自然言語事例ベースの仕様書文解析の評価

2Ｌ－１０
高山 泰博　伊藤 山彦　鈴木 克志
三菱電機（株）情報技術総合研究所

１．はじめに

自然言語処理の分野では技術の新しい適用対象の開拓が求められている[1]。我々は、事例ベース推論が自然言語理解の研究を端緒として生まれた点に着目して、自然言語表現の事例の検索を仕様書文の解析に応用したシステムを提案した[2]。従来の自然言語処理の応用は主に事務処理的な分野が対象であったのに対し、このシステムは受注生産型の機械製品であるエレベータの設計支援に用いる。電子化した仕様を基にした設計作業環境で、仕様中の言語データと設計結果の対を蓄積し、事例ベース構築の問題の解決を試みている。本稿では試作システムの概要、初期事例ベース構築時の実験結果を述べる。

２．エレベータ設計業務の概略と仕様書の解析

受注設計の業務は《製品仕様》を作成する【概略設計】と、部品表への展開や個別設計の【詳細設計】からなる。概略設計で電子化した製品仕様はネットワーク経由で工場へ伝送する（図１）。

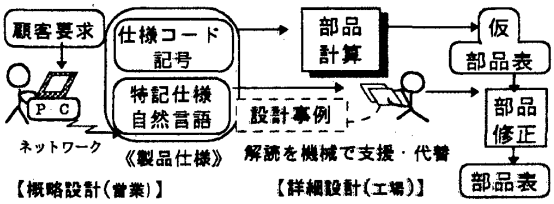

図１　受注生産型製品の設計業務の概要

製品仕様には記号で記述する標準化された部分と言葉で記述する例外的な「特記仕様」の部分がある。自然言語で自由に表記した特記仕様は部品表への展開が計算できない。そこで、過去の設計事例を参照することによって、特記仕様の文を計算機で解析して各特記文に関連する部品を推論し、設計作業を支援する。

An Evaluation of Natural Language Case-based Retrieval for the Specification Analysis
Yasuhiro TAKAYAMA, Takahiro ITO, Katsushi SUZUKI
Mitsubishi Electric Corporation.
5-1-1 Ofuna, Kamakura, Kanagawa 247, JAPAN

３．自然言語事例ベース検索による仕様書解読

仕様解読処理は、次の手順で言語事例ベース中の設計事例から類似仕様を取り出し、事例に関連付けた部品群から設計に必要な情報を推論する（図２）。

（１）ＢＮＦ(Backus Naur Form)定義による記号列の抽出
（２）専門語辞書による用語抽出と異表記吸収
　【例】乗場押釦（異表記）→ 乗場ボタン（正表記）
（３）用語索引による事例ベースからの事例候補の検索
（４）事例文との入力文の照合による優先度付け
（５）入力文の仕様に対応する関連部品の推論

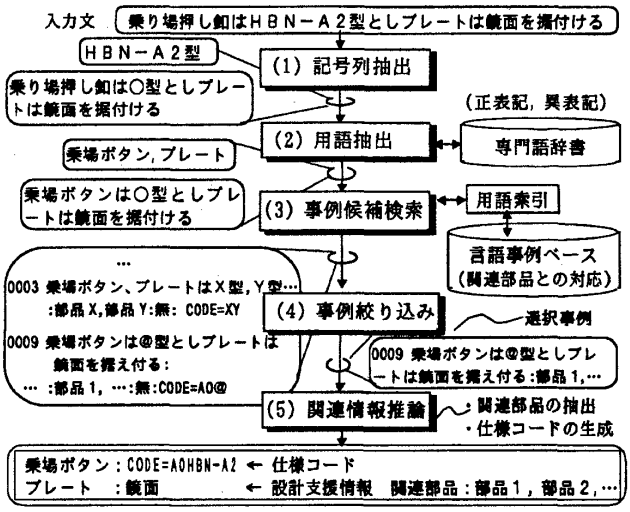

図２　仕様解読処理の構成

４．初期事例ベースの構築実験と分析

図２の構成で仕様書解読を行なう場合、初期事例ベースをいかに構築するかが一つの課題である。事例ベース構築には図２の(1)記号列抽出と(2)用語抽出のモジュールを用い、図３の処理を行なう。

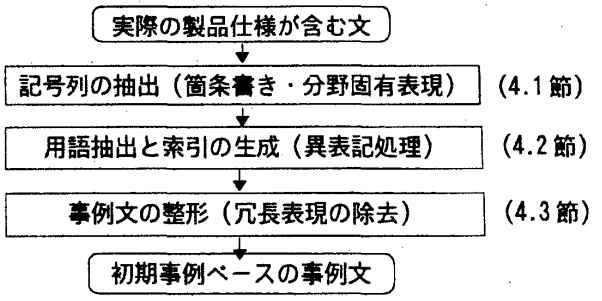

図３　初期事例ベースの構築処理の流れ

以下、初期事例ベース構築の実験について述べる。

4.1 記号列抽出による初期事例文の抽出

今回実験の対象とした準標準機種では、1件の仕様に対して5～6行程度の特記仕様が含まれている場合が多い（仕様内容でばらつき有り）。ある期間中に実際の設計に用いた約3400行（1行は最大40文字、複数行に渡る文も含まれる）の特記仕様から実験用の初期事例文を作成した。まず、3400行から単純な図面参照だけの文等を除いて、1599文を抜き出し、更に表1の記号列を抽出した。

表1 記号列抽出（1599文）の結果

箇条書きを表わす行頭の約物	667文
分野(エレベータ)固有の記号表現 (階床表現:例「1～6階」,「3FL」)	668文

表1の記号列を除去したうえで、sort処理等を行ない、初期事例文として、840文を選定した。

4.2 事例索引の生成と専門語辞書の表記の揺れ

次に、840の事例文に対し、約5500見出しの専門語辞書を用い、用語抽出（索引生成）を行なった。抽出総数は2796語（1つの文が同じ用語を2回以上含む場合、1索引とすると2587索引）であり、1つの事例文当り平均約3語が索引となった。辞書の登録見出しはほぼ適正と考えられる。

しかし、抽出した用語と事例文を分析すると、索引として抽出すべき以下の語があることが判った。

(a) 表記の揺れが組み合わさった専門語

【例】「かご呼一括取消ボタン」
[ひらかな・カタカナ]： かご,カゴ
[カタカナ・漢字] ： ボタン, 釦
[送り仮名]： 呼,呼び; 取消:取消し,取り消し

(b) 単純な書き換え処理が出来ない一般語

【例】「扉」→乗場の戸(乗場扉),かご室の戸,機械室扉

省略表記した語が複数の意味に対応する場合は単純に表記の揺れとして、あらかじめ処理できない。

上記の分析を基に、異表記を中心に専門語辞書に180見出しを追加して、再度索引の生成を試みた。

表2 索引生成（840文）の結果

試行	辞書見出し	抽出総数	索引数	異表記数
1回目	5502	2796	2587	155
2回目	5682	3255	3019	420

表2が示すように、2回目の索引生成では、異表記処理が働いた数が抽出総数の約13％にも及ぶ。

異表記処理は入力者が不特定多数である場合に特に重要となる。これは、一般にネットワーク上で流通する文書中の言語データの処理においても同様であると考える。

4.3 類似度計算のための事例文の整形

仕様解読の処理（図2）では、最長共通部分列(Longest Common Subsequence:LCS)[3]の長さの、文全体の長さに対する割合で類似度を定義し、事例の候補を順位付けする[2]。この方法は、用語辞書で吸収しなかった、送り仮名やカタカナ等の表記の揺れに対処できる。しかし、文の長さが類似度に影響を与えるため、類似度計算を行なう前に、以下のような冗長な表現を除去しておく必要がある。

【例】（冗長な文末の表現）
～手配とする。 ～手配ください。 ～手配願います。
～手配乞う。 ～手配下さい。 ～手配願う。

4.1節で用いた記号列抽出処理はBNFで定義したパタンを抽出する。そこで、上記の下線部のような文末表現をあらかじめ定義しておき、記号列抽出モジュールを事例文の整形処理にも用いる。この処理は、解読処理時に入力文に対しても行なう。

5．まとめ

自然言語事例ベース検索方式を実文章（エレベータの仕様書文）の解析に用いる場合の初期事例ベース構築における言語処理の実験と考察を述べた。今後、以下を実施し、設計作業支援に役立てていく。

- 解読処理(検索能力と関連部品推論)の定量評価
- 蓄積事例の仕様入力へのフィードバック
- 特注機種における複雑な特記仕様の処理

謝辞

本研究に協力いただく、当社稲沢製作所の市岡洋一,堀場一夫,平田政信ほかの各氏に感謝します。

参考文献

(1)野村ほか:自然言語処理研究の動向と問題点，情報処理学会研究会報告，NL100-6，Vol.94，No.28，pp.41-48(1994).
(2)高山，鈴木:自然言語事例ベースの仕様書文解析の応用，第52回情報処理学会全国大会 2B-4，分冊3,pp.25-26(1996).
(3)Cormen,T.H.,et al.:"Introduction to Algorithms", pp.314-320, MIT Press(1990).