Japanese version of reprint from "New Generation Computing (NGC)."
Accepted at 26/6/2015 by JSAI (Japan Society of Artificial Intelligence)

特集論文 「実践 Linked Open Data」

TEXT2LOD
～ テキスト情報の LOD 化に向けた Web API の開発 ～

TEXT2LOD
– Development of Web API for Triplification of Text Information –

川村 隆浩 *1
Takahiro Kawamura
(株) 東芝 研究開発センター
Corporate Research & Development Center, Toshiba Corp.

大須賀 昭彦
Akihiko Ohsuga
電気通信大学 大学院情報システム学研究科
Graduate School of Information Systems, University of Electro-Communications, Japan

keywords: information extraction, triplification, linked open data, LOD

Summary

Linked Open Data (LOD) is recently attracting attention as a vast amount of distributed knowledge base on the Web. Thus, semi-structured data such as tables and hierarchical data in several domains have been triplified to the LOD. In the research area, however, triplification of unstructured data such as text and sensor data is actively studied as the next target. Therefore, we developed a Web API for mainly extracting triples from text data, which is useful for the triplification of text data. We defined two steps for the text triplication. The first step is the extraction of phrases, which correspond to triple $< subject, verb, object >$, location and time from a natural language sentence, and the second one is a conversion of the extracted phrases to the existing (or new) resources and properties in the LOD. In this paper, we first describe the service specification corresponding to the first step, technical background, and evaluation of the current extraction accuracy, then finally introduce some use cases of the service. Although this service adopts a novel combination of a restrictive method using ontology-based rules and an example-based machine learning method using conditional random field, based on probability distribution, the main cotribution of the service is in practical aspect, that is, mash-up of several natural language processing techniques as a text triplification service, and deployment as a Web API freely available for public use so that non-expert easily use it.

1. は じ め に

近年，Web を巨大な分散知識ベースとする Linked Open Data(LOD) が国内外で普及しつつある．それに併せて，これまで表形式や階層構造で表されてきた，いわば半構造化データの LOD 化が各分野で盛んに進められている．一方で，研究面では次のターゲットとしてテキスト情報やセンサーデータなど非構造化データの LOD 化に注目が集まっている [Usbeck 14]．そこで本論では，LOD 普及のきっかけとなることを期待して，テキスト情報を LOD 化する際に役立つ，テキストから主にトリプルを抽出する Web API の開発，公開について述べる．

我々が考えるテキスト情報の LOD 変換処理は以下の 2 つのステップからなる．

(1) 自然文からのトリプル $< Subject, Verb, Object >$，および Location や Time に該当する文字列の抽出
(2) 抽出された文字列の既存（または新規）リソース，プロパティへの変換

ステップ 1 は，自然言語処理における Semantic Role Labeling（SRL，意味役割付与），および Named Entity Recognition（NER，固有表現抽出）に相当する処理である．SRL とは，機械翻訳や質問応答など自然言語処理アプリケーションの精度向上を目的とした研究分野の 1 つであり，自然文の中から「誰が (Subject, Agent)，何を (Object)，誰に (Patient)，どうした (Action, Predicate)」というような単語間の意味的な関係を抽出する処理である．格文法 [Fillmore 68] においては，文の意味構造を動詞–深層格–名詞という関係の集合として捉えており，SRL で抽出される関係は格文法における深層格に相当する．但し，抽出対象とされる関係は研究によって相違がある [Mooney 14, Sammons 14a]．また，主に日本語処理では，Predicate Argument Structure Analysis（PAS，述語項構造解析）とも呼ばれ，述語（または事態を意味する名詞，事態性名詞）とそれに対応する項を抽出し，ガ格，ニ格，ヲ格と呼ばれる表層格のラベルを付与する処理として定義されることもある [吉川 13]．また，時間や場所は

†1 現在は，国立研究開発法人 科学技術振興機構 情報分析室に所属．

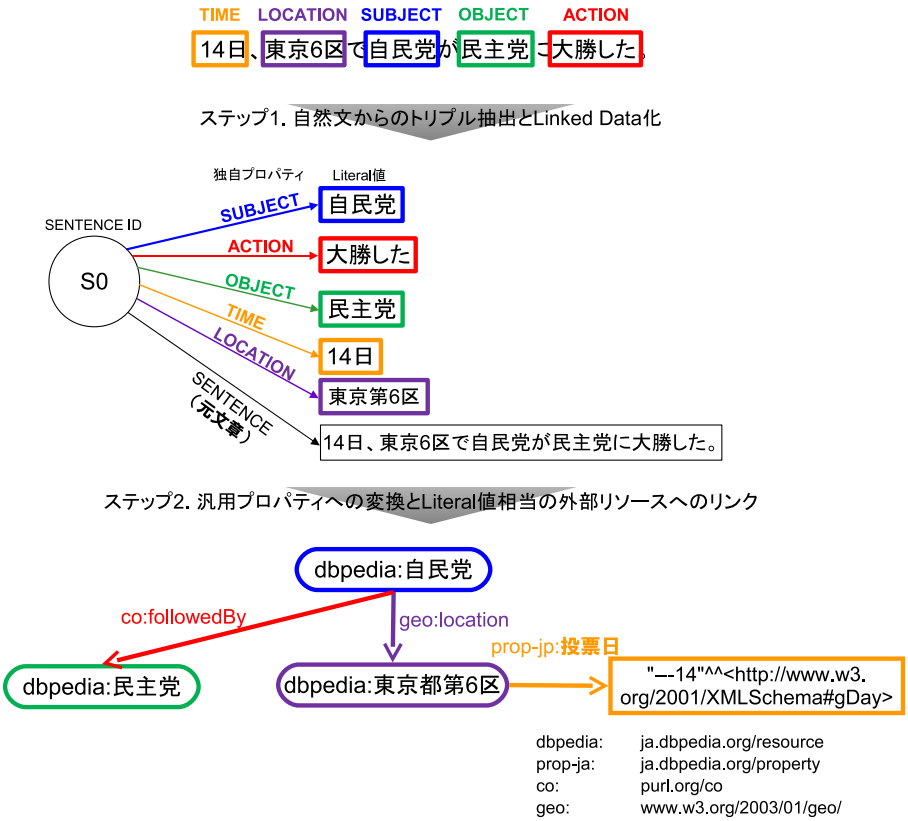

図 1 自然文の LOD への変換例

時間格 (Time)，場所格 (Location) として扱われる場合もあるが，多くの場合，SRL では対象とされない．一方で，時間や場所は NER の一環として，person や organization と共に time, location として抽出対象とされることが多い．

ステップ 2 は，主に名詞句を DBpedia Spotlight[*1]などを用いて，文字列から DBpedia 上の該当するリソースに変換する処理である．更に，動詞句を Linked Open Vocabularies(LOV)[*2]や BBC Ontologies[*3]上の該当するプロパティに変換し，リソース同士をリンクしてもよいだろう．実際のリソース，プロパティ変換では，部分文字列マッチングや Jaccard 係数を用いた String マッチング処理，WordNet オントロジーやドメイン辞書を用いた Semantic マッチング処理が行われることが多い．

こうした処理により，自然文を大きく 2 つの形式の LOD に変換することを意図している．1 つは，名詞句のみを変換するものであり，リソースとしては何かしらの ID を使ったものである．Freebase や Kira[Kira 12] らの研究で用いられている形式である．もう 1 つは，動詞句もプロパティとして変換し，Subject(Agent) と Object をプロパティでリンクする形式であり，DBpedia などで用いられている形式である．但し，全ての文がこの形式に変換できるわけではない．図 1 に変換例を示す．または，[Kawamura 14a] を参照してほしい．

LOD については提唱者の Tim B. Lee より 4 原則 [Lee 06] が示されている通り，可能な限り汎用的なプロパティを使うこと，外部リソースにリンクすることが実際の利用において重要である．ステップ 1 は，文章構造からトリプル $< S, V, O >$ と Time，Location を抽出し，Linked Data を表す RDF (Resource Description Framework) / XML (eXtensible Markup Language) 形式に変換するが，ここで使われているプロパティセットは本サービス独自のものであり，値は文字列（Literal 型）として入っており，外部リソースにリンクしていない．そのため，より良い LOD とするためには，ステップ 2 において Literal 値に相当する外部リソースへのリンク付け，独自プロパティの汎用プロパティへの変換が必要である．これにより，広く外部の方に使ってもらえる LOD となると考えている．しかし，ステップ 1 のみであっても作成した Linked Data 自体の公開は行わず，文章構造を表す内部グラフデータとしてさまざまな外部サービスに活用することができる．また，ステップ 2 はどのリソース，プロパティに割り当てるかが対象ドメインに大きく依存するため，汎用的なサービスとして一般化することが難しい．そこで，本研究ではステップ 1 にフォーカスしたサービスを開発，公開した．そして，本論ではステップ 1 の変換精度と，変換された内部データのサービス活用実績について述べる．以下，2 章でステップ 1 のサービスの仕様や変換処理の実装，現状の精度について述べる．また，3 章では本サービスを活用した研究事例について紹介し，4 章で関連研究を示す．最後に 5 章でまとめと今後の課題について述

*1 spotlight.dbpedia.org
*2 lov.okfn.org
*3 www.bbc.co.uk/ontologies

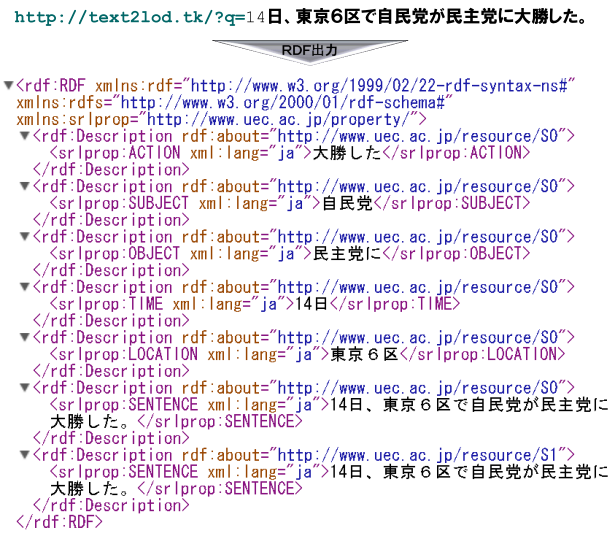

図 2 サービス入出力の例

べる．

2. TEXT2LOD サービスの開発

本章では，提案サービスの仕様，実装の詳細，ユースケースについて述べる．

2·1 サービス仕様

本サービスは，REST ベースの Web API として以下のアドレスで無償公開されている．

http://text2lod.tk/?q=

パラメータ q に続けて変換したい文章（一文単位）を HTTP GET で送信すると，RDF に変換された文を XML 形式で返却する．図 2 に入出力の例を示す．

抽出する関係は，Illinois 大の Sammons ら [Sammons 14b] と同様に Subject（ガ格またはハ格の名詞句），Action（述語の主辞を含む動詞句），Object（ヲ格またはニ格の名詞句）のみの SVO 形式とした．これは，本サービスの目的は言語学的に厳密な SRL の開発ではなく，文章からトリプルを抽出して LOD 化を助ける点にあるためである．尚，RDF における SVO が，文法学上の主語・述語・目的語の意味とは異なる場合もあることはよく知られているが，ここでは補語もプロパティ値としてトリプルに組み入れるため，便宜上，Object として抽出している．また，NER として Location（場所）と Time（時間）も併せて抽出している．Subject や Object が複数存在する場合はそれぞれ Subject や Object を別個に出力する．複文の場合は，ID(={S0, ..., Sn}) の異なる複数のトリプルを出力する．連体節など形容詞的な Action を含む文も主節と従属節とで複数のトリプルを出力する．この場合，従属節の Action に掛かる Subject と Object が揃っていれば，別 ID とする．Subject，Object のいずれかならば，Action の掛かる主節の名詞へリンクし，同 ID としている．但し，Linked Data では受動態のプロパティも頻繁に用いられるため（例えば，DBpedia Japanese において用いられている rdfs:isDefinedBy や wdrs:describedby，dbpedia-owl:influencedBy，prov:wasDerivedFrom など），格交替が起こった際も主語と目的語の入れ替えは行わず，受動態の動詞句を Action プロパティの値として持つ Linked Data を生成し，ステップ 2 において受動態のプロパティに変換されることを意図している．

制限事項として，公開した Web API は複数文の一括処理には対応していないため，文を跨いだ指示語による共参照関係の解決には対応していない．文内であれば，指示代名詞は直前の Subject または Object に置き換えられる．但し，我々のサンプル調査によれば，ニュース記事等において “それ” や “あれ” といった指示語の頻度は低い．また，同様の理由から文を跨いだゼロ照応関係（対応する格がない場合．特に日本語ではガ格が省略されることが多い）の解決にも対応していない．しかし，同一文内のガ格に関しては，直前の Subject が存在すれば，それが設定される．

2015 年 1 月現在，本サービスは Amazon EC2 (Elastic Compute Cloud)，Instance type: t2.micro (Intel Xeon プロセッサー 2.5GHz–3.3GHz, 1 VCPU, 1 GiB メモリ) で稼働しており，処理時間は文の長さに依存するが一文あたり 0.1–0.2 秒程度である．

2·2 変換処理の実装

図 3 に LOD への変換処理の流れを示す．

学習フェーズでは，まず自然文を形態素解析器 MeCab[*4] と係り受け解析器 CaboCha[*5] に掛けて単語（形態素）に分割し，それぞれの品詞や係り受け関係，および CaboCha による固有表現抽出結果を取得する．そして，以下の規則に沿ってチャンク（句）を生成し，文構造を単純化する．

- 助詞*n
- 助動詞*n
- 名詞＋助詞
- 名詞＋助動詞
- 動詞＋助動詞
- 動詞＋助動詞＋助詞
- 名詞*n
- 動詞*n
- 接頭詞＋名詞
- 記号*n
- 名詞＋記号（句読点以外）
- 記号（句読点以外）＋名詞
- 指示語＋名詞
- 形容詞＋助詞
- 副詞＋助詞

*4 mecab.sourceforge.net
*5 code.google.com/p/cabocha/

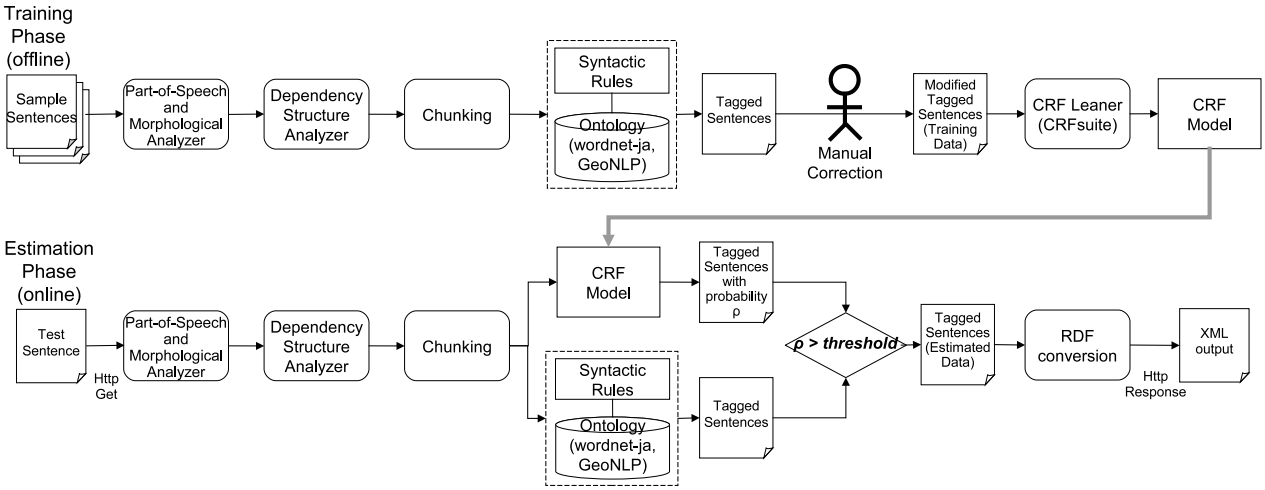

図 3 変換処理の流れ

ここで，*n は同品詞の単語の 2 回以上の繰り返しを表す．名詞や動詞，形容詞，副詞と結合された品詞は，その品詞句の一部として扱われ，上記ルールが繰り返し適用される（名詞＋助詞＋助詞⇒名詞句，など）．但し，助動詞 “である” や “です” は，is-a プロパティに変換するために独立して扱うなど，いくつかの例外は存在する．更に，辞書ベースでの連続語の結合も行っている（“緑の党”，など）．

次に，ルール適用によってチャンク間の関係を抽出する．ここで抽出される関係ラベルは，最終的に出力される Subject，Object，Action，Location，Time に加えて，Modifier（修飾句），Because（因果関係），Other（その他）の 8 種類である．修飾句としては，形容詞，副詞，品詞結合後の名詞の内，助詞 { の，に，な，の，で，ような，ように，ようで } で終わるような形容動詞，連体詞などが該当する．但し，ニ格の名詞句は補語となる場合もあり，その場合は前述したように Object として扱われる．また，修飾節は独立したトリプルとして扱われる．その他は，いずれにも判定されなかったものであり，主に記号と一部の副詞である．ルールは，各チャンクに対して付けられた品詞レベルでのラベル，CaboCha による固有表現ラベル（IOB2 形式．B-PERSON や I-DATE など），およびオントロジーを参照したラベルと，チャンク間の文法的な制約関係の組み合わせから構成されている．例えば，[品詞ラベル：NOUN ∧ endWith(“が”) ⇒ Subject]，[品詞ラベル：NOUN ∧ endWith(“を”) ⇒ Object]，[品詞ラベル：VERB ∧ 文末 ⇒ Action]，[固有表現ラベル：LOCATION ∧ endWith(“では”) ⇒ Location]，[固有表現ラベル：TIME ∧ endWith(“に”) ⇒ Time]（Location，Time も判定し直している），[オントロジーラベル：人 ∧ Action：“移動” ⇒ Subject] といったルールが約 300 件登録されている（endWith は語尾に付くことを意味する）．チャンクに対して，オントロジーラベルを付与するにあたっては，事前にいくつかの概念クラス（人，地名，時間表現など）を定義している．それぞれのクラスにはインスタンスとして約 1000 の単語が登録されている．例えば，時間表現クラスであれば，{ 来年，翌年，昨年，先般，昨今，今般，上旬，中旬，下旬，...} などが含まれている．そこで，例えばチャンクの文字列が，“近ごろ” であった場合，それ自体がインスタンスとして登録されていなくても，WordNet Japanese[*6] で類義語（または上位語）を取得し，登録済みのインスタンスとマッチさせることで（この場合，類義語 “昨今” がマッチする），該当チャンクのオントロジーラベルが時間表現クラスであることを判定する．尚，日本国内の地名に関しては GeoNLP[*7] を一部，参照しており，国内外で約 6000 の地名がインスタンスとして登録されている．尚，海外の地名に関してはカタカナまたは漢字表記にのみ対応している．そして，ルール適用結果を 2 名の研究員が目視で確認，修正し，学習用の正解データを構築した．尚，元文章はニュース記事，ソーシャルメディア等から政治，経済，技術，芸能等に関して収集した約 2000 文である．

これを系列ラベリング問題と見なして Conditional Random Fields（CRF，条件付き確率場）[Lafferty 01] の学習器 CRFsuite[*8] に通して学習モデルを構築した．特徴量としては，単語毎の単語名と品詞名，および 8 種類のいずれかの関係タグから，以下の 19 種類を生成した．

- $w[t-2], w[t-1], w[t], w[t+1], w[t+2]$,
- $w[t-1]|w[t], w[t]|w[t+1]$,
- $pos[t-2], pos[t-1], pos[t], pos[t+1], pos[t+2]$,
- $pos[t-2]|pos[t-1], pos[t-1]|pos[t], pos[t]|pos[t+1], pos[t+1]|pos[t+2]$,
- $pos[t-2]|pos[t-1]|pos[t], pos[t-1]|pos[t]|pos[t+1], pos[t]|pos[t+1]|pos[t+2]$

ここで t は文内における単語の位置，w は単語名，pos は品詞名を表す．| は単語または品詞の連続（n-gram）を表す．これらに正解データとして 8 種類のいずれかのタグ

*6 nlpwww.nict.go.jp/wn-ja
*7 geonlp.ex.nii.ac.jp
*8 www.chokkan.org/software/crfsuite

を先頭に付けたものが，1 単語分の学習ベクトルとなる．

推定フェーズでは，まず学習フェーズと同様に対象文を形態素解析と係り受け解析にかけてチャンクを生成する．そして，CRF 用の特徴量を生成し，8 種類いずれかの関係タグの推定を行う．一方で，学習データ構築の際に使用したルールも適用し，ルールによるタグの判定も行う．その上で，8 種類の関係タグそれぞれに設定した閾値によって CRF 推定結果か，ルール判定結果のいずれかの値を採用して最終結果とする．ルール判定結果と CRF 推定結果の切り替え手法，および精度については次節にて述べる．

最後に，Modifier は後に続く Subject や Object，Action，Time と結合する．これはステップ 2 にてリソース，プロパティに変換する際に，String マッチングや Semantic マッチング処理のための手掛かり情報とするためである．また，Other は削除する．そして，前章で述べた仕様に沿ってトリプルを構成し，RDF を出力する．公開サービスは，一文単位での処理のため本稿では詳しく触れないが，文内に因果関係節があれば，Because も併せて出力される（例：明日は雨なので、会社を休む。）．

2·3 精 度 評 価

本節では，ニュース記事，およびソーシャルメディアから無作為に抽出した 100 文（但し，1 つ以上の述語が含まれている文に限る）を対象に精度評価を行った結果を示す．まず，表 1，表 2 にルール単体での判定精度，CRF 単体での推定精度をそれぞれ示す．尚，推定値と正解値の総数が同じであるため，加重平均に関しては適合率と再現率は同値となっている．

表 1 関係別の抽出精度（ルール判定結果）

(%)	Subject	Object	Action	Loc.	Time	加重平均
適合率	92.77	79.55	93.28	50.00	56.67	91.52
再現率	92.77	95.54	86.38	100.00	77.27	91.52

表 2 関係別の抽出精度（CRF 推定結果）

(%)	Subject	Object	Action	Loc.	Time	加重平均
適合率	94.12	84.10	95.08	50.00	100.00	93.00
再現率	96.39	89.73	90.27	33.33	18.18	93.00

次に，CRF による各関係の推定確率分布を前節の学習用データ 2000 文について調べたところ，標準偏差で最大 18.1%，中央値で 62.3%，(算術）平均で 61.3%と属性毎に大きく異なることが分かった．そこで，ルール判定結果と CRF 推定結果の切り替え閾値は全体で一律の値を設定するのではなく，属性毎に設定することとした．そして，CRF 単体の方がルール単体よりも精度が高いことから，属性毎の推定確率の分布上，ボリュームゾーン以上であれば CRF を採用するため，閾値を-1σ（平均値-標準偏差 ×1，正規分布を仮定すると上位 84%相当）と設定した．その結果を表 3 に示す．ルール単体，CRF 単体の精度を上回っていることが確認できる．また，閾値を単純に属性毎の推定確率の平均値に設定した場合は，ルール判定結果の影響が大きく，却って CRF 単体よりも精度が下がることを確かめた（加重平均で 92.4%）．

表 3 関係別の抽出精度（閾値-1σ での統合結果）

(%)	Subject	Object	Action	Loc.	Time	加重平均
適合率	92.86	86.23	97.85	66.67	70.00	94.16
再現率	93.98	95.09	88.72	66.67	63.64	94.16

実験結果を文単位で細かく見ていくと，規則に従った文章であればやはりルール判定が強いことが確認できる．また，Location の精度が他プロパティと比べて低い理由は，主に海外の地名情報の不足が原因と思われる．Time の精度不足は Modifier との判別の難しさにある．例えば，「○○は将来（Time），重要だ」や「○○は将来（Modifier）の課題だ」などがそれに該当する．

3. サービス活用事例

本章では，テキスト情報の LOD 化の意義を明確にするため，これまでに本サービスを用いて開発した応用システム，サービスについて述べる．

震災時ユーザ行動の分析 [Nguyen 12] では，Twitter 上の tweet を Linked Data 形式で構造化し，行動系列や因果関係を辿ることで住民行動の把握や避難計画に役立てることを試みた．先の震災時に得られた教訓の 1 つとして，リアルタイムでの情報共有が挙げられるだろう．そこで，我々はユーザの行動を OWL（Web Ontology Language）によるオントロジーとして定義し，震災時の tweet35 万件から 15 種のユーザ行動属性を抽出，ユーザ行動に関する Linked Data によるグラフデータベースを構築した．更には，グラフ上でのユーザ行動の性質（ユーザの類似度，行動の連続性，成功行動か否か）を考慮した行動ベース協調フィルタリング手法を考案し，Twitter において呟かれなかった行動（欠損行動）を推測して行動ネットワークを補完する手法を開発した．本サービスに関しては，現在，防災ソリューションの一環として川崎市で実証実験を計画している．

ソーシャル×マスメディア比較 [川村 14c] では，両メディア情報を Linked Data として形式化し，比較を容易にすることで，デマや偏向報道など情報の信頼性判断のサポートを行った．昨今，ソーシャルメディアにおけるデマの拡散や，マスメディアにおける偏向報道・情報操作の疑いなど，ユーザ自身が情報の信頼性について自ら判断することが求められてきている．そこで我々は，一般ユーザが膨大なメディア情報を多角的な観点から比較することを支援するため，ユーザに代わってソーシャル，マス両メディアから特定の話題に関する情報を抽出，見える化し，特定の観点に基づく比較ポイントを提示するシステムを開発した．本システムでは，Twitter 上の tweet

とマスメディアのニュース記事から 13 の属性情報を持つ事象情報を抽出して Linked Data 化し，多様性，希少性，偏在性，因果関係の 4 つの観点に沿って SPARQL クエリーを用いた比較を実現している．本サービスに関しては，某新聞社との協業が検討されている．

コールセンターログ分析 [Kawamura 14a] では，ユーザからのクレームとソーシャルメディアでの口コミを比較することで，炎上に繋がる可能性のある製品欠陥を早期に発見することを試みた．近年，製品の不具合等に関してコールセンターに寄せられるクレームの件数が急増しており，メーカーは対応に苦慮している．特に，クレーム対応は顧客への初期対応が重要であり，対応を誤るとすぐにネット上で炎上し，製品の評判，販売に大きく影響すると言われている．しかし，コールセンターに一報が入った段階ではその不具合がユーザの使用状況等によって発生するものなのか，機種全体に共通する不具合なのかを切り分けることが難しい．そこで，事前にソーシャルメディアの製品に関する評判情報を Linked Data 化し，クレームの内容とサブグラフマッチングさせることで，そのクレームが氷山の一角であるのかどうかを自動的に判定するシステムを開発した．同様の分析が当社内で活用されている．

モバイル音楽推薦サービス [Wang 14] では，日々の tweet やユーザの行動ログ，歌詞情報を LOD 化しておき，ユーザコンテキストから n 次の隔たりを辿ることでその場にあった音楽を推薦するシステムを開発した．昨今，ユーザのニーズ，コンテキストを取り込んだ音楽推薦システムが数多く提案されている．しかし，コールドスタート問題（使い始めは何も推薦されない）や，推薦曲が似たようなものばかりになってしまう点が指摘されている．そこで我々は，まず Last.fm，Yahoo!ローカル，Twitter そして LyricWiki から情報を集めて大規模な LOD を構築し，スマートフォンのセンサーで得られたユーザの現在の状況から連想関係を辿ることで楽曲を推薦するサービスを開発した．ここで連想関係とは，あらかじめ定義された複数のルールに基づいて LOD 内のリンクを辿る SPARQL クエリーである．そして，セレンディピティの観点からユーザ評価を行い，一定の効果を示した．

ニュースキュレーションサービス [横尾 14] では，はてなブックマークなどユーザのお気に入りの記事から，構造的に類似性の高い記事を推薦するサービスを開発した．近年，ユーザの行動やソーシャルメディア上での発言を興味・関心として分析し，ニュース記事を推薦するキュレーションサービスが普及している．膨大な情報から自分で必要なものを探さなくても，自身の興味に関連した情報が手に入ることで利用者が増加している．しかし，既存のコンテンツベースの記事推薦システムでは頻出する語句を重視しており，語句間の関係を特徴として用いていない．そのため，頻出語が相互に関係を持たなくとも，文内のいずれかに存在すれば推薦されてしまうことがある．そこで，ニュース記事を Linked Data 化し，語句間の意味的リレーションを用いることで，ユーザのお気に入り記事と確かな関連性のある記事を収集・推薦するシステムを開発した．

4. 関 連 研 究

本章では，SRL 技術に関する関連研究を示す．しかし，英語と日本語でテストデータが異なることに加えて，抽出対象とする関係（Semantic Role）がそれぞれの研究で異なっているため，精度面の直接的な比較は難しく，主にアプローチや機能面の比較に留まる．

SRL へのアプローチは大きく 2 つに分けられるだろう．1 つは，主に動詞が取り得る格構造のパターンに関する大規模言語資源，Fillmore らが構築した FrameNet や VerbNet，日本語では京都大学格フレーム辞書や動詞項構造シソーラスを用いて，述語が取り得る名詞との関係に制約（選択制限）を掛ける方法である．あるいは，neo-Davidsonian 形式の一階述語論理 [Terence 90] を用いて制約を掛ける方法もあるが，ベースとなるコンセプト集合の知識記述に大きなコストがかかるという問題がある．もう 1 つは，用例に基づく手法であり，格構造付きの用例（例文）を集め，機械学習技術などを用いて入力文と最も類似度の高い用例から確率的に関係を求める手法である．

英文解析における代表的な研究の 1 つとしては，Gildea らによる 50000 文の述語や品詞，文内での位置などを Bayesian Network を用いて学習させた研究 [Gildea 02] が挙げられるだろう．ここでは FrameNet を対象に 80.4%の正解率を実現している．昨今では，NIST（National Institute of Standards and Technology）主催の TAC 2011 RTE-7 で 1 位を獲得した NEC Laboratories America らの SENNA[Collobert 11a, Collobert 11b] がよく知られている．SENNA は Deep Learning を用いて，形態素解析や係り受け解析，NER，SRL などのタスクを行うシステムであるが，特徴として大量のラベルなしテキストデータの言語モデルを利用して上記各タスクを統合的に学習している点にある．SENNA は CoNLL-05 で開催された SRL competition のデータを対象に F 値 75.49%を達成している．

より平易な文章へのアプローチとしては，Microsoft Research Asia らによる tweet を対象とした研究 [Liu 10, Liu 11] が挙げられる．ここでは，フォーマルな文章（ニュース記事など）との内容的な類似性（コサイン類似度など）から tweet 群をカテゴライズし，フォーマルな文章の解析結果を利用して砕けた文章の解析を行うという，一種の Self-Supervised Learning アプローチが取られている．それらをドメイン毎に CRF で学習し，F 値 66.0%を達成している．

日本語の解析に関しては，前述した Cabocha の解析結

果を利用した述語項構造解析器 SynCha[*9]が挙げられる．NAIST テキストコーパスと日本語語彙体系を用いて，動詞と格要素の共起尺度などを SVM で学習したものである．解析単位が一文単位ではなく複数文の処理に対応しており，文を跨いでゼロ照応関係を解決する点に特徴がある（前文照応解析）．一方，日本語語彙体系を必要としないよりポータブルな Yucha[*10]の開発も進められている．

また，KNP[*11]は，京都大学コーパスから得られた統計的情報と 69 億文から構築した格フレームに基づいて 係り受け解析と述語項構造解析を行う．一文単位であるが，文全体を見て最適な構文・格構造を決定する．また，格フレームに含まれる全ての格が解析対象となってる．

一方で，上記がいずれも述語とガ，オ，ニ格の出力であるのに対して，ASA[*12]は動作主，対象，相手，経験者といった約 80 種類の意味役割（深層格）を出力する点に特徴がある．

これらの既存研究に対して，本サービスはオントロジーを活用したルールによる制約手法と，CRF を用いた用例からの学習手法とを確率分布に基づいて統合する方法を提案している点に特徴がある．尚，一階述語論理と統計的アプローチを組み合わせた統計的関係学習法 Markov Logic Network を用いたアプローチ [吉川 13, Meza-Ruiz 09] など，考え方としては本アプローチと類似の研究も見られるが，同様のアプローチは見つかっていない．既存の複数の自然言語処理ツールを組み合わせて，本サービスと同様のサービスを構築することは可能である．しかし，そうしたツールは使い方や解析結果が専門的であり，LOD 構築を行う専門外の人がそれを使いこなし，更に RDF として出力させるには一段の労力が必要とされるだろう．そのためか，現時点でそうしたサービスは見つかっていない．したがって，本サービスの主たる貢献は，自然言語処理に関する専門知識がない人でも容易に使えるよう，独自の SRL 方式によってテキスト情報の LOD 化サービスを実現し，Web API として公開した実践面にあると考えている．例として，3-tier システム（プレゼンテーション層（ユーザインタフェース，UI），アプリケーション層（ビジネスロジック），データ層（データベース，DB））として構成されることが多い Web サービスにおいて，本サービスはいわば DB のコンテンツを構築するためのものであり，本サービスをツールとして使うことで，Web サービスの開発者はビジネスロジックと UI の設計，開発に集中できるという利点が挙げられる．そのため，前章のコールセンターログ分析では，開発者は新たに自然言語処理について勉強することなく，検討開始から設計，開発，評価までを約半年という短い期間で終えることができた．

5. まとめと今後の課題

本論では，LOD 普及のきっかけとなることを期待して，テキスト情報を LOD 化する際に役立つ，テキストから主にトリプルを抽出する Web API の開発，公開について述べた．具体的には，まずテキスト情報を LOD 化するステップを 2 つに分け，ステップ 1 に相当する本サービスの仕様，技術的詳細に続けて，現状の抽出精度について説明，最後にサービスユースケースを紹介した．

今後の課題としては，プロパティ毎の精度のばらつきを均すため，表 3 において特に精度の低い Location と Time の精度向上が必要と考えている．精度のゴールは，応用システム・サービスの使い方や目的によって異なるが，本サービスの出力精度がある一定の想定内であれば，応用システム・サービス側で事前に相応の対応（精度補償や使い方の工夫など）を取ることが可能と思われる．しかし，ステップ 2 でプロパティやリソースを変換すると，どのプロパティ・リソースが元のどのプロパティ{Subject, Object, Action, Loc., Time} に相当するのか分からなくなってしまうため，プロパティ間の精度のばらつきを抑えることが必要と思われる．そこで，現状，適合率・再現率共に 90%程度を得ている Suject, Object, Action に合わせて Location と Time の精度を向上させたいと考えている．方策としては，大規模な学習データの構築，およびオントロジーラベルと文法規則の組み合わせからなるルールに，4 章で紹介した格フレームから得られる知識を取り込むことなどが挙げられる．

また，併せてステップ 2 のサービス化にも取り組んでいきたい．ステップ 2 における名詞句へのリソースの割り当てには大きく 2 つの処理があると思われる．1 つは，文字列（RDF におけるリテラルノード）を既存のリソースへ割り当てる処理である．1 章で述べた DBpedia Spotlight は，自然文の中に DBpedia リソースを自動的にアノテートするオープンソースプロジェクトである．基本的なアルゴリズムとしては，部分文字列マッチングを用いてアノテートすべき句の選択（Phrase Spotting）とリソース候補の選択（Candidate Selection）を行った上で，コサイン類似度と TF-IDF を用いて周辺のコンテキストに基づいてリソースを決定（Disambiguation）している．現在も複数の改良が行われているが，2015 年 1 月現在，日本語版は公開されていない．もう 1 つは，複数のリテラルノードの同一性を判定して，新規リソースに集約させる処理である．この問題に対して我々は，Linked Data 内のリテラルノードの同一性を，文字列類似度，意味的類似度，周辺ノードの構造的類似度から判定 (名寄せ) する手法を開発した [Kawamura 14b]．更に，動詞句へのプロパティ割り当てに関しては，対象ドメイン毎に応じたプロパティ選択が必要である．今後，こうした技術のサービ

*9 www.cl.cs.titech.ac.jp/~ryu-i/syncha
*10 hayashibe.jp/yucha
*11 nlp.ist.i.kyoto-u.ac.jp/index.php?日本語構文解析システム KNP
*12 cl.it.okayama-u.ac.jp/study/project/asa/about_asa.html

ス化も検討していきたい．

◇ 参 考 文 献 ◇

[Collobert 11a] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa: Natural Language Processing (Almost) from Scratch, Journal of Machine Learning Research (JMLR), Vol. 12, pp. 2461-2505 (2011)

[Collobert 11b] R. Collobert: Deep Learning for Efficient Discriminative Parsing, Proceedings of 14th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 224-232 (2011)

[Fillmore 68] C. J. Fillmore: The Case for Case, Universals in Linguistic Theory, pp. 1-88 (1968)

[Gildea 02] D. Gildea and D. Jurafsky: Automatic Labeling of Semantic Roles, Computational Linguistics, Vol. 28, No. 3, pp. 245-288 (2002)

[Kawamura 14a] T. Kawamura, S. Nagano, and A. Ohsuga: Deployment of Semantic Analysis to Call Center, Proceedings of 13th International Semantic Web Conference (ISWC 2014), Industry Track (2014)

[Kawamura 14b] T. Kawamura, S. Nagano, and A. Ohsuga: Literal Node Matching based on Image Features toward Linked Data Integration, Proceedings of 2014 IEEE/WIC/ACM International Conference on Active Media Technology (AMT 2014), pp. 174-186 (2014)

[川村 14c] 川村 隆浩, 越川 兼地, 中川 博之, 清 雄一, 田原 康之, 大須賀昭彦: メディア情報の Linked Data 化と活用事例の提案, 電子情報通信学会論文誌, Vol. J96-D, No. 12, pp. 2987-2999 (2013)

[Kira 12] K. Radinsky, S. Davidovich, and S. Markovitch: Learning Causality for News Events Prediction, Proceedings of 21st World Wide Web Conference (WWW 2012), pp. 909-918 (2012)

[Lafferty 01] J. D. Lafferty, A. McCallum, and F. C. N. Pereira: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proceedings of 18th International Conference on Machine Learning (ICML 2001), pp. 282-289 (2001)

[Lee 06] Tim B. Lee: Linked Data, http://www.w3.org/DesignIssues/LinkedData.html (2006)

[Liu 10] X. Liu, K. Li, B. Han, M. Zhou, L. Jiang, Z. Xiong, and C. Huang: Semantic role labeling for news tweets, Proceedings of 23rd International Conference on Computational Linguistics (Coling 2010), pp. 698-706 (2010)

[Liu 11] X. Liu, K. Li, M. Zhou, and Z. Xiong: Collective semantic role labeling for tweets with clustering, Proceedings of 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011), pp. 1832-1837 (2011)

[Meza-Ruiz 09] I. Meza-Ruiz and S. Riedel: Jointly identifying predicates, arguments and senses using markov logic, Proceedings of 2009 Annual Conference of the North American Chapter of the ACL (NAACL 2009), pp. 155-163 (2009)

[Mooney 14] R. J. Mooney: CS 388: Natural Language Processing, http://www.cs.utexas.edu/~mooney/cs388/slides/srl.ppt (2014)

[Nguyen 12] T. M. Nguyen, T. Kawamura, Y. Tahara, and A. Ohsuga: Building a Time Series Action Network for Earthquake Disaster, Proceedings of 4th ACM International Conference on Agents and Artificial Intelligence (ICAART 2012), pp. 100-108 (2012)

[Sammons 14a] M. Sammons: Semantic Parsing (Semantic Role Labeling), http://cogcomp.cs.illinois.edu/page/project_view/7 (2014)

[Sammons 14b] M. Sammons: Semantic Role Labeling Demo, http://cogcomp.cs.illinois.edu/page/demo_view/srl (2014)

[Terence 90] P. Terence: Events in the Semantics of English, MIT Press (1990)

[Usbeck 14] R. Usbeck, A. C. N. Ngomo, M. Roder, D. Gerber, S. A. Coelho, S. Auer, and A. Both: AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data, Proceedings of 13th International Semantic Web Conference (ISWC 2014), pp. 457-471 (2014)

[Wang 14] M. Wang, T. Kawamura, Y. Sei, H. Nakagawa, Y. Tahara, and A. Ohsuga: Music Recommender Adapting Implicit Context Using 'renso' Relation among Linked Data, Journal of Information Processing, Vol. 22, No. 2, pp. 279-288 (2014)

[横尾 14] 横尾 亮平, 川村 隆浩, 清 雄一, 田原 康之, 大須賀 昭彦: 語句間の意味的リレーションに基づくキュレーションエージェント, 電子情報通信学会論文誌, Vol. J98-D, No. 6, pp. 982-991 (2015)

[吉川 13] 吉川 克正, 浅原 正幸, 松本 裕治：Markov Logic による日本語述語項構造解析, 自然言語処理, Vol. 20, No. 2, pp. 251-271 (2013)

〔担当委員：乙守 信行〕

2015 年 01 月 09 日 受理

著 者 紹 介

川村 隆浩（正会員）

1994 年早稲田大学大学院 理工学研究科 電気工学専攻 修士課程修了．同年，(株) 東芝入社．同社 研究開発センター等に所属．2001〜02 年 米国カーネギ・メロン大学 ロボット工学研究所 客員研究員 兼任．2003 年より，電気通信大学大学院 情報システム学研究科 客員准教授 兼任．2007 年より，大阪大学大学院 工学研究科 非常勤講師 兼任．2015 年より，国立研究開発法人 科学技術振興機構 情報分析室 主任調査員．現在に至る．博士 (工学，早稲田大学)．2012 年 ISWC 10-Year Award 受賞．2013 年人工知能学会研究会優秀賞受賞．本学会理事，代議員等を歴任．主としてセマンティック Web，エージェント技術の研究・開発に従事．情報処理学会会員．

大須賀 昭彦（正会員）

1981 年上智大学 理工学部 数学科卒業．同年，(株) 東芝入社．同社研究開発センター，ソフトウェア技術センター等に所属．1985〜89 年 (財) 新世代コンピュータ技術開発機構 (ICOT) 出向．2007 年より，電気通信大学 大学院情報システム学研究科 教授．2012 年より，国立情報学研究所 客員教授 兼任．博士 (工学，早稲田大学)．主としてソフトウェアのためのフォーマルメソッド，エージェント技術の研究に従事．1986 年度情報処理学会論文賞，2013 年度本学会研究会優秀賞受賞．IEEE Computer Society Japan Chapter Chair，本学会理事，日本ソフトウェア科学会理事，同学会監事等を歴任．情報処理学会，電子情報通信学会，日本ソフトウェア科学会，電気学会，IEEE Computer Society 各会員．