開発プロセス上の文書成果物の分析における自然言語処理の適用

日本アイ・ビー・エム株式会社 東京基礎研究所
荻野 紫穂，竹内 広宜，中田 武男
{shiho, hironori, nakada}@jp.ibm.com

1. 背景

近年、オフショア開発の増加などによって、システム開発の過程で発生する仕様書や設計書などの文書成果物が、矛盾なく、かつ、誤解を与えないように書かれているかどうかを吟味する重要さに、注目が集まっている[12)]。

要求工学の分野では、体系化された知識源であるオントロジーや制限言語を用いて書き手が同質の文書を作成することにより、仕様書の質の向上を目指す取り組みが、多く見受けられる[2)3)8)9)]。例えば3)では、対象ドメインで使用する動詞を限定し、その動詞句の内容を詳細に分割する形でオントロジーを作成する。仕様書作成の際にそのオントロジーを利用することにより、機能に関する記述の見通しのよさを図っている。こうした取り組みは、主に、新規に単体の文書を作成する際の使用を想定して開発されており、すでに作成してしまった過去の文書の再利用や、開発のために作成した文書群全体の一貫性の保持などの目的には、あまり向いていない。また、オントロジーや制限言語のリソース開発は、それ自体にコストがかかるという問題もある。

本研究では、要求工学で使われる分析観点に着目し、文書成果物中の問題点を自然言語処理技術を用いて検出することを試みる。本稿では、文書成果物に出現する各文の構文解析結果を、拡張格フレーム型文要約の表形式で出力する手法について述べる。また、実際の開発関連文書に提案手法を適用し、その有用性を報告する。

2. 仕様書が満たすべき項目と分析対象

2.1. 仕様書が満たすべき項目

仕様書などの文書成果物の分析において、自然言語処理の観点を取り入れること自体は、既に検討されている。7)は、仕様書に見られるあいまいさについて、様々な観点から観察しており、自然言語処理の伝統的なフェーズである「語彙」「構文」「意味」「運用」ごとに、それぞれ違ったあいまいさがあることにも触れている。自然言語処理の伝統的なフェーズごとに分類されたあいまいさは、文書群中のあいまいなポイントを検出する手がかりにはなるが、それだけでは、文書成果物の問題点の位置付けにはなり得ない。

多くの要求工学の教科書には、伝統的な自然言語処理の分類とは別な観点で、仕様書が満たすべき項目が述べられている[6)]。また、その逆に、実際の仕様書によく見られる問題点として、「あいまいさ」や「理解のしづらさ」を挙げる研究もある[7)]。

最近では、実用的な観点から、仕様書が満たすべき項目を具体的に述べている文献も見られる。12)では、オフショア開発で、日本語が母国語ではないメンバーと共に開発する際に共有される仕様書の書き方が、実例を挙げて紹介されている。また、5)のように、あいまいさを排除した要求仕様書の実例を作成しているグループもある。

先行研究で述べられている、仕様書が満たすべき条件は、研究の観点によって、少しずつ異なってはいるが、概念的なレベルでは、どれも、「あいまいでないこと」「矛盾がないこと」など、要求工学で定義されている項目と重なっている部分が多い。このため、本稿では、要求工学の教科書である 6)に挙げられている項目を採用し、我々が開発した手法によって、どの項目のどの部分を吟味するためのサポートができるかを考える。

2.2. 今回の分析項目

6)には、仕様書が満たすべき項目として、下記が挙げられている。これらの項目は IEEE Std 830-1998 として標準化されている。

- 妥当性（correctness）
- 非あいまい性（unambiguity）
- 完全性（completeness）
- 無矛盾性（一貫性）（consistency）
- 重要度と安定性のランク付け（ranked for importance and/or stability）
- 検証可能性（verifiability）
- 変更可能性（modifiability）
- 追跡可能性（traceability）

自然言語処理技術が、これらの項目全ての吟味に向いているわけではない。例えば、妥当性は、そもそも計算機での吟味自体が困難であることが 6)で述べられている。重要度と安定性のランク付けについても、市場価値などの外部要因を取り入れず、文書成果物のテキストだけを基に吟味するのは困難である。

我々は、上記の、仕様書が満たすべき項目のうち、単純な表現登録や一文ごとの単体文走査だけでは吟味が困難な完全性・無矛盾性・非あいまい性について、拡張格フレーム型文要約の手法を適用し、文書成果物の問題点を観察した。

3. 拡張格フレーム型文要約

3.1. 概要

本研究の拡張格フレーム型文要約では、ガ・ヲ・ニなどの格要素（英語の場合は subject や object や prepositional phrase など）や述語など、構文要素の種類

ごとの出力カラムを定義した。その上で、仕様書群中のテキストを構文解析し、その結果に基づいて、述語の個々の出現に対して、その箇所でその述語に係っている構文要素と、述語自体とを、各カラムに分類して拡張格フレーム型文要約として出力した。カラムにこだわらずに同じ語の出現箇所を揃えて見たい場合も多いため、全カラムの全名詞を出力する欄も設けた。また、仕様を確認する際には、テキストから抽出した情報以外、例えば図表部分や文書構造などを参考にする場合も多い。このため、文要約には、元の文書の記述に戻るためのリンクと、Keyword In Context (KWIC)[1] 形式の前後の文脈を添付した。

この形式は、「『印刷する』のガ格に当たる語のバリエーションを見たい」「あるモジュールが関係している動詞を全て見たい」など、文書群を吟味する上での観点を次々に変えたいという要求に対応することができる。また、着目語の周辺情報も、述語を中心とした簡潔な形で参照することができる。

拡張格フレーム型文要約の作成処理においては、日本語の構文解析には IBM Content Analyzer (ICA)[14]の日本語処理部を使用し、英語の構文解析には IBM T. J. Watson 研究所で開発されている XSG[10] を使用した。

3.2. 従来の技術との比較

文書成果物の問題箇所の発見は、情報抽出という観点において、Google[13]を始めとする検索エンジンや、ICA などのテキストマイニングシステムと、類似した性質を持つ。

検索エンジンの出力は、検索語を強調表示し、検索語の周囲の文脈を添付する形式で、KWIC を変形した出力とも言える。文書群中で、自分が見たい箇所の候補を大まかに洗い出すには非常に便利だが、ヒットした情報は、人手で整理し直す必要がある。

ICA を始めとするテキストマイニングの出力では、主語と述語、形容詞と被修飾名詞など、2-3 語間の簡潔な関係を、文書群から抽出して示すことが多い。こうした出力は、コールセンターでのエラー報告や、アンケートでの好悪記述など、内容の傾向把握に有効だが、仕様書を読む際には、周辺記述を次々に追いかけることも多く、簡潔な事実関係の記述だけでは足りないことも多い。

従来の拡張格フレーム型の要約は、質疑応答システムや検索システムへの応用を目的にして、単独の一文から情報を抽出するのではなく、短いテキスト、例えば、ニュース記事や文書のある一部分から単語を抜き出して、拡張格フレームのスロットに埋めるものが多い [4]。この方式は、文書全体の内容が正しい場合に、文書の要約を作成するには非常に向いているが、文書のあいまい性を吟味するには向いていない。本研究で用いた手法では、文書全体ではなく単体の述語ごとの拡張格フレーム型要約を作成し、その一覧を表形式で出力するため、文ごとの結果を見比べられるようになっている。

3.3. 機械翻訳資源を用いた日英共通フレームへの変換

日本語の構文解析と英語の構文解析とを、そのまま格要素ごとに出力するだけでは、日本語・英語双方の仕様書が混在する開発において、全てを同時に参照するための見通しがつけられない。この点を考慮して、我々は、日本語の構文要素を、英語の構文要素である subject や object や前置詞句に変換する手順を拡張格フレーム型文要約の作成処理に組み入れた。この変換には、機械翻訳の対訳辞書リソースを活用する。

仕様書全体を対象言語に翻訳するのは、非常にコストが高いが、日英共通の出力フレームを作成しておけば、専門用語の日英対応辞書がある場合には、出現単語だけをどちらかの言語に変換することによって、あるトピックが関連する部分を、英語・日本語を横断して参照することができる。仕様書が英語で書かれている場合は、XSG の構文解析結果をそのまま使用する。

4. 実際の仕様書に対する適用

拡張格フレーム型文要約を実データに適用した出力を、図 1～図 3 に示す。 この結果を用いて、文書成果物内の問題点を検出する実例を、以下に紹介する。

4.1. 完全性違反:格要素欠落の発見

文書成果物中の格要素の欠落は、完全性違反の一部と見なすことができる。格要素の欠落を発見する従来手法には、述語に対する必須格を辞書で定義し、文章中で述語に対して必須格が現れなければ、欠落と見なすものが多い。このためには、使用される全ての述語の必須格を定義しなければならない。また、必須格を過剰に定義すると、格欠落として抽出される例が多すぎて、吟味作業を妨げやすい。拡張格フレーム型文要約を使用する場合、文書群の中の、他の出現箇所と比べながら、格の欠落を吟味できるため、辞書定義の手間がなく、過抽出を避けることができる。

図 1 は、拡張格フレーム型文要約の出力を述語でソートして、「通知する」と「選択する」の部分を示したものである。「通知する」は、大抵の場合、英語で to に当たる「X に」が書かれている。しかし、「アドレスが…通知される」の例では、「X に」が欠落しているため、完全性違反である可能性が高い。一方で、「選択する」は、「X で」が書かれている例と書かれていない例の出現数がほぼ同じである。ここから考えると、「選択する」にデ格が書かれていなくても、格欠落ではない可能性が高い。

今回、設計文書群の調査では、文のトピックを示す「X は」「X が」に当たる subject が、50%の文において欠落していることも解った。日本語では、主語の省略は文法違反ではない。しかし、仕様書などの文書成果物は、日本語が堪能でない開発者や、専門分野の知識が豊富ではない翻訳者によって、書き手の意図とは違う主語が補われる可能性も高いため、注意が必要である。

4.2. 無矛盾性違反:表記揺れの発見

無矛盾性違反の一部である表記揺れを発見する手法の一つに、狭い範囲で同じ単語と共起している語の

セットを、候補として集める手法がある [11]。例えば、テキストセットの中で、ある動詞にガ格で係っている名詞を複数集めて、初期候補とするやり方である。この手法では、共起している語同士だけではなく、その周囲の情報を次々と追いかけながら、候補を絞り込んでいくことが多い。拡張格フレーム型文要約を利用すると、述語を中心とした周辺情報を簡潔な形で一覧でき、更に周辺情報でソートを繰り返すことが容易なため、こうした表記揺れの発見に有効である。

図 2 は、英語の仕様書の解析結果を、述語でソートしたものである。上表では、common layer の前後で受け渡しされるものが、文書群の中で、compressed raster band と compressed band raster の2種類の表記の間で揺れているのが解る。上表で、common layer に compressed raster band を受け渡しているらしい input source に関する部分を下表でみると、input source layer が作るものは band raster image と書かれている。これら3つの compressed raster band, compressed band raster, band raster image という表記は、同じ概念の表記の揺れである可能性が非常に高い。

表記揺れの発見は、用語統一辞書の作成において、非常に重要である。我々は、現在、用語辞書作成における拡張格フレーム型文要約の活用により、無矛盾性と同時に追跡可能性を向上させることを考慮している。

4.3. 非あいまい性違反:あいまいな記述の発見

文書の内容に関する非あいまい性違反や無矛盾性の発見は、オントロジーを利用して、そのオントロジーに違反した記述がないか、一文の範囲で吟味する従来手法が多い [8]。拡張格フレーム型文要約を利用すると、対象となる文書群全体から、同じトピックに関する記述を抽出して吟味することで、オントロジーに依存せずに非あいまい性違反の吟味を行うことができる。

図 3 は、日本語文書の解析結果を出現名詞でソートした「フォント番号」の部分である。ここには、「フォント番号は XYZ フォントと同じ」という記述と「フォント番号は(何かの)番号を使用する」という二つの記述が文書の異なる部分から抽出されている。後者の出現の詳細を追うと、「(何かの)番号」とは「フォントディスクリプタ中の番号」であることが解る。これらの記述を合わせると、「XYZ フォントの番号がフォントディスクリプタ中に保存される」など、XYZ フォントとフォントディスクリプタとの関係が、文書群中のどこかに記述されているべきだが、そうした記述は発見できなかった。この結果から、XYZ フォントとフォントディスクリプタとの関連性が文書成果物に記述されず、あいまいなまま残されていることが解る。こうしたあいまいさは、読み手に混乱や誤解を与えやすいため、発見と修正のサポートが重要である。

5. まとめ

本稿では、拡張格フレーム型の文要約を用いて、仕様書の問題点の吟味が可能であることを示した。

2.2 で述べた項目のうち、今回、拡張格フレーム型文要約を用いた分析対象としなかったものについても、継続して、吟味手法を探っていく。例えば、追跡可能性の一部分は、用語統一や、参照先を示す表現の周囲の抽出により、ある程度対処できる見通しを立てている。用語統一には、4.2 で述べた拡張格フレーム型文要約の適用も有効と考えられる。参照先を示す表現の抽出は、変更可能性の吟味にも使用可能である。今後も、こうした技術を積み重ねて、文書成果物の質の向上の統合的なサポートを開発していく予定である。

参考文献

1) 長尾 真: 文字・単語列の処理, 言語工学 pp. 42-74, 昭晃堂 (1988)

2) 大西 淳: 要求フレームに基づいたソフトウェア要求仕様化技法, 情報処理学会論文誌 Vol. 31, No. 2 pp. 175-181 (1990)

3) 吉岡 真治他, 入出力を中心とした機能表現と動詞を中心とした機能表現の比較分析, 第 8 回設計工学・システム部門講演論文集,pp. 145-148 日本機械学会 (1998)

4) 池田崇博, 佐藤研治, 奥村明俊: 5W1H 情報の在否により結果を分類する情報検索システム, 第 59 回平成 11 年後期情報処理学会全国大会講演論文集 No.3 pp. 103-104 (1999)

5) http://www.sessame.jp/workinggroup/WorkingGroup2/POT_Specification.htm

6) 大西 淳, 郷 健太郎: 要求仕様の特性, 要求工学ソフトウェアテクノロジーシリーズ 9, pp. 102-108, 共立出版 (2002)

7) Berry, D. M, and Kamsties, E: Ambiguity in Requirements Specification, Perspectives on Software Requirements, pp. 7-44, Kluwer Academic Publishers, Netherlands (2004)

8) 海谷 治彦, 佐伯 元司: 要求分析におけるオントロジーの活用法, 電子情報通信学会技術研究報告, ソフトウェアサイエンス(SS) Vol.105, No.25, pp. 7-12 (2005)

9) 山梨 敦志, 松浦 佐江子: 自然言語処理を利用したユースケース記述推敲支援, 第 68 回情報処理学会全国大会, 6A-3 (2006)

10) McCord, M. C.: Using Slot Grammar, IBM Research Report RC23978 (W0607-019) (2006)

11) 那須川 哲哉: テキストマイニングを使う技術/作る技術―基礎技術と適用事例から導く本質と活用法, 東京電機大学出版局 (2006)

12) 幸地司: 外国人プログラマ向け設計書の留意点, EngineerMind, Vol. 9, pp. 62-71 (2008)

13) http://www.google.co.jp

14) http://w3-06.ibm.com/jp/domino60/mkt/DB202.NSF/doc/00292120

原文(KWIC)				ターゲットの係り受け				動詞句にかかる名詞句、動詞句						
本文左側	名詞句		本文右側	名詞句		動詞句								
	自立語部分	付属語部		自立語部分	付属語部分	動詞句	動詞	topic	obj	v	to	from	by	others
・これをカレントフォントとして	フォントマネージャー	に	通知することによりフォントを選択することが出来る	フォントマネージャー	に	通知することにより	通知する	---	これを	して	フォントマネージャーに	---	---	---
・フォントの指定はフォントの選択によって得られたフォントデスクリプターを	フォントマネージャー	に	通知することによって行う。	フォントマネージャー	に	通知することによって	通知する	---	フォントデスクリプターを	---	フォントマネージャーに	---	---	---
・フォントの指定はフォントの選択によって得られた	フォントデスクリプター	を	フォントマネージャーに通知することによって行う。	フォントデスクリプター	を	通知することによって	通知する	---	フォントデスクリプターを	---	フォントマネージャーに	---	---	---
・フォントの指定はフォントの選択によって得られた	フォントデスクリプター	を	フォントマネージャーに通知することによって行う。	フォントデスクリプター	を	通知することによって	通知する	---	フォントデスクリプターを	---	フォントマネージャーに	---	---	---
・フォント初期化関数にて、ROM Slot Formatの	アドレス	が	通知される。	アドレス	が	通知される。	通知する-れる	アドレスが	---	---	---	---	---	フォント初期化関数にて、
・どのフオントデバイスを使用して文字を描画するか	フオントスケールマネージャ	に	通知されるカレントフオント中のフォントフォーマットで決まります。	フオントスケールマネージャ	に	通知される	通知する-れる	---	---	---	フオントスケールマネージャに	---	---	---

原文(KWIC)				ターゲットの係り受け				動詞句にかかる名詞句、動詞句						
本文左側	名詞句		本文右側	名詞句		動詞句								
	自立語部分	付属語部		自立語部分	付属語部分	動詞句	動詞	topic	obj	v	to	from	by	others
・これをカレントフォントとしてフォントマネージャーに通知することにより	フォント	を	選択することが出来る	フォント	を	選択することが	選択する	---	フォントを	通知することにより	---	---	---	---
・XXで	フォント	を	選択するときはこの順に優先順位が下がっていく. 。	フォント	を	選択する	選択する	---	フォントを	---	---	---	XXで	---
・YY resourceにあるXXフォント名から該当する	フォント	を	選択する処理になります。	フォント	を	選択する	選択する	---	フォントを	---	---	---	---	---
・	XX	で	フォントを選択するときはこの順に優先順位が下がっていく. 。	HP	で	選択する	選択する	---	フォントを	---	---	---	XXで	---

図 1 拡張格フレーム型文要約出力: 完全性違反の発見

tree_id	subj	verb	obj	to	in	on	at	into	from	with	through
9	common layer	receive	compressed raster band						input source		
20	output layer	receive	compressed band raster						common layer		
81		receive	data						scanner		

tree_id	subj	verb	obj	to	in	on	at	into	from	with	through
5	input source	produce	band raster image								

図 2 拡張格フレーム型文要約出力: 無矛盾性違反の発見

	原文(KWIC)			ターゲットの係り受け			
		名詞句		名詞句			
元文書	本文左側	自立語部分 付属語部分	本文右側	自立語部分 付属語部分	動詞句	動詞句にかかる名詞句、副詞句	
基本設計書	・	フォント番号 が	同じでもフォントクラスが異なれば別フォントになる。	フォント番号 が	同じでも		
基本設計書	・ビットマップのフォントリンクは	フォント番号 で、	アウトラインのフォントリンクはフォント名でそれぞれソートされリンクされている。	フォント番号 で、	リンクされている。	フォントリンクは	ソートされ
基本設計書	・	フォント番号 は	元になるXYZフォントと同じとする。	フォント番号 は	同じとする。	XYZフォントと	
基本設計書	・〔5〕	フォント番号 は	フォントデスクリプター中の番号を使用す	フォント番号 は	使用する。	・〔5 〕	番号を

図 3 拡張格フレーム型文要約出力: 非あいまい性違反の発見