E-23

品番などを含む技術文書の形態素解析手法

A Morphological Analysis Method for Technical Documents including Product/Part Identifiers

荒木 円博† 尾口 健太郎†
Mitsuhiro Araki Kentaro Oguchi

1. はじめに

われわれは技術文書から，推論などの知識処理に利用可能な知識の抽出をめざしている．

文書から知識を抽出する方法として，あらかじめ抽出したい知識のフレームを定義しておき，フレームのスロットに値を埋めていく方法が考えられる．その場合，前処理として形態素解析と構文解析が必要になる[1].

形態素解析では，辞書にない形態素は未知語として扱われる．しかし，その影響で前後の形態素の品詞を適切に特定できない場合がある．したがって，未知語による悪影響を排除するには，形態素をすべて辞書に登録するのが望ましい．

しかし，技術文書には品番のように無限に存在し得る形態素も多く使われており，これらをすべて辞書に登録することはできない．

そこで，正規表現で形態素を定義，解析できる手法を開発し，品番などを扱えるようにした．

2. 知識抽出システムの構想

2.1 システムの概要

近年，技術文書がコンピュータ上で蓄積・再利用されるようになってきたが，現状の文書形式は人が読むのには適しているものの，コンピュータが解釈するのには適していない．

しかし，技術文書をコンピュータが解釈しやすい形式にできれば，例えば CAD ツールを使って設計を行っている状況で，文書化されたノウハウをもとに適切なガイダンスを与えることが可能になる．

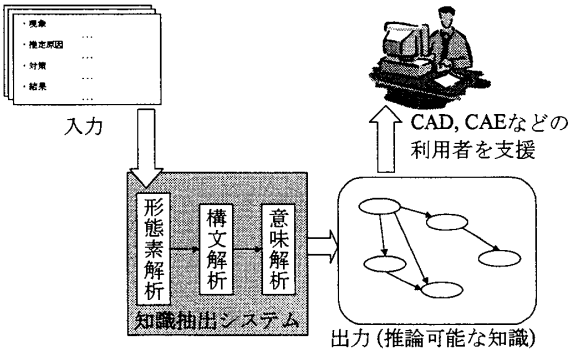

図1 構想中のシステム概要

そこで，われわれは技術文書を入力として，推論可能な知識を抽出するシステムを開発する．システムは形態素解析と構文解析を行い，意味解析としてフレームのスロットに値を埋め込む（図 1）．解析結果の外部記憶での表現形式や推論のための表現形式は Semantic Web[2][3]に基づく予定である．

2.2 問題の特徴

技術文書には品番や型番が多く使われている．

これらの語句は無限に存在し得るため，形態素解析器の辞書に登録することができない．形態素解析器は辞書に登録されていない形態素を未知語あるいは詳細が不明な名詞として扱う．そのため，品番などの前後にある形態素の品詞情報が一意に決定できない場合が生じ，後段の構文解析や意味解析に悪影響を与える．

逆に，品番などに対して詳細な品詞情報を与えることができれば，前後に出現する形態素の品詞情報を一意かつ詳細なものにしやすい．すなわち，品番などを適切に扱うことによって，構文解析や意味解析の精度向上が期待できる．

3. 品番などを含む技術文書の形態素解析手法

3.1 考え方

形態素解析器が利用する辞書は，基本的には形態素の品詞情報と字面文字列の対から成り立っている．一方，品番や型番は字面のパターンが正規表現で定義できる．したがって，文字列の代わりに正規表現を使って辞書のエントリを定義できるようにし，それに基づいて形態素解析を行うようにすれば，目的を達せられる．

そこで，既存の形態素解析器に前処理と後処理を付加することによって，正規表現対応の形態素解析器（以後，拡張形態素解析器と呼ぶ）を構成する．

3.2 拡張形態素解析器の構成

拡張形態素解析器は，既存の形態素解析器（形態素解析器本体）とその辞書（本体辞書），正規表現で字面を指定した形態素（特定語）を定義する特定語辞書（表 1），特定語を形態素解析器本体で処理するのに都合の良い文字列（代替語）に置換する特定語置換器，代替語を特定語に戻す特定語復元器，および実際に出現した特定語と代替語の対応表（特定語対応表：表 2）から構成される（図 2）．

特定語辞書は特定語の正規表現と代替語の対を列挙したものである．特定語の品詞情報は本体辞書で代替語に対する品詞情報として与える（表 3）． 特定語対応表は代替語をキーとし，特定語の出現順による列を値とするハッシュ表である．

† (株) 豊田中央研究所, Toyota Central R&D Labs., Inc.

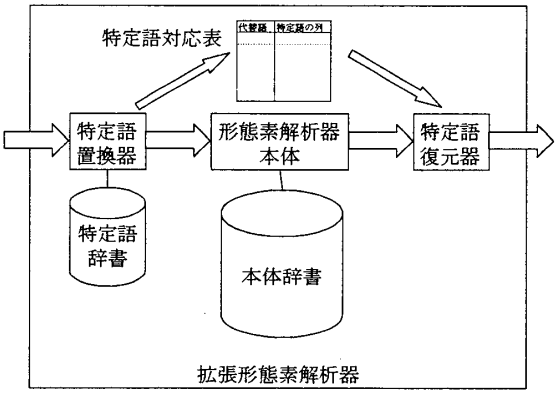

図2 拡張形態素解析器の構成

表1 特定語辞書の例

正規表現	代替語
2S[BD][0-9]+	代替語－低周波Ｔｒ品番
2S[AC][0-9]+	代替語－高周波Ｔｒ品番

表2 特定語対応表の例

代替語	特定語の列
代替語－低周波Ｔｒ品番	[2SD460]
代替語－高周波Ｔｒ品番	[2SC11 2SC382]

表3 本体辞書での特定語品詞情報の例

品詞情報	代替語
名詞－固有名詞－品番－トランジスター低周波	代替語－低周波Ｔｒ品番
名詞－固有名詞－品番－トランジスター高周波	代替語－高周波Ｔｒ品番

3.3 拡張形態素解析器の動作

拡張形態素解析器は以下のように動作する．

拡張形態素解析器への入力文字列は，まず特定語置換器によって処理される．特定語置換器は特定語辞書の最初のエントリを使って入力文字列の中で正規表現に照合した部分を，代替語に置換するとともに，特定語対応表に格納する．例えば，入力文字列に "2SD460" が含まれており，表1の特定語辞書が定義されている場合は，正規表現 "2S[BD][0-9]+" に照合するので，"2SD460" が「代替語－低周波Ｔｒ品番」に置換され，表2のように「代替語－低周波Ｔｒ品番」をキーとして "2SD460" を先頭要素とする値が特定語対応表に登録される．

この処理を照合する部分がなくなるまで繰り返し，さらに特定語辞書の残りのエントリについても同様に処理する．

特定語置換器で一通り置換の終わった入力文字列は形態素解析器本体によって通常の形態素解析が行われる．この時，代替語も表3のような本体辞書の定義に基づいて形態素に分解され品詞情報が付与される．

形態素解析器本体から出力される各形態素の情報は1つづつ特定語復元器に送られる．特定語復元器は形態素の字面で特定語対応表を検索する．その結果，特定語の列が見つかれば形態素の字面を列の先頭の項目で置き換え特定語を復元する．この時，列の先頭の項目を取り去り，列の2番目の項目を新たな先頭とする．この繰り返しによって代替語がすべて特定語に復元される．

3.4 拡張形態素解析器の実装

Windows 2000 上に Java 言語 (J2SE 1.4.0 SDK) を使って拡張形態素解析器を実装した．

形態素解析器本体には茶筅[4] (cha21244.exe) を，本体辞書には茶筅用の ipadic-sjis-2.4.4 を利用した．茶筅とのやりとりは茶筅を別プロセスとして起動し，標準入出力を介して実現した．

特定語辞書の正規表現の処理には J2SE 1.4 から使えるようになった java.util.regex パッケージを利用した．特定語辞書そのものは XML ファイルとして表現した．

4. まとめ

技術文書には品番や型番など無限に存在し得るが字面のパターンが正規表現で定義できる単語が多く含まれている．こうした単語を適切に形態素解析できる拡張形態素解析器を設計，実装した．

今後は拡張形態素解析器を利用して，文書からの知識抽出システムの実現を進めていく．

参考文献

[1] 新山祐介，徳永健伸，田中穂積：自然言語を理解するソフトウエアロボット：傀儡，情報処理学会論文誌，Vol.42, No.6, pp.1359—1367 (2001).

[2] http://www.w3.org/2001/sw/

[3] 萩野達也ら：特集 セマンティック Web，情報処理，Vol.43, No.7, pp.707—750 (2002).

[4] 松本裕治ら：http://chasen.aist-nara.ac.jp/