要求文解析によるソフトウェアの自動呼び出し手法

矢野 篤志 † 奥村 紀之 †
† 長野工業高等専門学校 電子情報工学科

1 はじめに

本研究では, 日本語文の入力から, 求めているソフトウェアを呼び出したり, 新たにコンピュータにインストールされたソフトウェアの種類を自動的に判別し, データベースを更新できるシステムを開発する.

2 研究概要

本研究では, 自然言語文章の入力に対し, ユーザの要求に合うようなソフトウェアを呼び出し, 新たにインストールされたソフトウェアの自動分類も行うシステムを提案する.

本システムは, 職種判断システム [1] の職種名選出処理手法を利用して, ソフトウェアの選出を行い, その時に用いる知識ベースの自動更新を行うものである. 職種判断システムとは, ユーザーからの入力に対して適切な職種名を返すシステムである. 図1に, 本システムの構成を示す. ここで, ソフトウェア選出及び分類に使用される知識ベースを"ソフトウェア分類知識ベース"とし, ソフトウェアの選出と分類で共用する.

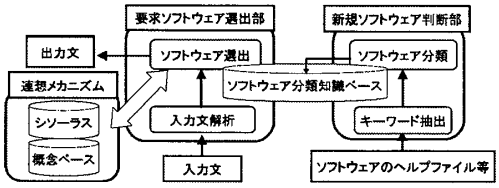

図 1: システム構成

2.1 入力文解析

本システムでは, 入力文に対し形態素解析器 MeCab[2] を用いて形態素解析を行い, 要求名詞と要求動詞を取り出す.

要求名詞と要求動詞について, "○○を××したい." といったような入力があった場合, "○○"に対応する目的語 (一般名詞) が要求名詞であり, "××"に対応する用言が要求動詞である. 要求名詞には目的語 (一般名詞) を, 要求動詞には動詞, 形容詞, サ変接続名詞をそれぞれ一語ずつ入力文から取り出す.

2.2 ソフトウェア選出

本システムでは, 職種判断システムの手法を基にソフトウェア選出を行うため, ソフトウェア分類知識ベースを作成する.

ソフトウェア分類知識ベースは, "ソフトウェア名", "キー動詞", "動詞", "代表語"の4個のフィールドで構成されている. "ソフトウェア名"には, ソフトウェア名が登録されている. "キー動詞"には, "ウェブブラウザ"ならば"閲覧"のように, そのソフトウェアが, どのような目的に使われるかを表す語が登録されている.

"動詞"には, そのソフトウェアと密接な関係がある動詞が登録されている. "代表語"には, "ウェブブラウザ"ならば"インターネット" のように, そのソフトウェアが何をするのか, 何に使われるのかを表す複数の語が入る.

表 1: ソフトウェア分類知識ベース (初期9個)

ソフトウェア名	キー動詞	動詞	代表語
文章作成ソフト	(無し)	書く	文章
ウェブブラウザ	閲覧	見る	インターネット
⋮	⋮	⋮	⋮

2.3 キーワード抽出

新規にインストールされたソフトウェアの種類を分類するために, ソフトウェアに付属するヘルプファイルや Readme から, そのソフトウェアの特徴を表す語, "キーワード"を抽出する. その方法として, 本研究では情報検索システムで広く用いられている"$tf \cdot idf$"を用いて, その上位20位をキーワードとして抽出する.

"$tf \cdot idf$"では, 文章"d"における索引語"t"の重みは式 (1) で与えられる.

$$tf(t,d) \cdot idf(t) = tf(t,d) \cdot (\log \frac{N}{df(t)} + 1) \quad (1)$$

"idf"の算出には, 母数として, 大量の文章集合が必要となる. そこで, 本研究では, "idf"を算出するための文章集合として, 辞書や新聞等から得られた一般的に利用する大量の語が登録されている概念ベース [3] を使用する. キーワード抽出を行う文章は, ソフトウェアに付属するヘルプファイルや Readme とするが, その中でも特にソフトウェアの概要, 若しくは"はじめに"の部分を用いる. キーワードは, ストップワードを除いた形容詞, 名詞, 動詞とする. ストップワードとして, 複数のヘルプファイルにおける頻出語上位100位を利用する.

概念ベースを利用するためには, 未知語の問題が挙げられる. 未知語とは, MeCab 用の辞書及び, 概念

An Automatic Call Software Method by Requirement Analysis
† Atsushi Yano
† Noriyuki Okumura
(noriyuki_okumura@ei.nagano-nct.ac.jp)

ベースに登録されていない語のことである．概念ベースに登録されていない語は"*idf*"を取得することができない．

一方，概念ベースは，新聞や辞書等の一般的な文章に利用される語と，その語が持つ概念を対応させたデータベースである．その概念ベースに登録されていない未知語は一般的な文章には出現し難く，希少性の高い語であるといえる．"*idf*"は，語の希少性を示すものである．そのため，"*idf*"の値を，上限値の 1 に設定する．

2.4 ソフトウェア分類

キーワード抽出で得られたキーワードを用いて，ソフトウェア分類知識ベースを参照し，対象のソフトウェアの種類を分類する．分類手順を図 2 に示す．

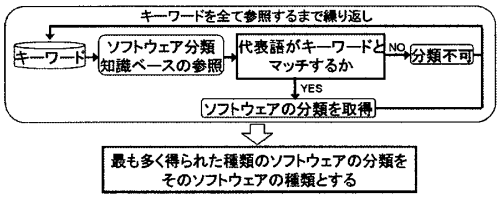

図 2: ソフトウェア分類

3 評価

本章では，ソフトウェア選出，キーワード抽出，知識を追加した場合のソフトウェア選出の評価を行った結果を記す．

3.1 評価：ソフトウェア選出，キーワード抽出

ソフトウェア選出手法の評価を行うため，自然言語文章からのソフトウェア呼び出し評価を行った．使用した文章の内容は，ソフトウェア分類知識ベースに登録されているソフトウェア名を連想可能な文とする．

例：絵を描きたい ⇒ ペイント，… etc

先の条件に合う文章 30 文を作成し，ソフトウェア選出を行った結果を図 3 に示す．(○：正解，×：不正解)

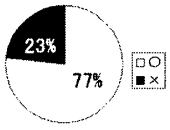

図 3: ソフトウェア選出結果

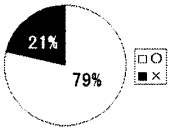

図 4: キーワード抽出結果

キーワード抽出手法の評価を行うため，ウェブ (窓の杜，Vector) 上から，新たにダウンロードしたソフトウェア (47 個) に付属しているヘルプファイル等を用いてキーワード抽出を行った．その結果得られたキーワードについて，ソフトウェアを分類する際，正しく分類可能なキーワードを得られた数が，誤った分類となるキーワードの数より多い場合を"○"とし，逆の場合を"×"とした時の評価結果を図 4 に示す．また，キーワード抽出時に使用したストップワードリストは，上述のヘルプファイル等を用いて作成した.(4562 語)

3.2 評価：知識追加後のソフトウェア選出

現在採用しているソフトウェア分類知識ベースへの知識の追加手法を図 5 に示す．

3.1 のキーワード抽出の結果，得られたキーワードと分類を用いて，ソフトウェア分類知識ベースに新たにダウンロードしたソフトウェアの知識を追加し，3.1 で使用した文章と同じ文章を用い，ソフトウェア選出を行った結果を図 6 に示す．

(○：正解のみ，△：不正解を含む ×：正解無し)

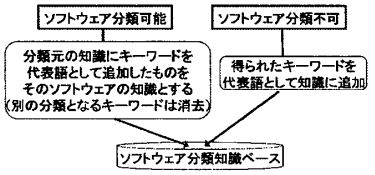

図 5: 知識の追加法

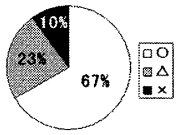

図 6: 知識追加後のソフトウェア選出結果

4 考察

ソフトウェア選出を評価した結果，77% の正解率を得られた．不正解となった文章には，形態素解析により得られた要求名詞の内容がソフトウェア分類知識ベースに存在しないという原因があった．しかし，この結果は，元のシステム [1] に近い精度が得られているため，高い精度が得られたといえる．

キーワード抽出を評価した結果，79% の正解率が得られた．不正解となった評価データには，ソフトウェアの分類を行えないが，未知語のため，"$tf \cdot idf$"の値が高くなる語により，キーワードが設定されるという原因があった．

知識追加後のソフトウェア選出では，67% の正解率が得られた．精度自体は，元のソフトウェア選出の評価と比べて下がっている．しかし，キーワード抽出での分類の結果，分類がされなかったソフトウェアについてもソフトウェア選出を行えた場合があった．

以上の結果より，7~8 割程度の精度を持ったシステムが開発できた．

5 おわりに

本研究では，評価実験により，従来手法と同等の精度を持ち，学習機能を付加したシステムを提案した．

参考文献

[1] 川波正典，大江奈緒子，渡部広一，河岡司．概念連想に基づく職種判断システムの構築－状態文処理機能の拡張－．第 18 回人工知能学会全国大会論文集 2D2-03, 2004 年 6 月．

[2] 工藤祐．Mecab. http://www.mecab.sourceforge.net/.

[3] 渡部広一，奥村紀之，河岡司．概念の意味属性と共起情報を用いた関連度計算方式．自然言語処理，Vol. 13, No. 1, pp. 53–74, 2006 年 1 月．