

(40) シームレスな相互運用のための建設用語とコア構成要素の対応付けに関する研究

Research on Coherence between Terms of Construction Fields and Core Components for Seamless Interoperability

北野光一¹・中村健二²・田中成典³

Kitano Koichi, Nakamura Kenji, Tanaka Shigenori

抄録： 情報処理技術の発展に伴い、様々な業界において電子商取引が盛んに行われるようになった。そこで、国連の下位機関である UN/CEFACT とコンピュータと通信の標準化団体である OASIS は、異業界間の電子商取引を円滑化するために、XML を用いたインターネット上の企業間電子商取引のための国際的な仕様群である ebXML を提案した。ebXML に則って電子商取引用データを作成することで、様々な分野間でのデータ交換が可能となった。一方、土木建設分野におけるデータ項目は、ebXML と対応付けがなされておらず、国際的な情報交換や、複数企業間でのデータ連携において様々な問題が発生している。そこで、本研究では、タグ用建設用語の候補を法律、要領や規定集等の信頼性の高い文書から抽出し、ebXML と自動的に関連付ける手法を提案する。

キーワード： 電子データ交換, ebXML, コア構成要素, ビジネス情報エンティティ, オントロジー, シソーラス, 形態素解析, 概念ネットワーク

Keywords : Electronic Data Interchange, ebXML, Core Component, Business Information Entity, Ontology, Thesaurus, Morphological Analysis, Concept Network

1. はじめに

情報処理技術の発展に伴い、情報システムを用いて、国や分野を横断した電子商取引が行われるようになった。しかし、国や分野によって、同様の概念が異なる語彙で表現されることやデータ項目に違いなどが発生し、データ流通の障害となっていた。そこで、国連の下位機関である UN/CEFACT とコンピュータと通信に関する標準化団体である OASIS が共同で、XML を用いたインターネット上の企業間電子商取引のための仕様群である ebXML¹⁾ を提案した。このことにより、ebXML に則って電子商取引用データを作成することで、様々な分野間でのデータ交換が可能となった。

一方、土木建設分野においても、CALS/EC の推進により公共工事の受発注等において電子納品や電子商取引が推進されている。しかし、土木建設分野におけるデータは国土交通省や各自治体、企業が独自に定めた仕様に基づいて構築されている。そのため、国際的な情報交換や企業が構築した情報システム間でのデータ連携において様々な問題が発生している。

これらの背景を踏まえ、我々は、建設分野において一般的に活用される用語セットを建設業界の XML タグ用の標準用語集であるコア用語（以下、タ

グ用建設用語）と定義し、これらの用語を土木用語大辞典から選定する研究²⁾を行ってきた。しかし、土木用語大辞典のみからタグ用建設用語を抽出した場合、既存仕様の用語との親和性が保証されておらず、タグ用建設用語として公開しても、一般的に普及しないものとなる危険性が高い。そのため、これまでの研究の発展として、本研究では、タグ用建設用語の候補を Web 上の信頼のおけるデータ（法律、要領、規定集など）から抽出する手法と抽出したタグ用建設用語の候補を国際標準である ebXML と自動的に関連付ける手法を提案する。

2. 研究の概要

(1) 研究内容

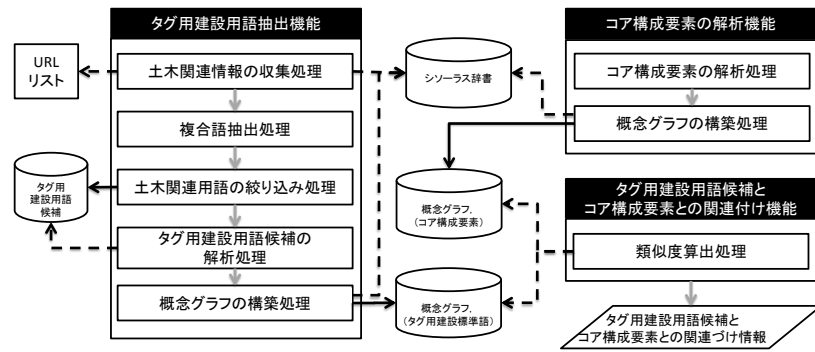
本研究では、「既存研究で抽出したタグ用建設用語が、土木用語大辞典に含まれる語にのみ依存する課題」と「抽出したタグ用建設用語と国際標準である ebXML との関連性が考慮されていない課題」の 2 つに対応した、タグ用建設用語の抽出手法を提案する。本手法では、Web 上の信頼のおけるデータからタグ用建設用語を抽出し、抽出した用語と意味的に類似した ebXML のコア構成要素^{3), 4)}を選択し、対応付けを行う。このことにより、土木建設分野にお

1： 学生会員 修士(情報学) 関西大学大学院 総合情報学研究科

(〒569-1095 大阪府高槻市霊仙寺町二丁目一番一号, Tel : 072-690-3213, E-mail : noahany@gmail.com)

2： 正会員 博士(情報学) 立命館大学 助手 情報理工学部 (〒525-8577 滋賀県草津市野路東一丁目一番一号)

3： 正会員 工学博士 関西大学 教授 総合情報学部 (〒569-1095 大阪府高槻市霊仙寺町二丁目一番一号)



図－1 処理の流れ

る用語と ebXML の用語との連携が可能となり、土木建設分野においても国際的な電子商取引や、他分野との円滑な情報交換が可能となると考えられる。

(2) 処理の流れ

本提案手法のシステムの流れを図 1 に示す。本システムは 3 つの機能で構成されており、タグ用建設用語抽出機能では、まず、Web 上の信頼の置ける情報リソースから言語資源を収集し、複合語抽出を行ってタグ用建設用語の候補を抽出する。次に、抽出した用語 T に対して形態素解析を行い、用語に含まれる形態素集合 $M=\{m_1, \dots, m_n\}$ を取得する。そして、形態素に意味的な情報を付加するため、シソーラス辞書を用いて形態素の概念情報を取得する。なお、本研究では、シソーラス辞書として日本語の概念間の関係を記述した日本語 WordNet⁵⁾を用いる。このとき、1 つの形態素 m には 1 つ以上の概念情報が存在するため、概念集合 $C=\{c_1, \dots, c_p\}$ を取得し、形態素集合 M の要素全てに対応する C の集合を $C_M=\{C_1, \dots, C_n\}$ とする。次に、得られた概念集合 C の要素 c を起点として、上位の概念を繰り返し取得し、最上位の概念までを取得して概念グラフ g を作成す

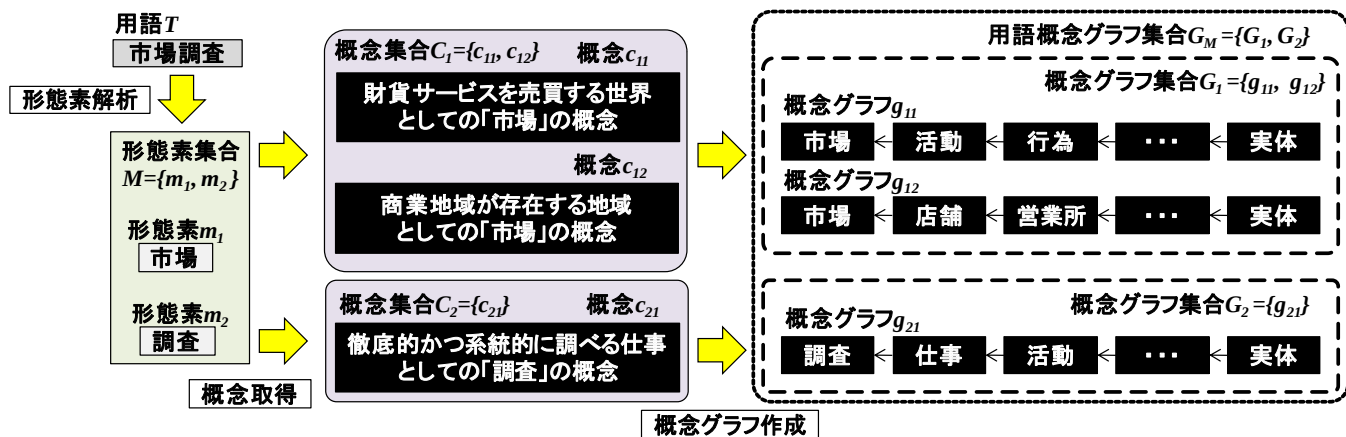
る。また、 C の各要素に対応するグラフ集合を $G=\{g_1, \dots, g_p\}$ とし、 C_M に対応する G の集合を G_M とする。概念グラフのイメージを図 2 に示す。図 2 に示した概念グラフは、「市場調査」の用語から作成した概念グラフを示す。コア構成要素の解析機能では、コア構成要素のオブジェクトクラス名の形態素を解析し、タグ用建設用語抽出機能と同様の処理を行うことで、概念グラフを構築する。最後に、タグ用建設用語候補とコア構成要素との関連付け機能では、タグ用建設用語抽出機能とコア構成要素の解析機能で作成した概念グラフ同士を比較し、類似度を算出する。そして、算出した類似度に基づき、タグ用建設用語候補とコア構成要素を関連付けて出力する。

3. タグ用建設用語候補とコア構成要素の類似度算出アルゴリズム

2.2 節で登場した概念グラフ $g_1=\{c_{11}, \dots, c_{1n}\}$ と $g_2=\{c_{21}, \dots, c_{2m}\}$ があるとき、 g_1 から g_2 への類似度を算出する手順を次に示す。

表－1 各指標の内容

指標	説明
指標 1	タグ用建設標準語の末尾の形態素から作成した概念グラフ集合 $G_1=\{g_{11}, \dots, g_{1n}\}$ とコア構成要素名の末尾の形態素から作成した概念グラフ $G_2=\{g_{21}, \dots, g_{2n}\}$ において、 $SIM_1(G_1, G_2)=1/\sum sim(g_{1i}, g_{2j}) / (Count(G_1) Count(G_2))$ を類似度の値とする。
指標 2	指標 1 と同様に作成した概念グラフ集合 G_1 と G_2 において、 $SIM_2(G_1, G_2)=1/\sum sim(g_{2j}, g_{1i}) / (Count(G_1) Count(G_2))$ を類似度の値とする。
指標 3	指標 1 と同様に作成した概念グラフ集合 G_1 と G_2 において、 $SIM_3(G_1, G_2)=1/\min(sim(g_{11}, g_{21}), \dots, sim(g_{1n}, g_{2m}))$ を類似度の値とする。
指標 4	指標 1 と同様に作成した概念グラフ集合 G_1 と G_2 において、 $SIM_4(G_1, G_2)=1/\min(sim(g_{21}, g_{11}), \dots, sim(g_{2m}, g_{1n}))$ を類似度の値とする。
指標 5	タグ用建設標準語の形態素集合 M_1 から作成した概念グラフ集合 $G_{M1}=\{G_{11}, \dots, G_{1p}\}$ とコア構成要素名の形態素集合 M_2 から作成した概念グラフ集合 $G_{M2}=\{G_{21}, \dots, G_{2q}\}$ において、 $SIM_5(G_{M1}, G_{M2})=\max(SIM_1(G_{11}, G_{22}), \dots, SIM_1(G_{1\min(p,q)}, G_{2\min(p,q)}))$ を類似度の値とする。
指標 6	指標 5 と同様に作成した概念グラフ集合 G_{M1} と G_{M2} において、 $\maxindex=\min(p, q)$ とするとき、 $SIM_6(G_{M1}, G_{M2})=\max(SIM_2(G_{11}, G_{21}), \dots, SIM_2(G_{1\maxindex}, G_{2\maxindex}))$ を類似度の値とする。
指標 7	指標 5 と同様に作成した概念グラフ集合 G_{M1} と G_{M2} と $\maxindex$ において、 $SIM_7(G_{M1}, G_{M2})=\max(SIM_3(G_{11}, G_{21}), \dots, SIM_3(G_{1\maxindex}, G_{2\maxindex}))$ を類似度の値とする。
指標 8	指標 5 と同様に作成した概念グラフ集合 G_{M1} と G_{M2} と $\maxindex$ において、 $SIM_8(G_{M1}, G_{M2})=\max(SIM_4(G_{11}, G_{21}), \dots, SIM_4(G_{1\maxindex}, G_{2\maxindex}))$ を類似度の値とする。
指標 9	指標 5 と同様に作成した概念グラフ集合 G_{M1} と G_{M2} と $\maxindex$ において、 $SIM_9(G_{M1}, G_{M2})=\sum SIM_1(G_{1i}, G_{2i}) / \maxindex$ を類似度の値とする。
指標 10	指標 5 と同様に作成した概念グラフ集合 G_{M1} と G_{M2} と $\maxindex$ において、 $SIM_{10}(G_{M1}, G_{M2})=\sum SIM_2(G_{1i}, G_{2i}) / \maxindex$ を類似度の値とする。
指標 11	指標 5 と同様に作成した概念グラフ集合 G_{M1} と G_{M2} と $\maxindex$ において、 $SIM_{11}(G_{M1}, G_{M2})=\sum SIM_3(G_{1i}, G_{2i}) / \maxindex$ を類似度の値とする。
指標 12	指標 5 と同様に作成した概念グラフ集合 G_{M1} と G_{M2} と $\maxindex$ において、 $SIM_{12}(G_{M1}, G_{M2})=\sum SIM_4(G_{1i}, G_{2i}) / \maxindex$ を類似度の値とする。



図－２ 「市場調査」の概念グラフ

1. c_{1i} が g_2 内に含まれている場合は、 g_1 から g_2 への類似度は $1/i$
2. c_{1i} が g_2 内に含まれていない場合は、 i の値に 1 を加え、再度手順 1 を実行
3. i が n と等しくなっても c_{1i} が g_2 内に含まれていない場合は、 g_1 から g_2 への類似度は 0

このとき、 g_1 から g_2 への類似度を求める関数を $\text{sim}(g_1, g_2)$ とする。本研究では、グラフの類似度を算出する関数 sim を用いて、タグ用建設用語とコア構成要素名との類似度を表 1 に示す複数の指標で算出する。

4. 評価実験

本研究では、提案手法の有効性を評価するため、「実験 1: タグ用建設用語の抽出実験」と「実験 2: タグ用建設用語とコア構成要素との関連付けアルゴリズムの精度評価実験」とを行う。それぞれの実験の詳細を次に示す。

(1) タグ用建設用語の抽出実験

a) 実験内容

本実験では、信頼における Web リソースからタグ用建設用語が抽出可能であるかを評価するため、道路に関する 16 の法令の文章を解析して、タグ用建設用語の候補の抽出を行う。なお、本実験では、土木施工管理関係法規集⁶⁾において、道路に関する項目の中で重要と設定されている法令を採用する。本実験の手順を次に示す。

1. 道路に関する法令の文章を取得
2. 複合語抽出手法⁷⁾を用いて、各種法令文書から複合語を抽出
3. 手順 2 で抽出した語の中から、既存研究²⁾の手法で抽出したタグ用建設用語に一致する語のみを抽出
4. 手順 3 の解析結果を分析し、タグ用建設用語の抽出の可能性を検討

手順 4 では、手順 2 の複合語に対して既存研究²⁾の手法を適用することで、タグ用建設用語の候補が抽出できる可能性があるかどうかを検証するため、土木用語大辞典から抽出したタグ用建設用語の候補の中から、手順 3 で抽出できた用語に着目して解析を行う。このことにより、法令から手順 2 にて抽出した複合語の中で、既存研究²⁾で定義した用語の特徴と一致した複合語は、タグ用建設用語の候補として採用可能であるかを検証することができる。

b) 結果と考察

本実験で用いた 16 の法令から複合語を抽出したところ、3,754 語を取得できた。取得した語の中で、既存研究²⁾の用語と一致する 161 語の一部を表 2 に示す。表 2 に示すとおり、いずれの用語もタグ用建

表－２ 既存研究の用語と一致する語（一部抜粋）

居住環境	停止線	開発行為
都市再開発	合成勾配	計画交通量
重量制限	軸重	規制標識
停車帯	商業地域	縦断勾配
損失補償	優先道路	緩和区間
河川区域	車両総重量	波浪
重要港湾	斜め駐車	補助標識
漁港	照明施設	冠水
禁止標識	流域下水道	市街化区域
設計図	直角駐車	支持力
縁石線	大型自動車	美観地区
遮音壁	地域住民	除雪機械
狭窄部	対向車線	安全地帯
都市高速鉄道	路上駐車	地区計画
視線誘導標	公共下水道	資金計画
隅角部	副道	景勝地
原因者負担金	再開発	市街化調整区域
都市施設	沿道整備計画	中央帯
緩衝建築物	警戒標識	道路台帳
交通量調査	案内標識	地上権
交通島	工業用水道	輪荷重
地下横断歩道	信号機	一般電気事業

設用語として重要であり、一般的に土木建設業界で広く利用される可能性のある用語であることが分かる。これらの結果から、16の法令から複合語抽出した3,754語に対して、既存研究²⁾と同様の解析を行うことで、土木建設業界で広く利用される用語を抽出することができると思われる。

(2) タグ用建設用語とコア構成要素との関連付けアルゴリズムの精度評価実験

a) 実験内容

本実験では、本研究で提案した概念グラフを用いた類似度の算出方法が有効であるかを検証するため、あらかじめコア構成要素との関連付けアルゴリズムが有用であるかを検証するため、あらかじめコア構成要素との関連が明らかとなっているビジネス情報エンティティの名称を用いて、コア構成要素との関連付け精度を評価する。本実験では、次世代EDI推進協議会が公開する「コア構成要素辞書(2009年B版)³⁾」中のコア構成要素集合体のオブジェクトクラス名(385件)、「業務情報項目辞書(2009年B版)⁴⁾」中の業界用語・同義語の項目をタグ用建設用語として採用する。また、実験の精度評価は、ビジネス情報エンティティに対して類似度が高いコア構成要素を関連付け、その中に正解が含まれるかどうかで評価する。また、評価する際の指標として、表1に示す12の指標を用いる。

b) 結果と考察

実験結果を表3に示す。表に示すとおり、1位の場合は、指標8の283(55.16%)が最良であることが分かった。指標8は、ビジネス情報エンティティを構成するすべての形態素を用いて類似度を算出し、類似度が最大となるコア構成要素を選択する指標であり、指標4の末尾の形態素のみを類似度算出に用いるものと比較して、良好な結果が得られていることが分かる。これらのことから、用語を構成するすべての形態素から概念グラフを構築し、類似度が最大となるコア構成要素を選択することで、優良な結果が得られることが分かった。また、誤判定したデータを確認したところ、シソーラス辞書に掲載されていない用語や、他のコア構成要素との差が現れず、正しいコア構成要素を判定できない事例がみられた。これらの課題については、今後類似度算出のアルゴリズムを改良して用語の共起ベクトルを考慮することで、解消する予定である。

5. おわりに

本研究では、Web上の信頼のおけるデータからタグ用建設用語を抽出し、抽出した用語と意味的に類似したebXMLのコア構成要素を選択し、対応付け

表-3 各指標の順位毎の対応付け精度

指標	1位	5位以内	10位以内
指標1	152 (29.62%)	205 (39.96%)	237 (46.19%)
指標2	111 (21.63%)	201 (39.18%)	222 (43.27%)
指標3	266 (51.85%)	276 (53.80%)	297 (57.89%)
指標4	271 (52.82%)	277 (53.99%)	283 (55.16%)
指標5	120 (23.39%)	192 (37.42%)	216 (42.10%)
指標6	106 (20.66%)	197 (38.40%)	224 (43.66%)
指標7	271 (52.82%)	277 (53.99%)	292 (56.92%)
指標8	283 (55.16%)	285 (55.55%)	288 (56.14%)
指標9	125 (24.36%)	197 (38.40%)	240 (46.78%)
指標10	94 (18.32%)	210 (40.93%)	241 (46.97%)
指標11	222 (43.27%)	251 (48.92%)	271 (52.82%)
指標12	226 (44.05%)	254 (49.51%)	274 (53.41%)

する手法を提案した。評価実験の結果、タグ用建設用語は正しく抽出されていることを証明した。

本研究により、Web上に存在する膨大な情報資源から、建設分野で広く一般的に利用される標準用語を抽出することが可能となる。さらに、コア構成要素との対応付け結果を提供することで、国際的な情報システム間連携を考慮したデータ構造の実現のために参考となる情報を提示可能となる。今後は、Webリソースから抽出したタグ用建設用語に対して、既存研究²⁾の手法を適用した結果を考察する予定である。

謝辞: 財団法人日本建設情報総合センターの方々には大変お世話になりました。ここに記して感謝致します。また、本研究の一部は、平成20～24年度私立大学戦略的研究基盤形成支援事業(研究課題「セキュアライフ創出のための安全知循環ネットワークに関する研究」)から助成を受けその成果を公表するものである。

参考文献

- 1) OASIS: ebXML, <<http://www.ebxml.org/>>, (入手 2011.7.29)。
- 2) 北野光一, 中村健二, 田中成典, 古田均: 情報の相互運用性を高めるためのコア用語の選定手法に関する研究, 土木情報利用技術論文集, 土木学会, Vol.19, pp.301-310, 2010。
- 3) JEDIC: コア構成要素辞書 2009年B版, <http://www.jipdec.or.jp/dupc/jedic/dictionary/dictionary_download.html>, (入手 2011.7.29)。
- 4) JEDIC: 業務情報項目辞書 2009年B版, <http://www.jipdec.or.jp/dupc/jedic/dictionary/dictionary_download.html>, (入手 2011.7.29)。
- 5) Kuroda, K., Bond, F., and Torisawa, K.: Why Wikipedia needs to make Friends with WordNet, The 5th International Conference of the Global WordNet Association, pp 9-16, 2010。
- 6) 国土交通省大臣官房技術調査課: 土木施工管理関係法規集, 2000。
- 7) 小山照夫: 日本語テキストからの複合語用語抽出, 情報知識学会誌, 情報知識学会, Vol.19, No.4, pp.306-315, 2009。