自然言語処理を用いた住所解析方式

1S－8

岩瀬成人　唐沢裕明　早坂秀雄　池田謙司
ＮＴＴ情報通信研究所

１．はじめに

顧客データをコンピュータで管理・検索する上で、住所情報は顧客を特定するための重要な入力条件の一つである。住所解析とはこの住所として入力された文字列を住所コードに変換する処理であり、入力の容易化、曖昧さへの対処が要求される。コード化対象の住所は①都道府県　②市区郡町　③町大字　④字丁目の４階層に分かれている。このため、従来の住所解析では(1)住所階層毎に区切って入力する方法　(2)各住所階層の先頭文字を入力してガイダンスにより確定する方法などが用いられていた。しかし、この様な方法では町大字、字が複数の単語から構成されることも多く、区切りを誤って入力する場合が多い（表１）こと、また、住所階層を誤ると各住所階層の先頭文字を入力しても目的とする住所は得られない等の問題点があった。そこで本稿では、自然言語処理技術を用いて住所階層を意識しないで住所を入力する方法を報告する。

２．自然言語処理技術適用の問題点

住所解析に自然言語処理を適用する上での問題点は形態素解析と意味解析の役割分担である。住所はほとんど地名と接尾詞（県、市、区、町等）で入力されるので一般的な自然文の文法は適用できない。従って、住所の包含関係（Ａ町がＢ市に含まれるとき包含関係があると呼ぶ）によって解を絞り込むことになる。字までの住所は全国で４６万件（別読みを含む）あり、１・２文字のカナ地名はほとんどすべての文字の組合せが存在する。形態素解析でこの辞書を用いて完全な包含チェックを行うことは解候補の件数が多くなりすぎて実用的な実行速度が得られない。

本稿では形態素解析に県レベルの包含チェックを行うことにより解候補の件数を絞り、絞った解候補に意味解析を行う手法を提案する。

表１　住所の入力誤り例

誤りの種類	入力例
町名に区が含まれる	誤：姫路市△飾磨区△英賀春日町 正：姫路市△飾磨区英賀春日町
町名に大字が含まれる	誤：熊本市△清水町△大字新池 正：熊本市△清水町大字新池
通りの区切りと異なる	誤：札幌市△中央区△北３条△西１丁目 正：札幌市△中央区△北３条西△１丁目

△：スペース

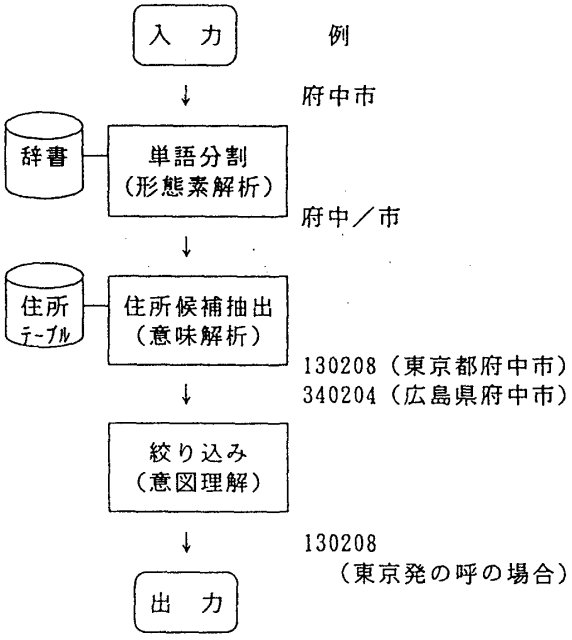

図１　住所解析の構成

A Study of Address Analysis Using Natural Language Processing
Shigehito Iwase, Karasawa Hiroaki, Hayasaka Hideo, and Kenji Ikeda
NTT Information and Communication Systems Laboratories
3-9-11, Midoricho, Musashino-Shi 180, Japan

3．住所解析方式

図１に住所解析の構成図を示す。各々のモジュールは以下の機能を持つ。

(1)単語分割：辞書を用いて入力された文字列を単語に分割し、市区町等の意味を付与する。いわゆる形態素解析を行う部分である。

(2)住所候補抽出：分割された単語から住所コードを求める。その時、各単語が包含関係にあるかチェックする。
一般の自然言語処理の意味解析に相当する。

(3)絞り込み：解析結果が複数ある時、発呼者の住所、濁音の一致数、接尾詞の一致数等により解を絞り込む。

単語分割で包含チェックを行う場合、単語辞書に住所コードを持ち込む必要がある。ところが、一つの単語には数多くの住所コードが対応する。例えば、「アイオイ」は全国の市と町だけでも70存在する。そこで、辞書が巨大になるのを防ぐため、ある単語が属する住所を県のレベルのビット列で持つ構成とした。例えば、「アイオイ」は県市区郡レベルでは兵庫県と徳島県、町大字レベルでは北海道、山形県、神奈川県・・・に存在することをビット列で表したフラグで持つ（表２）。住所コードをフラグで持つことにより辞書量の増加は１単語当たり１８バイトの増加(47bit*3) で済む。

また、この様な辞書構成にすることにより次の様な利点が生まれる。①包含チェックが単にビット演算で容易に実現できる。②辞書で県までの情報を持つため、住所候補抽出で必要な住所テーブルを県単位に分割することが出来る。従って、サーチするデータ数が減少し、テーブルアクセスが高速になる。

4．実験結果

本手法の有効性を検証するためにＣ言語でプロトタイプを作成した。住所テーブルは字まで含めて約４０万件、単語辞書は１６万単語になった。問い合わせ例１０００件を解析したところ平均の単語分割の解件数は２件であり、９９％は８件以内に収まった（図２）。入力の平均文字数は８文字であったので、包含チェックを行わなければ約２００件の解候補が出力されるので、約１／１００の絞り込み効果が得られた。

5．まとめ

住所解析に自然言語処理を導入することにより住所階層を意識しない入力方式を検討した。その結果、形態素解析に県レベルの包含チェックを行うことにより解候補の件数を１／１００に絞り込むことができ、有効性を確認できた。

表２　単語辞書の構成

項目	例
見出し語	アイオイ
品詞	名詞
意味	地名
県フラグ	見出し語の存在する県をビット列で表したフラグ。下図参照

県フラグの例

レベル＼県	北海道 青森 岩手 宮城 秋田 山形 福島 ・・・・・・ 沖縄
県	0 0 0 0 0 0 0 0
市区町大字	1 0 0 0 0 1 0 0
字	0 0 0 0 0 1 0 0

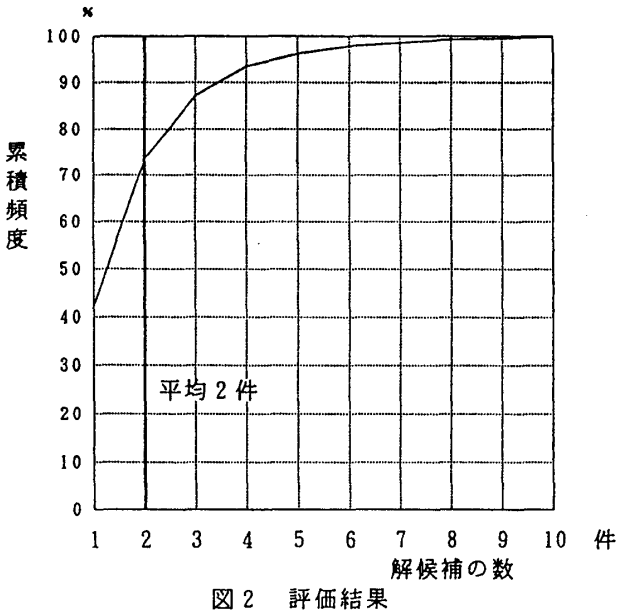

図２　評価結果