オントロジー形式による交通関係アノテーション

Bou Savong　鈴木 直樹　三輪 誠　佐々木 裕
豊田工業大学
{savong, sd16042, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1　はじめに

本論文では，オントロジー形式の関係（Ontology-Style Relation; OSR）アノテーションアプローチを提案する．従来の関係抽出（Relation Extraction; RE）データセットでは，関係はエンティティメンション間のリンクとしてアノテーションされる．対照的に，OSR アノデーションでは，関係はリンクではなく関係メンションとしてアノデーションされ，関係メンションからエンティティメンションへの *domain* と *range* という項のリンクが付けられる．このアノテーションには以下の利点が期待できる:

- 関係注釈はオントロジー RDF (*Resource Description Framework*)[*1]トリプルに簡単に変換できるため，アノテーション付きの関係を使用してオントロジーエントリを作成できる．
- 関係は関係メンションとしてアノテーションされているため，従来の RE コーパスの関係タイプ分類タスクの一部は *Named Entity Recognition (NER)* タスクとなり, 多くの関係タイプを使用する RE タスク [1] と比較すると，深層学習は多くの NER [2] で 80% を超える F 値を達成しており，非常に高性能である．
- OSR アノテーションはオントロジーの内容の明確なドキュメントとして利用できる．オントロジーの内容がテキストに紐づくことで，オントロジーの内容の理解に役立つ．

ケーススタディとして，すでに別形式でアノテーションされている日本の交通ルールのコーパス [3] を OSR アノテーションに変換し，新しい OSR-RoR（Rules of the Road）コーパスを構築した．変換のアノテータ間一致率を測ったところ，85〜87% であり，高い一致率で変換できることがわかった．

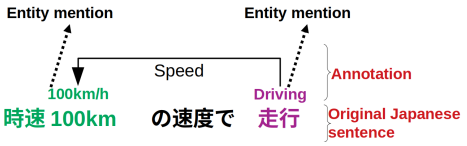
2　オントロジー形式の注釈付け

オントロジーの表現基盤は RDF である．RDF では，すべての情報は RDF トリプル (*subject*, *predicate*, *object*) で表現される．RDF スキーマ（rdfs）における 3 つの主要な述語には `rdfs:subClassOf`, `rdfs:domain`, と `rdfs:range` がある．オントロジーのクラスは，`rdfs:subClassOf` という名前の一般化関係によって階層構造になっている．たとえば，クラス C1 が C2 の一般化である場合，RDF トリプル形式では `(C2, rdfs:subClassOf, C1)` として表される．RDF では，トリプルまたはバイナリ関係の述語はプロパティと呼ばれる．プロパティは，オントロジーのノードとしても表される．たとえば，クラス C1 に C2 への関係 R1 がある場合，`(R1, rdfs:domain, C1)` と `(R1, rdfs:range, C2)` に表される．2 つのプロパティ間の一般化関係は `rdfs:subPropertyOf` で記述できる．

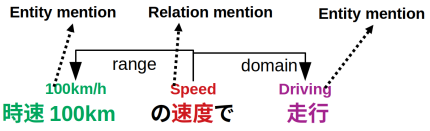
オントロジー形式の関係アノテーションと従来の関係アノテーションは，オントロジーとセマンティックネットワークとの関係に類似している．オントロジーではプロパティがノードとして表されるが，セマンティックネットワークではプロパティがラベル付きリンクとして表され，2 つの概念間のプロパティ/関係を記述するリンクラベルを使用して自由に構築できる．

本研究ではオントロジーと同じ形式での文書への関係のアノテーション (OSR アノテーション) を提案する．図 1は，従来のアノテーションと提案する OSR アノテーションの主な違いを示している．図 1aは従来の注釈であり，関係 *Speed* がリンクとしてアノテーションされている．一方で，図 1bは提案する OSR アノテーションであり，関係 *Speed* は，関係メンションとして注釈が付けられ，`domain` と `range` のリンクが *Driving* と *100km/h* それぞれに接続されている．アノテーシ

[*1] `https://www.w3.org/RDF/`



(a) 従来のアノテーション



(b) オントロジー形式のアノテーション

図1: 従来および提案するアノテーションの例

ョンの効率化のために，オントロジークラスとデータタイプは NE のカテゴリを参照すれば区別できるため，区別しないことに注意する．ただし，OSR アノテーションでは `rdfs:subPropertyOf` に注釈付けをしない．また，2 つの用語間の等価性を指定するための *Web Ontology Language (OWL)* クラス公理の 1 つである `owl:equivalentClass` を採用する．また，必要に応じて，`osr:partOf` を要素プロパティとして追加して，部分全体の関係を記述する．

3　対象文書と従来の関係アノテーション

河辺らは，安全運転に関する文書に，従来の関係表現形式でアノテーションを行った．アノテーションの対象となる文書は交通教則 [4] の規定を用いている．本研究では，このコーパスを改善したコーパス RoR (Rules of the Road) を用いた．表 1 に従来のコーパスの統計をまとめた．

4　オントロジー形式の関係アノテーション

4.1　オントロジークラス

RoR オントロジークラス (図 2) は，交通に関連する階層構造の概念である．用語には (1) ABSTRACT 概念, (2) CONCRETE 概念, (3) PROPERTY 概念 (relations), (4) VALUEs (データタイプ), (5) MODIFIERs の 5 つの主要なクラスを定義した．オントロジーでは値はデータタイプとして扱われるが，アノテーションにおいてはクラスとデータタイプは用語の概念の同じ階層に配置した．クラスの例を図 2に示す．

表1: 従来の関係注釈での RoR コーパスの統計

Type	Counts
# 章	11
# 節	49
# 文章	1,476
# 文字	68,655
# 用語タイプ	270
# 関係タイプ	99

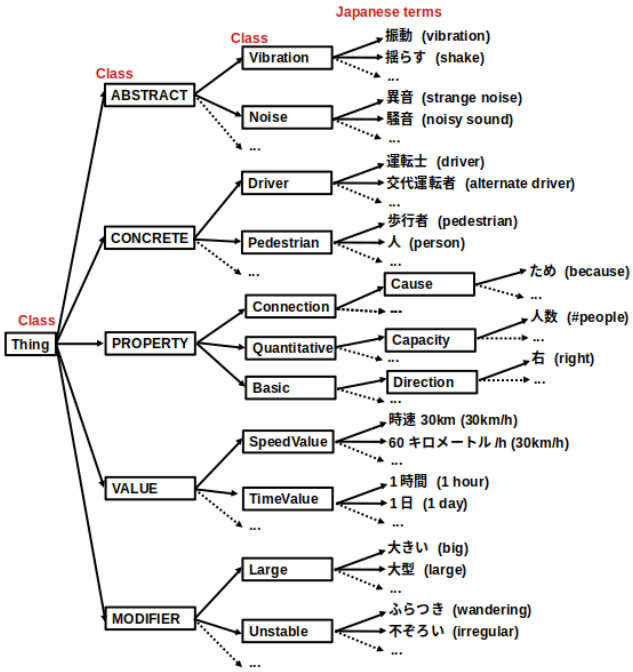


図2: 日本の道路交通法のクラス階層

4.2　オントロジー形式の関係

RoR コーパスでは，交通に関連する単語/フレーズは “term(s)” と呼ばれる．本研究では，従来の関係アノテーションを OSR アノテーションに変換する．リンクを使用して関係を維持するのではなく，関係メンション (relation mention) と呼ばれる中間用語を使用して，エンティティメンション (entity mentions) と呼ばれる他の 2 つの用語間の関係を維持する．次に，オントロジー構造は，“domain” リンクと “range” リンクを使用して，関係メンションを関係の元となるエンティティメンションと関係先となるエンティティメンションにそれぞれ接続することによって採用される．

アノテーションスキーマの設計では，リンクラベルの数を最小限に抑え，データセット内の関係を表現す

る際に標準の RDF プロパティをできるだけ使用することを目指した．例外として，関係を表現する適切な中間関係のメンションが見つからない場合，エンティティメンションは，従来のように関係固有のリンクラベルによって直接リンクした．

OSR-RoR コーパスの統計を表 2にまとめた．他の OSR-RoR リンクは，適切な仲介関係のメンションに変換されるはずのリンクラベルですが，文に適切な関係のメンションがないため，依然として直接リンクのままである．これは，約 94% (=3,816/(3,816+247)) の関係が OSR 関係に正常に変換されたことを示している．新しい RoR コーパス内のリンクタイプの数は，元の RoR コーパス内のリンクタイプのわずか $\frac{1}{9}$ に大幅に削減されている（c.f. 表 1の # 関係タイプ）．

5　アノテーションの例

この節では，主要なアノテーションの例を紹介する．

5.1　subClassOf 関係

subClassOf 関係は，RDF で使用される標準プロパティである．したがって，中間の関係メンションを使用せずに直接注釈を付ける．例を図 3に示す．“指示表示” と “規制表示” は “道路標識” のサブクラスであるため，図に示すように subClassOf 関係が維持される．

5.2　Property 関係

用語が別の用語または別の用語の修飾子を記述するために使用される場合，両方の用語は Property の関係

表2: OSR-RoR コーパスの統計．OSR リンクは，“domain”, “range”, “subClassOf”, “partOf”, と “equivalentClass” リンクを示す．

タイプ	数
#ABSTRACT クラス	34
#CONCRETE クラス	15
#PROPERTY クラス	25
#VALUE クラス	11
#MODIFIER クラス	5
# 属性	4
# リンクの種類	11
#Entity mentions	8,835
#Relation mentions	3,816
#OSR リンク	11,180
# その他の OSR-RoR 固有のリンク	247

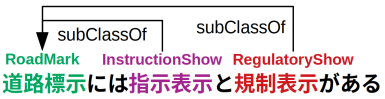


図3: subClassOf 関係の例

によって接続される．例を図 4に示す．“自分勝手” というエンティティメンションは，“通行” というエンティティメンションの特性を説明しているため，それらは関係メンションである “に” という “Property” 関係によって関連付けられている．次に，標準の RDF プロパティ “domain” および “range” を使用して，図に示すようにそれらの関係を維持する．

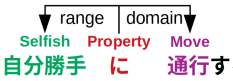


図4: プロパティ関係の例

5.3　条件関係

特定の条件で特定のアクションが実行される場合，そのような条件の関係（条件関係と呼ぶ）はデータセットで適切に示される必要がある．データセットでは，“Case” プロパティを使用して，テキストで見つかったすべての条件を表す．例を図 5に示す．図では，簡単のために，条件に直接接続されている関連用語と関係のみを示した．“とき = Case” という関係メンションはエンティティーメンション “通行” と “ 守る” 間の条件関係を作成するための関係メンションとして機能する．

図5: 条件付き関係の例

6　OSR アノテーションの評価

アノテーションの有効性の評価のため，Cohen’s kappa [5] を使用したアノテータ間一致率 (Inter-Annotator Agreements; IAA) を計測した．2 人のアノテータが同じ従来の関係注釈のセットを OSR 注釈に変換した．具体的には，OSR 注釈ガイドラインの説明を受けたのちに，2 人のアノテータが独立に従来の方法で注釈が付けられた 105 個の文を変換した,．IAA は表 3に示した．用語と関係の両方の Cohen’s kappa (κ-scores) のスコアは 85-87% であり，変換された結果

が “ほぼ完全に一致” レベルで一致していることを示した．不一致の部分を確認したところ，用語に関する主な原因は，選択した用語に日本の助詞を含めるべきかどうかの曖昧さが原因であった．また，関係については，誤って注釈が付けられた用語が原因であった．結果は，データセットの注釈ガイドラインが人間にとって明確であることを示している．

7 関連研究

これまで，多くの RE コーパスが RE タスク用に構築されている．*Automatic Content Extraction (ACE)* プログラム 2004 [6] では，個人名などの名前付きエンティティと，Part-Whole や User-Owner などの関係は一般的な英語，アラビア語，中国語の記事に注釈が付けられる．SemEval 2010 タスク [7] は，関係分類のみを対象としている．タスクは，文中の 2 つの与えられた 2 つのエンティティ $\langle e1 \rangle$ と $\langle e2 \rangle$ 間の関係を決定することである．関係タイプには，Content-Container および Entity-Destination が含まれる．これらは従来の関係アノテーションに基づいている．

従来の述語項構造やイベントのアノテーションとも異なっている．まず，注釈の対象が異なる．PAS およびセマンティック役割は名前付きエンティティを考慮しないため，長距離の引数を接続せず，我らよりも浅いセマンティック関係を処理する．イベントは通常，動的な関係を扱う．第二に，リレーションはバイナリリレーションであり，PAS，セマンティックロールおよびイベントは n -ary リレーションである．最後に他の注釈は RDF を考慮していない．

8 おわりに

本研究はオントロジー形式アノテーションという新しい注釈スタイルを提案した．ケーススタディとして，日本の交通規則の従来の関係抽出コーパスを OSR アノテーションに変換した．人間のアノテータによる一致率の評価では，OSR アノテーションへの変換においては，高いアノテータ一致率が達成できることがわかった．今後の課題としては，英語の RE コーパスを OSR アノテーションに変換し，利点を評価することが挙げられる．さらに，OSR アノテーションを介して 2 つの異なる情報ソースをリンクすることにより，テキストとオントロジーエントリを紐付けすることが考えられる．

本研究で作成した OSR-RoR コーパスを用いた用語抽出・関係抽出の評価に関する報告は，第 26 回言語処理学会年次大会の別の発表にて行う．

表3: IAA による評価 (Cohen’s Kappa)

用語注釈	関係注釈
0.8484	0.8719

謝辞

本研究の一部が JSPS 科研費 17K00318 により支援されたことに深く感謝する．

参考文献

[1] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *EMNLP 2017*, pp. 35–45, 2017.

[2] Khai Mai, Thai-Hoang Pham, Minh Trung Nguyen, Tuan Duc Nguyen, Danushka Bollegala, Ryohei Sasano, and Satoshi Sekine. An empirical study on fine-grained named entity recognition. In *COLING 2018*, pp. 711–722, August 2018.

[3] Kawabe Kazuhito, Miwa Makoto, and Sasaki Yutaka. Transportation terminology recognition for semi-automatic traffic ontology expansion. 言語処理学会第 21 回年次大会, pp. 135–138. ANLP, March 2015.

[4] National Public Safety Commission, Notice No. 3. Japan, October 1978. `https://www.npa.go.jp/koutsuu/kikaku/kyousoku/index.htm`.

[5] Cohen Jacob. A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*, pp. 37–46, 1960.

[6] George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. The automatic content extraction (ACE) program - tasks, data, and evaluation. In *LREC*, 2004.

[7] Yunfang Wu and Peng Jin. Semeval-2010 task 18: Disambiguating sentiment ambiguous adjectives. In *SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pp. 81–85, 2010.