大規模コーパスに基づいた未登録語の傾向分析

7G－7

飯盛可織†　佐川雄二†　大西昇†　杉江昇‡

†名古屋大学工学部情報工学科　‡名城大学理工学部電気電子工学科

1　はじめに

自然言語の解析を行なう場合、辞書に登録していない語（以下、未登録語）が現れることが多い。そしてこの未登録語は意味情報を持たないという点で、解析の大きな障害となる。現在の機械辞書の中には、十数万語もの見出し語を持つ大きなものも存在するが、年々増加していく新語をすべて登録し続けていくことは不可能である。書き言葉の場合では、英文はわかち書きがされているので未登録語を特定できるため、未登録語の品詞の推定[1]や、意味の推定[2]に関する研究も行なわれている。しかし、日本語はわかち書きがされていないため未登録語を含む文章を解析する際の一番の問題点は、未登録語の切り出しである。

自然言語処理の中でも特に話し言葉の解析においては音声認識の問題や、言い誤りなど解析システム上で未登録語となってしまうような語が多数出現することなど、書き言葉にはない多くの問題がある。そこで、未登録語が障害とならないような解析システムが必要とされる。話し言葉を対象としたシステムを考える場合、音の情報のみであり、漢字、カタカナなどの文字情報を考慮に入れることはできないので、まず最初に問題になるのは未登録語の限定および抽出である。

日本語は元々外来要素が多く、そのほとんどが名詞成分であり、約８割は原語が英語のものである[3]。このことから、未登録語においても外来語が多数存在していると考えられる。こういった、未登録語の傾向を積極的に利用することによって、未登録語を含む文章をよりよく解析ができるのではないかと考え、大規模コーパスを用いて話し言葉中に現れる、未登録語の傾向分析を行なうことにした。

Large Corpus-Based Analysis of the Characteristics of Unknown Words in Japanese

Kaori Isagai†, Yuji Sagawa†, Noboru Ohnishi†, Noboru Sugie‡
†School of Engineering,Nagoya Univiersity
Nagoya 464-01, Japan
‡Faculty of Science and Technology,Meijo University
Nagoya 468, Japan

2　対象

大規模コーパスと形態素辞書を用いて、コーパスから未登録語を抜き出し、その未登録語についての分析を行なった。

まずコーパスは、ＡＴＲ自動翻訳電話研究所によって作成されたＡＤＤ[4]を使用し、単語テーブルに登録されている語のうち、約６万語の中から未登録語を抽出した。

次に、用いた辞書であるが、ＩＣＯＴによって作成されたＴＲＩＥ辞書[5]を用いた。

3　方法と結果

3.1　品詞の分類について

まず、未登録語として現れたものについて品詞別の数についての分析を行なったところ、表１のようになった。

表 1: 品詞別の割合

品詞	出現回数	割合（%）
間投詞	２２６３	５０．２
数詞	７９６	１７．７
固有名詞	５１３	１１．４
普通名詞	３１７	７．０
接続詞	１２７	２．８
⋮	⋮	⋮
総数	４５０８	１００．０

3.2　外来語の割合について

品詞の分類を行なった結果、特徴として挙げられることには、

1. 名詞には、外来語（カタカナ語）が多い。（例：外来語の割合は、普通名詞中に８７％、固有名詞中に３０％）

2. 未登録語として現れる外来語は、圧倒的に英語からのものが多いと考えられるが、英語で２～５の単語で構成されている名詞とは限らない語（例：テクニカルツアー）が、日本語では一つの名詞の単語としての役割を果たしている。

3. 固有名詞に現れている語については、地名や人名などで辞書に登録されていないものが多いが、語尾に「駅」「先生」などの未登録語を切り出すための手がかりを持った語も多数存在する。

3.3 外来語の音素列について

以上の結果に基づいて、特に未登録語を含む文の処理の障害となりそうな外来語に焦点をあてて分析を行なった。

まず、単語間距離 [6] の比較を行なった。単語間距離とは、音素が同じ場合の距離は０、異なる場合は１として、単語の音素列の距離を計るものである。結果は、表２のようになった。

表 2: 単語間距離の平均値

		外来語を含まない単語	外来語
外来語を含まない単語	普通名詞	3.42	5.94
	固有名詞	6.02	7.11

やはり、日本語と日本語を比較した場合よりも、日本語と外来語を比較した場合の方が、全体的に単語間距離が大きいことがいえる。

このことから、外来語の音素列について、日本語の音素列にはにはない特徴を持っていないか調べるため、音素列についてのストリング・マッチングを行なった。表２は、外来語（３５９語）とそれ以外の語（２０７４語）それぞれについての、出現頻度の高い４以上の音素を含む音素列である。

表３からわかるように、外来語に限らず多く出てくる音素列の並びも存在すると考えられるが、外来語だけに多く現れる音素列、つまり、日本語だけの場合には、あまり現れない音素列がいくつか確認された。

表 3: 出現頻度の高い音素列

外来語	外来語以外
syon	zjun
iNgu	jaku
suto	teki
eRsyoN	zyou
puro	roku
uraN	syou
kusu	ousi
oRru	kyou

4 考察

大規模コーパスから未登録語を抽出してその傾向を調査した結果、品詞別には名詞、固有名詞が多く、しかもその大半は外来語であるということがわかった。

次に、外来語には外来語以外の語とは違う音素列の特徴を持つことがわかった。しかし、今回の結果から実際の音素列においても外来語の音素列にはある程度の特徴があるといえる。この特徴を未登録語を含む文の解析に積極的に用いることによって、未登録語が外来語であった場合、未登録語の切り出しが容易になると思われる。

参考文献

[1] 宇佐美重之ほか:未登録語を含む英文の構文解析システム電子情報通信学会技術研究報告　NLC90-49,pp.1-8(1991)

[2] 山田一郎ほか：英文における未登録語の意味推定の検討、情報処理学会自然言語処理研究会資料 NL93-9(1993)

[3] 玉村文郎：日本語における外来要素と外来語、日本語教育　第 74 号、pp.13-27(1991)

[4] 江原暉将ほか：ＡＴＲ対話データベースの内容、ATR Technical Report(1990)

[5] （財）新世代コンピュータ開発機構：ＴＲＩＥ辞書ユーティリティー (1991)

[6] 中川聖一、義永洋士：誤りを含んだ音素系列からの候補単語の検索、計量国語学、vol.14 、no.8 、pp.327-334(1985)