、本手法が有効である事が分かる。なお、単語分割が正しいのに冠称名を誤った例には、単語の多義が解消できないために誤解析した場合が多い。例えば、「フラワーショップ高橋」の場合、「フラワーショップ」は職業なので「高橋」という検索キーを作る。しかし、「フラワー美容室」の「フラワー」は商品なので、職業冠称になる。従って、「美容室」なるキーを作ってしまう。これについては、名義の職業が美容院であることを解析時に使用することを検討している。

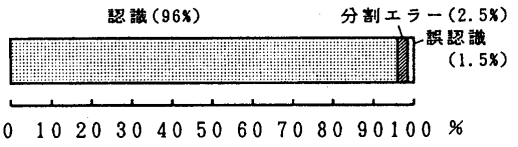

図2　冠称名の認識率

5. まとめ

企業名の先頭を省略される問い合わせに対処するため、言語処理技術を用いて、入力される表現を判断し、データベースにあらかじめ登録する方法を提案した。その結果、職業分類が美容院である企業では96%は正しく冠称名を認識していることが分かった。今後は、①職業情報を利用した単語の多義解消法の検討　②実際の問い合わせによる評価を行う予定である。

文　献

1)岩瀬、大山、橋田「企業名の普通名詞分割」信学論誌D　Vol.70-D No.4 pp.832-835 1987.4

2)松本他「Prologに埋め込まれたBottom Up Parser:BUP」情処学会自然言語研究会34-6, 1982