6E−8

職業名の自動分割法

高橋　克巳　　藤原　進
ＮＴＴ情報通信処理研究所

１．はじめに

’９０年１２月に、５０音別の電話帳をオンライン検索するサービスが開始された。職業別の電話帳に関しても効率よくオンライン検索する方式が研究されている。

職業別電話帳は、お店や会社を約１８００業種に分類し掲載しており、さらに類義語辞書である、職業さくいん（見出し語とよぶ）を設け、検索上の便宜をはかっている。

筆者らは職業別電話帳情報をキーワード検索する手法の研究を進めているが、職業名を表す、見出し語、および利用者の問い合わせは、ともに複合語であり、その有効な分割手法と解析が必要である。

本稿では、見出し語間での単語の重複情報を用いて、見出し語を自動的に分割する方法を検討し、見出し語の解析を行なったので報告する。

２．見出し語の特徴

２．１見出し語の例

職業別電話帳は、業種を約１８００に分類している。見出し語は分類語である１８００の業種名と、その類義語からなる約８０００語の集合である。

表１　見出し語の例

業種名	類義語
預かり業	一時預かり
	コインロッカー
	品物預かり
	荷物預かり
	自転車預かり
	物品預かり
マーケット	市場
	スーパーマーケット
	マート

２．２見出し語の特徴

見出し語には、

１）同一分類中には同じ単語が何度も使われる。

２）各見出し語で共通した末尾語が多い。

といった特徴がある。

特徴１）、２）について予備調査を行なった。調査により、次のことが明らかになった。

a）全職業分類中、７割をこえる、１２８５分類に重複単語が存在する。

b）出現回数の多い順から１２０個の末尾語（5回以上出現）は、合わせて２２３４の見出し語の末尾になっている。

以下に結果を示す。

表２　職業の各分類における重複単語の出現

重複単語１	６４２
重複単語２	３９３
重複単語３	１５８
重複単語４以上	９２
重複単語なし	４６５
総職業分類数	１７５０

（単位　職業分類数）

（表２について）

同一分類中で、２つ以上の見出し語に使用されている文字列（単語候補）を重複単語として数える。

表３　代表的な末尾語

道場	106	用品	69
工事	108	器具	61
機械	99	料理	54
販売	97	製品	48
製造	96	印刷	41
軽食	86	装置	39
団体	85	工業	37
材料	82	診療所	36
教室	78	貿易	32
学校	74	加工	31

（単位　見出し語数）

（表３について）

見出し語の末尾に出現する文字列（単語候補）を出現回数の多い順に、上位２０を紹介する。抽出処理は全見出し語について行なった。

Automatic Segmentation Method for Occupation Heading
Katsumi TAKAHASHI　Susumu FUJIWARA
NTT Communications and Infomation Processing Laboratries

ここで重複単語および末尾語を抽出する際に、以下の基準を設けた。

イ）２文字以上
ロ）包含関係で単語候補が複数現われた場合　（例「物預かり」と「預かり」）、短い方（「預かり」）を採用する。
ハ）カタカナの重複文字列は単語候補としない。
（例　マーケットとマートの「ー」）

３．見出し語の性質を利用した自動分割

以下に分割処理の概要を示す。処理はａ）からｃ）の順で行なった。

ａ）文字属性による分割
カタカナと漢字が混在する見出し語を、カタカナと漢字に分割する。点（・）かっこは分割する。

ｂ）重複文字列による分割
同一業種の見出し語内で重複する文字列を検出し、その業種内にかぎり単語候補とする。

ｃ）末尾語による分割
全見出し語にわたって、重複して出現する末尾語を抽出し、そのうち頻出する末尾語を、全見出し語内で、単語候補とする。ただし、条件イ）ロ）ハ）に加え、ニ）全体で５回以上出現したものを単語候補とする。

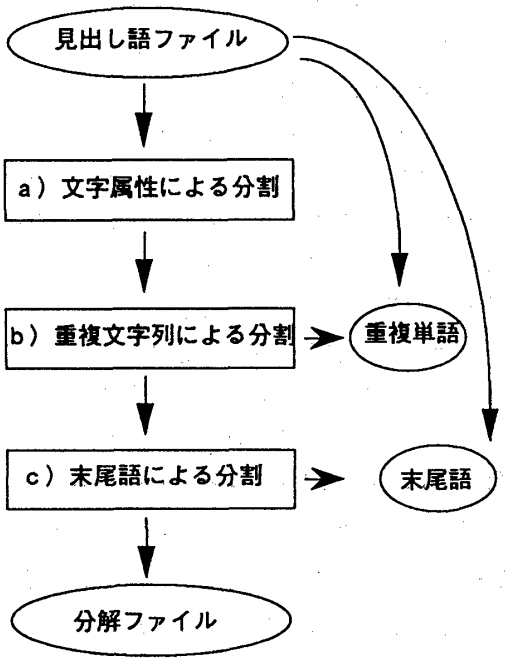

図１　見出し語の性質を利用した自動分割

４．実験結果と考察

本方法を用いて職業別電話帳の見出し語について分割処理を行なった。その結果、見出し語７９６４単語を、４３７５の単語に分割することができた。
処理別では処理ａ）で２０１４、ｂ）で４８６８、ｃ）で４８２見出し語が分割された。３処理合計では４７６４見出し語が分割された。

このうち２０％程度の見出し語は、分割が不十分でさらに分割することが必要であった。これは
・カタカナを単語候補から外している
・一文字の単語候補を外している（特に末尾語で一文字のものは多い）
・単語候補を獲得する処理が、ｂ）とｃ）のみでは不十分であったため
と考えられる。

また、数％の見出し語は、誤った分割処理、あるいは分割が一意に決定できない状況が生じている。これは見出し語だけではサンプル数が不足していることと、単語候補が正しいものであるかという検証手段がなかったためである。

５．おわりに

今回の検討により、職業名が関連した職業名の情報、及び単語の構成情報のみから、自動的に分割できる見通しを得た。また本処理で、職業の見出し語から、重複単語辞書および、末尾語辞書を自動的に作成することができた。
これらは、職業の解析および分類に有用であり、職業情報検索システムにおいて、利用者からの問い合わせの解析に利用できるので、検索の性能を向上させることができる。
今後はこれらの辞書を使った、未分割語の再分割処理を検討し分割の精度をあげる検討を行なう。また両辞書は出現頻度という重みを持った集合であるので、４．で指摘した分割の検証への利用を検討したい。

参考文献

１）岩瀬ほか：番号案内における職業推論の検討
1990年通信学会春期全国大会
２）宮崎：係り受けを用いた複合語の自動分割法
1984年情報処理学会論文誌
３）杉森：レ・パージュ
1985年ベルブックス