3B4-4

日本語 DODDLE に基づく ロケット運用オントロジーの構築と検索システムへの適用

Constructing a Rocket Operation Ontology and it's Application of an Information Retrieval System based on Japanese DODDLE

小野 穣[*1] Yutaka Ono　洪 潤基[*2] Yunki Hong　森田 武史[*2] Takeshi Morita　川村 正則[*3] Masanori Kawamura　小出 誠二[*3] Seiji Koide　山口 高平[*1] Takahira Yamaguchi

[*1]慶應義塾大学 理工学部 Faculty of Science and Technology, Keio University

[*2]慶應義塾大学 大学院 理工学研究科 Graduate School of Science and Technology, Keio University

[*3]株式会社ギャラクシーエクスプレス技術部 R&D Division, Galaxy Express Corporation

In this paper, we propose an information retrieval system using domain ontologies and a domain ontology construction tool called Japanese DODDLE. In order to obatin appropriate retrieval results, many research of semantic search have been done. The use of ontologies is one of the means to realize the seamantic search. Although retrieval results can be improved with ontologies, it usually takes many costs for users to construct domain ontologies. This paper discusses how to integrate search result refinement and domain ontologies refinement, presenting a methodology to construct domain ontologies incrementally with less costs, using Japanese DODDLE. In our case studies, we developed a rocket operation ontology using Japanese DODDLE and loaded the ontology to an information retrieval system. Through the case studies, we show that the methodology can be promising.

1. はじめに

本研究では，オントロジー搭載型検索システムおよび日本語オントロジーを半自動で構築可能なツールである日本語 DODDLE の提案を行う．より精度の高い検索を実現するために，文書の意味を考慮した意味検索に関する研究が多数行われている．意味検索を実現する方法の一つとしてオントロジーの利用が考えられるが，オントロジーは構築コストが高いという問題がある．本研究では，最初に検索システムに搭載するオントロジー（初期オントロジー）の構築コストを削減するために，日本語 DODDLE を用いる．また，領域オントロジーを検索システムに搭載し，絞込検索および拡大検索により検索精度を向上させる．ケーススタディとして，日本語 DODDLE を用いたロケット運用オントロジーの構築について述べる．また，ロケット運用オントロジーを用いた文書検索実験及びその評価について述べる．

2. オントロジー搭載型検索システム

図 1 に，オントロジー搭載型検索システムの全体図を示す．専門文書検索に用いる領域オントロジーを手動で構築する作業は困難であるため，本研究では初期領域オントロジーを日本語 DODDLE を用いて半自動構築し，検索システムに搭載する．検索システムは，(株) ギャラクシーエクスプレス社 (以下 GX 社) で開発された GXFinder を用いる．GXFinder には，検索された文書がユーザが求めていた文書かどうかをログとして保存する機能がある．ユーザが検索作業を繰り返しながら，ログを領域オントロジーに反映させていくことによって，専門文書検索に特化した領域オントロジーを構築できると共に，検索精度の向上も期待できる．本研究の特徴は，検索システムの検索精度の向上と領域オントロジー構築の問題を，領域オントロジー構築ライフサイクルと文書検索を関連付けて解決することである．

連絡先: 森田武史，山口高平，慶應義塾大学 大学院 理工学研究科 〒223-8522 神奈川県横浜市港北区日吉 3-14-1，Tel: 045-566-1614，E-mail: {t_morita, yamaguti}@ae.keio.ac.jp

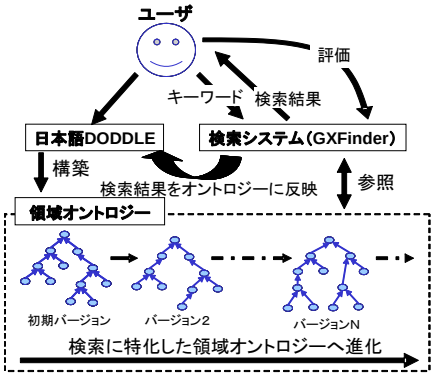

図 1: オントロジー搭載型検索システム

3. 日本語 DODDLE

日本語 DODDLE は，EDR 電子化辞書 [日本 01] と対象領域に関する日本語専門文書を用いて，日本語を概念の表記としてもつ領域オントロジーを半自動で構築可能なツールである．日本語 DODDLE は，DODDLE-OWL [Morita 04] のアーキテクチャを基礎としている．図 2 に日本語 DODDLE のシステムフローを示す．日本語 DODDLE は，入力モジュール，オントロジー構築モジュール，オントロジー洗練モジュール，視覚化モジュール，変換モジュールの５つのモジュールから構成される．はじめに，ユーザは入力モジュールにおいて，入力概念を選択する．オントロジー構築モジュールは，オントロジーの基礎となる初期概念階層と概念対のセットを，EDR 電子化辞書と日本語専門文書を参照しながら，入力概念を基に生成する．初期概念階層は IS-A 階層として構築される．概念対の

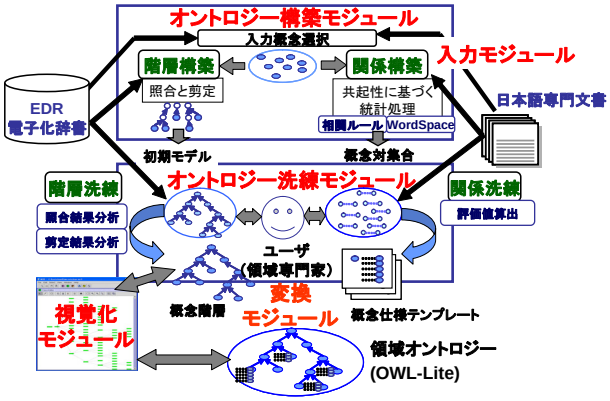

図 2: 日本語 DODDLE のシステムフロー

セットは共起性に基づく統計処理を用いて獲得される．これらの概念対の中から特に重要な概念対を選択し，その間の関係を定義したものが非階層関係（概念定義）となる．オントロジー洗練モジュールでは，オントロジー構築モジュールで構築された初期オントロジーをユーザとやりとりしながら洗練していく．初期オントロジーを洗練するために，概念変動と概念対のセットの評価の管理を支援する．初期概念階層は一般的なオントロジーから生成されるため，ユーザは概念変動と呼ばれる問題を考慮しながら，初期概念階層を特定の領域に調整する必要がある．それは，特定の部分の概念が領域によって変化することを意味する．概念変動管理のために，日本語 DODDLE は照合結果分析と剪定結果分析の２つの戦略を適用する．オントロジー構築モジュールで生成された概念対のセットから重要概念対を評価するための指標として，WordSpace 法による文脈類似度と相関ルールによる信頼度の２つの共起性に基づく統計的な手法を用いた重要概念対の評価方法を用いている．最終的に構築されたオントロジーは，変換モジュールによって OWL 形式に変換される．

4. 領域オントロジーを用いた検索

GXFinder は，日本語 DODDLE により構築された OWL 形式の領域オントロジーを搭載し，概念階層を利用した，絞込検索および拡大検索を行うことができる．

4.1 絞込検索

絞込検索は，検索結果数が膨大な場合に検索結果数を絞込み，ユーザの目的に合った文書を見つけやすくする．検索キーワードを表記として持つ概念の下位概念について，それぞれの概念表記を OR で結合し，検索を行う．例えば，「油」をキーワードとして絞込検索を行うとする．「油」の下位概念には「残油」「既存油」「禁油」「作動油」「防油」が存在するとする．GXFinder で絞込検索を実行すると，上記５つのキーワードを OR で結合し，検索を行う．何階層下まで展開するかはユーザが指定することができる．

4.2 拡大検索

拡大検索は，検索結果数が少なすぎる場合や検索したいキーワードをユーザが漠然としか思い浮かばない場合に，関連する文書を多くユーザに提示し，目的に合った文書を見つけやすくする．拡大検索は，検索キーワードを表記として持つ概念の上位概念および兄弟概念について，それぞれの概念表記を OR で結合し検索を行う．例えば，「油」の上位概念は，「油脂」であり，兄弟概念には「灯油」，「軽油」などが存在するとする．拡大検索を行うと上記のキーワードが展開され，検索が実行される．「油」のみで検索する場合に比べ，油に類似するキーワードを含む文書を検索することができ，目的の文書が見つからない場合でも，目的と類似する文書を見つけることが可能であると考えられる．絞込検索と同様に，何階層上まで展開するかはユーザが指定することが可能である．

表 1: キーワード抽出結果

品詞名	語数
一般名詞	32,415 語
その他の名詞	4,470 語
動詞	1,232 語
未知語	3,689 語
合計	41,806 語

5. ケーススタディ

本ケーススタディの目的は，領域オントロジーを用いた拡大検索および絞込検索の有用性を確かめることである．本ケーススタディは GX 社の協力の下，ロケット運用に関する領域オントロジーを構築し，それを GX 社内文書検索システム GXFinder に搭載し，検索実験を行った．以下では，日本語 DODDLE を用いたロケット運用オントロジーの構築および通常のキーワード検索と領域オントロジーを用いた検索の比較実験について述べる．また，実験の考察について述べる．

5.1 ロケット運用オントロジーの構築

ロケット運用オントロジーの構築は，1 人のユーザが約 30 時間かけて，1．キーワード抽出，2．不要語の削除，3．多義性解消，4．階層構築の手順で行った．

1．キーワード抽出

GX 社豊洲分室で作成されたロケット運用に関する 2845 の日本語文書 (ワード，エクセル，pdf，テキスト等) から形態素解析システム Sen [SEN] によってキーワードを抽出する．その中から一般名詞・動詞・その他の名詞・未知語を抽出した．この際，名詞が連続してつながっている場合，それらをまとめて名詞の複合語とした．例えば，「発射装置」という単語を形態素解析すると「発射」と「装置」に分割されるが，「発射」と「装置」の品詞は共に名詞であるため，「発射装置」は名詞の複合語とみなす．表 1 に抽出された単語の品詞および語数を示す．

2．不要語の削除

Sen により形態素解析解析されたキーワードの中には，切られ方の誤った語や領域に不要な単語などが含まれている．日本語 DODDLE に入力する前に１文字の語，長すぎる単語（20 字以上），文字化けしたと思われる単語については，プログラムにより自動的に削除した．また，明らかに切られ方のおかしい語や領域にとって不要な単語については，ユーザが手動で削除を行った．本作業は，形態素解析の精度が向上することにより軽減されると考えられる．最終的に 32,814 語を日本語 DODDLE の入力とした．

3．多義性解消

日本語 DODDLE の入力モジュールにおいて，EDR 電子化辞書中からユーザは入力単語に対応する領域にとって最も適切な概念を選択する．

大部分の複合語は，それを表記として持つ概念が EDR 電子化辞書中に存在しない．日本語 DODDLE では，部分照合を

行うことによって，多くの複合語の多義性解消を可能にしている．日本語 DODDLE の多義性解消方法は完全照合と部分照合の 2 種類がある．完全照合は，入力単語と EDR 電子化辞書中の概念が持つ表記が完全に一致することを意味する．部分照合は，入力単語と EDR 電子化辞書中の概念が持つ表記が部分的に一致することを意味する．完全照合しなかった入力単語については，Sen を用いて形態素解析を行い，先頭の単語を順に除いて EDR 電子化辞書中の概念と対応付けを試みて，最長一致した単語に対応する概念と対応付けを行う．

例えば,「ロケット発射装置」という入力単語について多義性解消を行うとする.「ロケット発射装置」が完全照合しなかった場合，形態素解析を行い,「ロケット」と「発射」と「装置」に分解する．はじめに,「発射装置」について照合を行い，次に「装置」について照合を行う．この例では,「発射装置」を表記としてもつ概念は EDR 電子化辞書中に存在せず,「装置」を表記として持つ概念が EDR 電子化辞書中に存在する．よって,「ロケット発射装置」の意味として,「装置」を表記として持つ概念を候補としてユーザに提示する．その際に,「ロケット発射装置」を「装置」概念の下位概念とするか,「装置」概念の別表記とするかをユーザは選択可能である．本ケーススタディでは，概念数が膨大であったことから，部分照合した単語については，すべて照合した概念の下位概念とした．

本来は入力単語の文書中の出現位置から入力単語の意味は決定すべきであるが，今回のケーススタディでは文書が非公開であり，多義性解消の際には文書を参照することができなかった．そのため本ケーススタディでは，単語からロケット運用という分野に最も適していると推測される概念を選択した。

部分照合した複合語の多義性解消については，同様に部分照合する複合語の多義性解消結果に統一した．個々の複合語には，ある部分照合した複合語の意味とは異なる意味をもつものも存在する可能性があるため，本来は個別に多義性解消を行うべきである．しかし，今回は入力単語数が膨大で，すべての複合語について多義性解消を行うことが困難であったため，同様に部分照合した複合語の意味はすべて同一であると仮定した．

4．階層構築

階層構築は日本語 DODDLE が自動で行う．日本語 DODDLE 特有の機能として，多義性解消時に部分照合された複合語について，語尾および語頭による階層化を行っている．

語尾が等しい複合語（部分照合した概念が等しい入力単語）は，兄弟概念として階層化される．例えば,「排出設備設置」及び「供給設備設置」が EDR 電子化辞書中の「設置」概念に部分照合した場合，両者は「設置」概念の下位概念として定義される．

部分照合した入力単語の語尾以前（照合しなかった部分）の文字列を表記として持つ概念が構築中の領域オントロジー内に存在する場合，その上位概念と入力単語の語尾を組み合わせた概念を入力単語の上位概念として定義する．例えば,「計器」の下位概念に「レーダ」,「センサー」,「ゲージ」という概念が定義されているとする.「モデル情報」,「レーダ情報」,「センサー情報」,「ゲージ情報」という複合語を階層化する場合，語尾による階層化では,「情報」の下位概念に,「モデル情報」,「レーダ情報」,「センサー情報」,「ゲージ情報」が定義される．ここで，複合語の語尾以前の単語である,「レーダ」「センサー」「ゲージ」については，領域オントロジー中に共通の上位概念である「計器」が定義されている．よって,「計器」と複合語の語尾の「情報」を組み合わせて,「計器情報」という表記を持つ概念を作成し，その下位概念に「レーダ情報」,「センサー情報」,「ゲージ情報」を定義しなおす．これにより,「モデル情報」と「レーダ情報」,「センサー情報」,「ゲージ情報」という計器に関する情報に分類することができる．

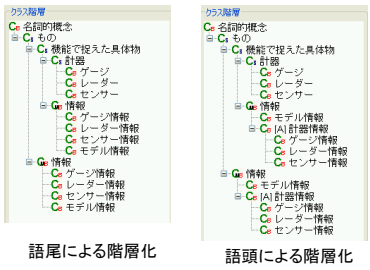

図 3: 語尾および語頭による複合語の階層構築例

図 3 に，語尾および語頭による複合語の階層構築例を示す．

本来は，階層構築後，領域に特化した形に階層を修正すべきである．しかし，本ケーススタディでは，ユーザが初期領域オントロジーを洗練するために十分な領域に関する知識を持っていなかったこと及び概念数が膨大であるために，階層洗練戦略が示唆する概念変動が生じていると思われる箇所が数千カ所におよび確認が困難であったことから，概念階層の洗練は行っていない．

表 2 に，ロケット運用オントロジーにおける入力単語数，完全照合数，部分照合数，未知語数， 概念数を示す．

表 2: ロケット運用オントロジーにおける入力単語数，完全照合数，部分照合数，未知語数，概念数

入力単語数	31817
完全照合数	4982
部分照合数	26835
未知語数	997
概念数	34451

5.2 実験方法

5.1 節で述べた手順で構築したロケット運用オントロジーを GXFinder に搭載し，検索実験を行った．検索は，ロケット運用に詳しい専門家が行った．何もない状態から検索対象となる文書を専門家が思い浮かべることは困難であるため，領域オントロジーの階層を専門家に見てもらいながら検索対象となる文書を想定してもらう．検索対象となる文書を専門家が決定後，はじめにキーワード検索を行う．次に検索結果数が多すぎる場合には絞込検索を，少なすぎる場合には拡大検索を行う．上位 10 件および 20 件について評価を行う．検索結果の評価には，適合率，再現率，F 値を用いた．

5.3 実験結果

本節では 2 つの実験結果を示す．実験 1 は,「発射管制卓」に関する文書を探したいが,「発射」が思い浮かばず,「管制卓」で検索を行うという想定で行った．実験 2 は,「ターミナルカウントダウンシーケンス」に関する文書を探したいが,「ターミナル」が思い浮かばず,「カウントダウンシーケンス」で検索を行うという想定で行った．図 4 に「管制卓」および「カウントダウンシーケンス」周辺の概念階層を示す．

表 3 に実験 1 および実験 2 における通常検索ヒット数，絞込検索ヒット数，上位 10 件および 20 件の通常検索および絞込検索における正解文書数，全文書中の正解文書数を示す．また，表 4 および表 5 に表 3 より求めた，検索結果上位 10 件

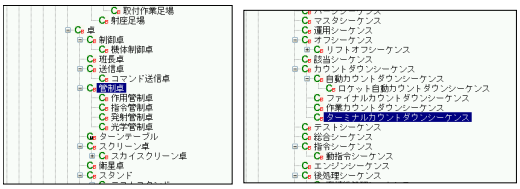

図 4: 「管制卓」および「カウントダウンシーケンス」周辺の概念階層

表 3: 実験 1 および 2 の検索結果

	実験１	実験２
通常検索ヒット数	86	82
絞込検索ヒット数	44	46
通常検索上位 10 件の正解文書数	1	3
通常検索上位 20 件の正解文書数	3	3
絞込検索上位 10 件の正解文書数	3	8
絞込検索上位 20 件の正解文書数	4	11
正解文書数	4	22

および 20 件における通常検索と絞込検索における適合率，再現率，F 値を示す．

実験 1 および 2 の結果より，絞込検索は通常検索に比べ適合率，再現率，F 値のすべてにおいて良い結果を得ることができた．絞込検索では，領域オントロジーの階層構造を利用して，より詳しいキーワードに絞り込んで検索をすることにより，ユーザが検索キーワードを明確に思い浮かべることができない状態でも，目的の文書を探すことを容易にすることが可能であると考えられる．拡大検索については，ユーザが全文書を詳細に把握していることもあり，うまく機能する例題が得られなかった．

5.4 考察

本ケーススタディを通して，未知語の処理，階層構築，拡大検索，ログの活用の主に 4 つの問題が見えてきた．

本ケーススタディでは，未知語の処理は行っていない．専門家は 2 時間程度の検索実験の間に，誤って階層化されている概念，誤って形態素解析された単語，未知語の概念階層中に挿入されるべき位置などを多く指摘していた．未知語の中には領域に特化された概念が含まれている可能性が高いため，今後，専門家が未知語を概念階層中の適切な位置に挿入するのを支援する機能を考える必要がある．

日本語 DODDLE の階層構築方法では対応できない場合が本ケーススタディを通していくつか見られた．例えば，「〜ガス」というガス会社を階層構築する場合，EDR 電子化辞書中に「ガス会社」としての意味が定義されいない場合には，語尾の「ガス」で階層構築されるため，「可燃ガス」や「ヒドラジンガス」などの気体としてのガスの兄弟概念として位置づけられる．また，「ガス」の下位概念に「有毒ガス」が存在するが，ロケット運用における有毒ガスを階層化することは，現在の日本語 DODDLE の階層構築方法では困難である．

本ケーススタディでは，拡大検索がうまく機能しなかった．原因として，キーワードに対応する概念の兄弟概念が非常に多く，拡大検索を行うと幅広くキーワードが展開されてしまい，検索ヒット数が膨大になりすぎたことが考えられる．兄弟概念数を減らす方向でオントロジーを洗練することで，拡大検索が

表 4: 実験 1 の検索結果

	再現率	適合率	F 値
通常検索 上位 10 件	0.250	0.100	0.143
絞込検索 上位 10 件	0.750	0.300	0.429
通常検索 上位 20 件	0.750	0.150	0.250
絞込検索 上位 20 件	1.000	0.200	0.333

表 5: 実験 2 の検索結果

	再現率	適合率	F 値
通常検索 上位 10 件	0.136	0.300	0.188
絞込検索 上位 10 件	0.364	0.800	0.500
通常検索 上位 20 件	0.136	0.150	0.143
絞込検索 上位 20 件	0.500	0.550	0.524

有効に機能すると考えられる．兄弟概念数を減らす方法の一つとして，日本語 DODDLE では語頭による複合語の階層構築を行っている．語尾のみによる複合語の階層構築と比べて，兄弟概念数を減らすことが可能だが，まだ不十分である．今後より兄弟概念数を減らすためには，構築中の領域オントロジーだけでなく，汎用オントロジー中の概念階層も考慮して，複合語の概念階層の洗練を行う方法が考えられる．

本研究では，GXFinder により取得されたログを領域オントロジーに反映させることはできていない．検索結果のログが徐々に蓄積されるため，検索されたキーワード周辺の階層をログを元に洗練することができれば，検索結果および領域オントロジーを洗練できると考えられる．

6. おわりに

本研究では，オントロジー搭載型検索システムおよび日本語オントロジーを半自動で構築可能なツールである日本語 DODDLE の提案を行った．ロケット運用分野におけるケーススタディを通して，日本語 DODDLE を用いることにより，初期領域オントロジーを半自動で構築できることを示した．また，領域オントロジーを用いた絞込検索が，通常検索よりも精度が高いことを示した．拡大検索については，初期概念階層の兄弟概念数が多すぎるために，検索結果が膨大となり，適切に機能しなかった．今後は，拡大検索が有効に機能するように，また，より領域に特化したオントロジーを構築できるように，領域オントロジーの洗練に取り組む予定である．

参考文献

[HP Labs 03] HP Labs, : Jena Semantic Web Framework (2003), http://jena.sourceforge.net/downloads.html

[Morita 04] Morita, T., Shigeta, Y., Sugiura, N., Fukuta, N., Izumi, N., and Yamaguchi, T.: DODDLE-OWL: OWL-based Semi-Automatic Ontology Development Environment, *Evaluation of Ontology-based Tools* (2004), http://mmm.semanticweb.org/doddle/

[SEN] SEN, : http://ultimania.org/sen/

[中川 03] 中川 裕志，森 辰則，湯本 紘彰：出現頻度と連接頻度に基づく専門用語抽出，自然言語処理，pp. 29–35 (2003), http://gensen.dl.itc.u-tokyo.ac.jp/

[日本 01] 日本電子化辞書研究所：EDR 電子化辞書 (第 2 版) 仕様説明書 (2001)