業務概念体系（オントロジー）に基づく概念クラス図構築支援

Supporting Conceptual Class Modeling with Domain Ontology

神谷慎吾[†‡]　石橋昌彦[†]　森田武史[†]　山口高平[†]
Shingo Kamiya[†‡]　Masahiko Ishibashi[†]　Takeshi Morita[†]　Takahira Yamaguchi[†]
†慶應義塾大学大学院　理工学研究科
‡(株)ＮＴＴデータ　技術開発本部
†Graduate School of Science and Technology, Keio Univ.
‡Research and Development Headquarters, NTT Data Corp.

要旨
概念スキーマ（概念クラスモデル）構築は上流工程の要の一つであるが、現状は要員スキル依存で品質のばらつきが大きい。そこで、オントロジー技術に基づき業務概念体系をデータ構造化し、本情報を活用して要求文書等から概念クラス図案を導出することでモデラーを支援する方法を提案する。

1. はじめに

情報システム開発の受託側であるＩＴ企業は、一般に発注側業務分野の背景知識に乏しい。発注側との業務知識差に起因してしばしば外部仕様の抜け漏れや錯誤が生じ、大きな手戻りの元凶となる。この現状を打破する一方法として、先駆者が苦労して獲得した業務背景知識を形式知として外部化し、後続類似案件支援や後進育成等に生かす路線が考えられる。業務背景知識は多岐に渡ることから、当該知識の中核を成す業務概念体系（業務上の基本概念群及びそれらの概念構造）に関する形式知化と活用を初期目標とし、上記路線の有効性を見極めることとした。業務概念体系を形式知化する枠組みとしては、オントロジー技術が適合する。開発工程としては概念クラス図（概念スキーマ）作成工程に対応することから、本工程の作業支援体系（手順や支援環境等）を構築し、その有効性を評価する。

我々は既に上記アプローチに沿って、全体手順や個別技法の確立、プロトタイプ開発、簡単な仕様例を用いた適用評価等を進めてきている [1], [2], [3]。個別技法の各論についてはこれら先行報告の引用にとどめ、本報告では、情報システム開発合理化の観点から、全体的な流れの紹介と応用上の有効性に関する考察を示す。以下では、まず 2 節で全体的な流れと個別手順を概説し、3 節で簡単な適用評価に基づく有効性や残存課題の考察を行う。最後に、4 節でまとめとして今後の方向性等を述べる。

2. 概念クラス図構築支援の枠組み

標記支援体系（図１）の中核は、対象業務分野の基本概念（専門用語とその意味）とそれらの構造（関連等）を蓄積する領域オントロジー（一種の DB）である。加えて、領域オントロジーを補佐する汎用オントロジー（電子化辞書等、一般的な単語とその意味を構造化した DB）も用いる。これら DB に、領域オントロジー構築支援系（手順と支援環境、以下オントロジー系）及び概念クラス図作成支援系（以下クラス図系）が付加されている。前者は領域オントロジー自体を初期構築或いは一括充実させる枠組みであり、複数の開発プロジェクトをサポートする共通技術支援部門等による遂行を想定している。後者は、個別情報システム開発プロジェクトの概念クラス図作成工程を支援する枠組みである。

領域／汎用オントロジーのスキーマ（図２）は類似しており、用語の綴りと概念（意味）の構造とから成る。概念の構造はクラス図に類似しており個々の概念がクラスに相当するが、クラス名が用語綴り（単語）として分離され、綴りと概念との関連が多対多になる点が特徴である。これは、類義語（標記ゆれを含む）や多義語を適切に扱うためである。国語辞典的な一般用語を格納する汎用オントロジーの用語綴りは通常単純（単純語＝単一の単語）だが、領域オントロジーに登録される専門用語は一般に複合語であるため、複合語を構成する要素単語群の並び構造を併せて保持する。概念間の関係中、IS_A（継承）関係は共通的だが、その他の一般的な関連は目的に応じて異なってくる。通常入手可能な汎用オントロジーは電子化「辞書」であるため、添付される一般関連は、例えば格関係など自然言語処理的応用

を想定した内容である。一方、領域オントロジーのスキーマ設定は利用目的に応じて任意であることから、今回の目的では、概念クラス図と同様な関連をそのまま格納することとした。

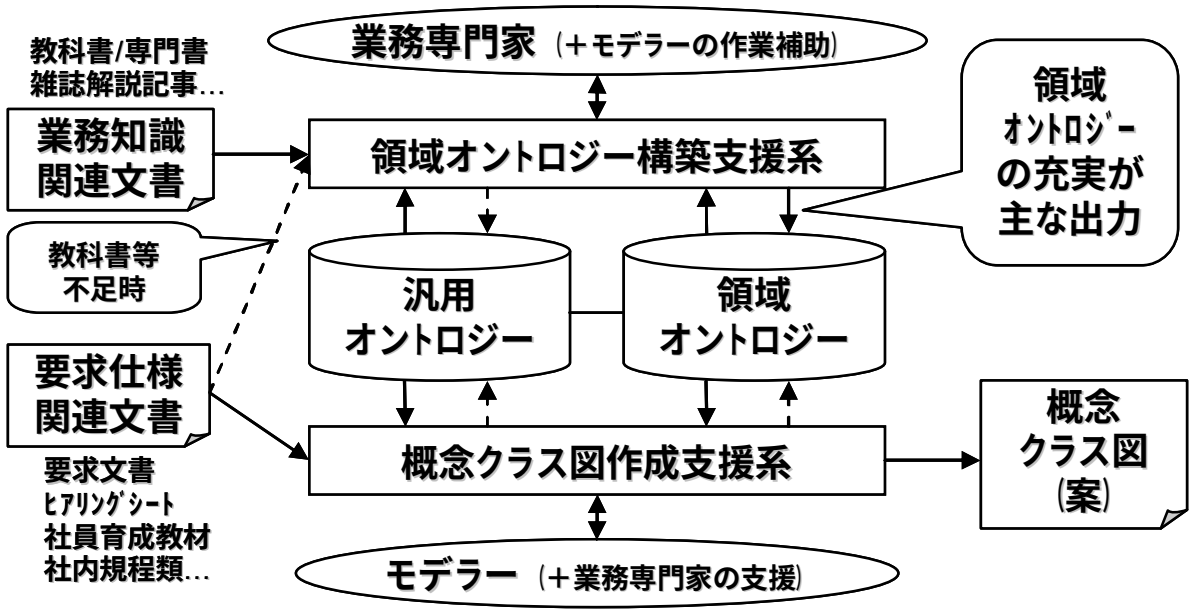

図1　概念クラス図構築支援体系

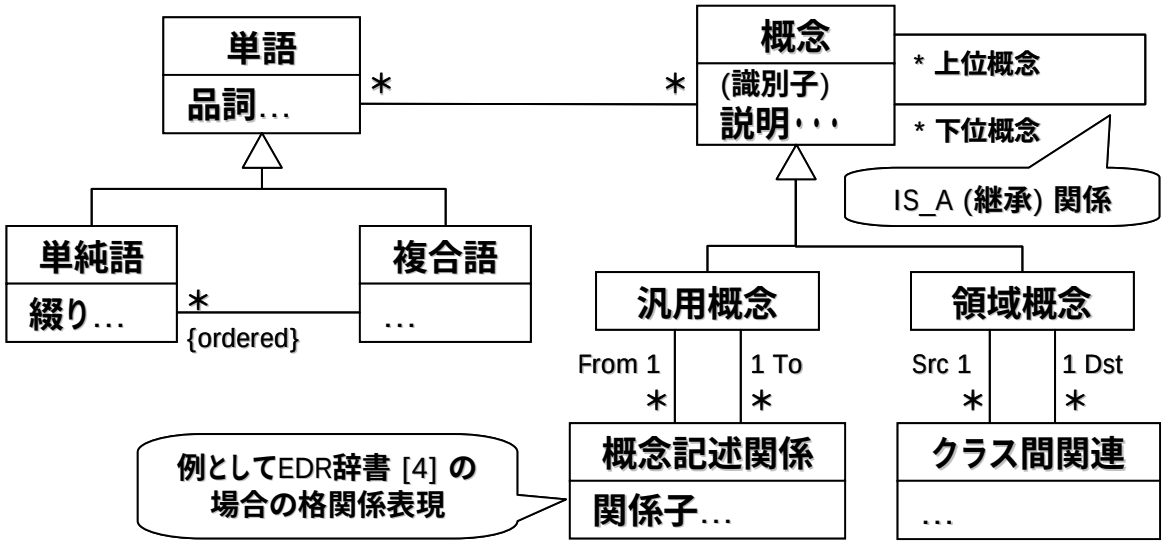

図2　領域／汎用オントロジーのスキーマ概要

オントロジー系とクラス図系の手順は類似しており、図３のように進む。個々の処理ステップはほぼ同様だが、オントロジー系では主に汎用オントロジーを用い業務専門家支援下で領域オントロジーに情報を付加するのが主眼であるのに対し、クラス図系では主に領域オントロジー内の情報を活用して（必ずしも業務専門家でない）モデラーが特定案件の概念クラス図を作成する作業を支援するのが主眼となる。以下では、各ステップを概説する。

用語抽出では、入力文書から対象業務分野の専門用語（主に複合語）を抽出する。入力文書は文がアウトライン構造化された形式と想定する。単純な場合は単に文の列である。処理としては、まず入力文書を形態素解析して単語を抽出する。今回は茶筌 [5] を用いた。次に複合語抽出ツールを用いて複合語も抽出する。今回は言選 [6] を用いた。このとき、汎用オントロジーとの照合等のため、複合語内の単語構造を保持しておく。例えば複合語「出庫依頼票」について (出庫, 依頼, 票) を保持する。複合語抽

出ツールが出力する複合語重みも記録する。また、モデラーによる介入（洗練作業）支援や用語対の共起情報利用等のため、抽出した用語の出現箇所も保持する。更に、抽出した用語群と領域オントロジー中の複合語綴りとの照合を行い、照合一致した用語には重みを加点する。自然言語処理に完璧を期するのは困難ゆえ、次ステップに進む前にモデラー等が結果を吟味する。主な作業は、複合語の単語連結誤りの訂正や、明らかに専門用語でない一般語の捨象等である。

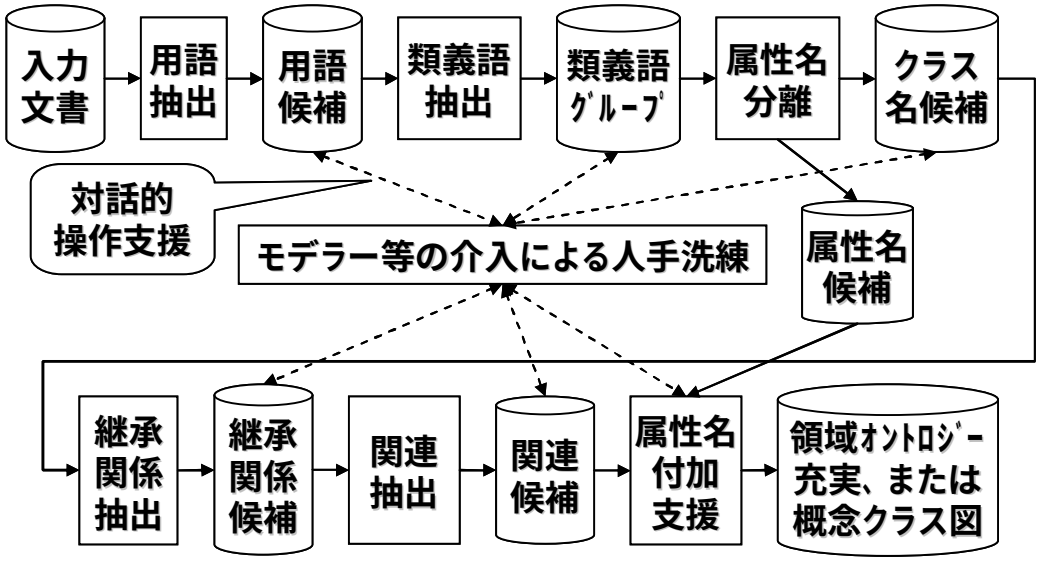

図3　オントロジー／クラス図系の処理手順概要

類義語抽出では、誤解の元となりやすい類義語の混在を排除して用語を統一するため、まず自動的に類義語候補を抽出する。領域／汎用オントロジーにおいて同一概念に繋がる用語群が類義語候補となる。複合語の場合は、含まれる要素単語毎の類義語（含、標記ゆれ）もチェックする。例えば「コンピュータ機器」と「コンピューター機器」等である。この後、やはりモデラー等が介入して洗練を行う。主な作業は、類義語か否かの最終判断、類義語の場合にどの用語に統一するかの指定等である。

属性名分離では、上記手順で得られた用語群から属性名を切り分けてクラス（エンティティ）的な用語に絞り込む。これにより、クラス間の構造を扱う継承関係や関連の同定処理を効率化する。処理方法は、まず、領域／汎用オントロジーを用いて用語に接続する概念を取得し、本概念の上位概念を辿る。このとき予め選定してある幾つかの上位概念に当たったなら、比較的高い確度で、当該用語が名称、数量、時間に関する用語と判定できる。これらを属性名候補として分離する。最後はモデラー等が介入して分類エラーの補正を行う。

継承関係の同定は、２つの方法の組合せにより行う。１つは、オントロジー中にある継承関係を用いる方法であり、もう一つは複合語の短縮による方法である。後者の例は、「優良顧客」の冒頭から単語を削って得られる「顧客」が、「優良顧客」の上位概念とみなす等である。最後はモデラー等が介入して正誤判断を行う。支援環境が抽出した継承関係候補が、最終的に類義語と判断される場合もある。

関連の同定は、基本的には領域オントロジー中に格納された関連に照合するか否かによる。オントロジー中の関連に十分な付加情報を保持しておけば、多重度や関連端名候補も支援可能である。しかしながら、オントロジー系のように領域オントロジーが充実していない段階の作業では、本方法は十分に機能しない。そこで、確度は低くなるが、汎用オントロジーのみを用いる方法も併用する。一つは用語対の共起情報を用いた絞込みで、関連を持つ用語対は入力文書中で近接して現れるはずとの仮説に基づく。加えて、汎用オントロジー中の格関係を用いた照合処理を行う。通常、格関係は「名詞－動詞」の関係であることから、入力文書中の動詞も抽出保持しておき、「名詞－動詞－名詞」の２段照合により関連候補を抽出する。最後はモデラー等が介入し、過不足の補正を行う。

属性名付加では、予め分離しておいた属性名を戻し適切なクラスに配置する。基本は共起度と領域オ

ントロジー照合による方法だが、領域オントロジー不足時は現状では共起度による支援のみとなり、モデラー等による補完量が多くなる。しかしながら、属性名については、元文書を参照すると図表等にて明記してあることがしばしばあり、関連同定ステップよりは容易に補完できると想定される。

3. 試行適用結果と考察

上記手順を支援するプロトタイプ環境を構築し、古典的な酒類販売会社在庫問題 [7] を用いて試行評価をいった。使用した汎用オントロジーは EDR 辞書 [4] である。上記アプローチの有効性を確認できた点もあるが、課題も残る。以下、議論を生じたステップ毎に考察を示す。

用語抽出は単純だが手作業では面倒なステップであり、自動抽出の恩恵を示しやすい。しかし、少数ではあるが、「旨依頼者」のような不正解の抽出、「日時」抽出漏れのような正解の看過が見られた。モデラー等にとって不正解の訂正は比較的容易だが、抽出漏れの解消では形態素解析結果や入力文書の全スキャン等を行うため補正工数が大きい。かつ、本ステップで漏れた用語は後で挽回するステップがないため、最終結果まで漏れることとなり影響も深刻である。正解率向上より漏れ防止を重視した用語抽出方法や、漏れの補正を効率化する支援方法等を追加検討する必要がある。

属性名分離ステップでは、概ね前述の技法で分離され、単純手作業からの解放は確認できた。しかしながら、「空コンテナ搬出マーク」のように名称、数量、時間の上位概念では判定しにくい属性名が残存した。属性名分離ルールを逐次追加できる機構が必要と考えられる。

概念クラス図構築の肝は関連の抽出にあるが、領域オントロジー不足段階での支援方法が不十分であり、十分に良い結果が得られない。そこで、現在、幾つかの追加アプローチを検討中である。第一に、[8] にあるような深い自然言語処理を行う方向がある。第二に、汎用オントロジーにも含まれる継承関係を用いて用語を人、物、場所、事象等に分類し、[9] にあるような分析パターンを適用する方向がある。第三に、汎用オントロジー中にない複合語において最後の単語だけを照合に用いる現行方法（例えば「出庫指示」→「指示」）を越え、他の要素単語（前例では「出庫」）も加味する方向がある。

4. まとめ

継承関係抽出までのステップは、現行方法の延長で十分に支援効果が得られる感触を得た。関連同定についても、領域オントロジー充実下では有意な支援効果を示せると考える。大きな課題は、領域オントロジー自体を構築する局面等、領域オントロジー不足段階での関連同定法の強化である。また、現状のプロトタイプはステップ毎機能がシームレスに連携されていないため、実務向け環境からはほど遠い。今後は、よりシームレスな環境を整備し、本環境を用いた定量的効果測定に進む予定である。

参考文献

[1] 山口高平, 樗松理樹, 青木千鶴, 関内律恵子, 加賀山茂, 吉野一, “計算機可読型辞書を利用した領域オントロジー構築支援環境”, 人工知能学会誌, Vol.14, No.6, pp.1080-1087, 1999.

[2] 峰岸巧, 石橋昌彦, 福田直樹, 飯島正, 山口高平, “オントロジーに基づくソフトウェア開発上流工程支援”, The 19th Annual Conference of the Japanese Society for Artificial Intelligence, 2005.

[3] 神谷慎吾, 石橋昌彦, 森田武史, 福田直樹, 飯島正, 山口高平, “オントロジーを用いた分析クラス図作成支援”, KBSE2005-15, 信学技報, Vol.105, No.208, pp.25-30, 2005.

[4] 日本電子化辞書研究所, EDR 電子化辞書(第 2 版)仕様説明書, TR2-006(改), 2001.

[5] 松本裕治他, 形態素解析システム『茶筌』version 2.3.3 使用説明書, 奈良先端科学技術大学院大学, 2003.

[6] 中川裕志他, “専門用語（キーワード）自動抽出システム”, http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb.html.

[7] 山崎利治, “共通問題によるプログラム設計技法解説”, 情報処理学会誌「情報処理」, 25 巻 9 号, 1984.

[8] 原田実, 野村佳秀, 山本幸二, 大野雅志, 田村浩樹, 高橋史郎, “自然語要求仕様からオブジェクト指向設計図を自動生成するシステム CAMEO”, 情報処理学会論文誌, Vol.38, No.10, pp.2031-2039, 1997.

[9] ジル・ニコラ, マーク・メイフィールド, マイク・アベニー, ストリームラインオブジェクトモデリング, (株)デュオシステムズ(訳), 今野睦(監訳), ピアソン・エデュケーション, 東京, 2002.